Machine Learning for Finance

Beginner's Guide to Explore Machine Learning in Banking and Finance

SAURAV SINGLA

bpb

Machine Learning for Finance

Beginner's Guide to Explore Machine Learning in Banking and Finance

Saurav Singla

www.bpbonline.com

FIRST EDITION 2021

Copyright © BPB Publications, India

ISBN: 978-93-89328-622

All Rights Reserved. No part of this publication may be reproduced, distributed or transmitted in any form or by any means or stored in a database or retrieval system, without the prior written permission of the publisher with the exception to the program listings which may be entered, stored and executed in a computer system, but they can not be reproduced by the means of publication, photocopy, recording, or by any electronic and mechanical means.

LIMITS OF LIABILITY AND DISCLAIMER OF WARRANTY

The information contained in this book is true to correct and the best of author's and publisher's knowledge. The author has made every effort to ensure the accuracy of these publications, but publisher cannot be held responsible for any loss or damage arising from any information in this book.

All trademarks referred to in the book are acknowledged as properties of their respective owners but BPB Publications cannot guarantee the accuracy of this information.

Distributors:

BPB PUBLICATIONS
20, Ansari Road, Darya Ganj
New Delhi-110002
Ph: 23254990/23254991

MICRO MEDIA
Shop No. 5, Mahendra Chambers,
150 DN Rd. Next to Capital Cinema,
V.T. (C.S.T.) Station, MUMBAI-400 001
Ph: 22078296/22078297

DECCAN AGENCIES
4-3-329, Bank Street,
Hyderabad-500195
Ph: 24756967/24756400

BPB BOOK CENTRE
376 Old Lajpat Rai Market,
Delhi-110006
Ph: 23861747

To View Complete
BPB Publications Catalogue
Scan the QR Code:

Published by Manish Jain for BPB Publications, 20 Ansari Road, Darya Ganj, New Delhi-110002 and Printed by him at Repro India Ltd, Mumbai

www.bpbonline.com

Dedicated to

My wife Vidhu Gupta

About the Author

Saurav Singla is a Senior Data Scientist, a Machine Learning Expert, a Technical Writer, a Data Science Course Creator, a Mentor, a Speaker, and a Podcaster.

He has fifteen years of comprehensive experience in statistical modeling, machine learning, natural language processing, deep learning, and data analytics.

He has worked in senior roles for JBSA, Valiance, Siemens, and Harvey Norman.

He has demonstrated success in developing and seamlessly executing plans in complex organizational structures. He has also been recognized for maximizing performance by implementing appropriate project management tools through analysis of details to ensure quality control and understanding of emerging technology.

Outside work, Saurav volunteers his spare time for helping, coaching, and mentoring young people in taking up careers in the data science domain.

He has created two courses on Udemy, with over twenty thousand students enrolled in it. He regularly authors articles on data science.

About the Reviewer

Pedro Lewis Blesa Crucefix has spent years researching new sources of information, data science techniques, and opportunities offered by AI in organizations across all industries. Pedro graduated in business, majoring in Strategy from the University of Catalonia, and holds a master's degree in Business Analytics from the University of Westminster. He has worked with companies such as Ladbrokes, Disney, and The Automobile Association. His most recent experience includes identifying opportunities in data extracted from connected cars.

Acknowledgement

There are a few people I want to thank for the continued and ongoing support they have given me during the writing of this book. First and foremost, I would like to thank my parents, my wife, and two sons, Aarav and Parth, for putting up with me while I was spending many weekends and evenings on writing—I could have never completed this book without their support.

This book wouldn't have happened if I hadn't had the support from my publisher, BPB Publications. I would like to thank them for giving me this opportunity to write my first book for them.

Preface

The field of mechanical autonomy has significantly progressed with new wide-ranging and innovative accomplishments. One is the ascent of huge amount of information, which offers a greater chance to incorporate programming ability with mechanical frameworks. Another is the use of advanced sensors and associated gadgets to screen environmental factors like temperature, pneumatic force and light; the sky is the limit. These achievements have helped develop advanced robots for use in assembling, health care, security, etc.

Robots are generally used to perform straightforward and redundant tasks such as in vehicle production and in businesses involving hazardous situations. Several parts of mechanical technology include human-made consciousness; robots might be furnished with what might be compared to human faculties (e.g., vision, touch, and the capacity to detect temperature). Some are even fit for basic straightforward leadership. Research in mechanical autonomy has led to the design of robots with a level of independence that will allow versatility and basic leadership in an unstructured domain. The present mechanical robots don't look like humans; a robot with human-like appearance is called an android.

At present, we are habituated to the Internet in numerous ways in our everyday life. For example, to explore an obscure place, we use Google Maps. We use social media to express our musings or sentiments. Or, to know the latest news, we access online news websites. If we attempt to comprehend the impact of science in our life, we will see that, really, these are all applications of AI and machine learning. Of late, there has been increased interest in machine learning, and more and more people are realizing the extent of new applications empowered by the ML approach. It fabricates a guide to make contact with the gadget and make the gadget reasonable to react to our reactions. As our life gets progressively computerized, in this book, we will discuss some of the mind-blowing applications of machine learning.

The fields of adaptive machining, profound learning, and computerized reasoning are quickly expanding, and will probably do so for a long time to come. The advancements have been phenomenal, opening new ways to deal with long-standing innovation challenges (e.g., progress in computer vision and picture investigation). These innovations alone are opening new pathways and applications in the field of ICME/MGI.

Over the nineteen chapters in this book, you will learn the following:

Chapter 1 discusses the process of how machines are taught, and the relation between machine learning, AI, deep learning, data mining, and statistics. It then focuses on the application of machine learning in the finance domain.

Chapter 2 discusses Naive Bayes, normal distribution, and automatic cluster detection with Gaussian process in depth. It then focuses on the application of machine learning in the cybersecurity domain.

Chapter 3 discusses the difference between structured and unstructured data and how advanced ML tools have helped businesses make quicker decisions. It then focuses on how ML tools are helping business put data in order so that it can be processed properly and in a more orderly manner.

Chapter 4 gives a tour of NLP, and it reviews the advantages and applications of NLP.

Chapter 5 is a key chapter as it discusses, in depth, computer vision and its applications. It then reviews the neural network architecture of computer vision. It then also discusses the application of computer vision in image recognition, biometric recognition, and software vulnerabilities.

Chapter 6 gives an in-depth discussion on neural networks, and how they work, and their different types and benefits. It then reviews gradient boosting machines and gradient descent.

Chapter 7 describes some of the different types of sequence modeling techniques, and reviews the ML modeling procedure, which involves feature engineering and selecting a model, model training, validating the model, and testing the model on new data.

Chapter 8 discusses which ML algorithm is suitable for your business problem. It then shows what a data reduction technique is. It also reviews the concept, types, and application of reinforcement learning.

Chapter 9 describes why finance is at the forefront of technology. It further discusses how machine learning can benefit the finance industry. It also reviews some use cases and the impact of ML on the finance industry.

Chapter 10 describes the impact of technology on FinTech companies. It then describes the key challenges faced by FinTech companies.

Chapter 11 reviews ML use cases in the banking sector. It also discusses the application of machine learning in cybersecurity, credit scoring, algorithmic trading, and robo-advising.

Chapter 12 describes the different ways information can be stolen. It then discusses the different security measures to stop this fraud.

Chapter 13 discusses data mining and data visualization, and their real applications.

Chapter 14 discusses the advantages and disadvantages of machine learning.

Chapter 15 discusses in depth the application of machine learning.

Chapter 16 reviews the impact of machine learning not only on the financial sector but also on our lives, economy, and humanity as a whole.

Chapter 17 discusses the applications of AI in banking.

Chapter 18 discusses in depth traditional algorithms such as regression, k-means clustering, k-nearest neighbor, principal component analysis algorithm, polynomial fitting and least squares algorithm, forced linear regression algorithm, support vector machine algorithm, conditional random fields algorithm, and decision tree algorithm.

Chapter 19 lists some frequently asked questions about machine learning.

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Did you know that BPB offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.bpbonline.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at :

business@bpbonline.com for more details.

At **www.bpbonline.com**, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on BPB books and eBooks.

BPB is searching for authors like you

If you're interested in becoming an author for BPB, please visit **www.bpbonline.com** and apply today. We have worked with thousands of developers and tech professionals, just like you, to help them share their insight with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

The code bundle for the book is also hosted on GitHub at **https://github.com/bpbpublications/Machine-Learning-for-Finance**. In case there's an update to the code, it will be updated on the existing GitHub repository.

We also have other code bundles from our rich catalog of books and videos available at **https://github.com/bpbpublications**. Check them out!

PIRACY

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at :

business@bpbonline.com with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**.

REVIEWS

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions, we at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Table of Contents

1. **Introduction 1**
    - Introduction 1
    - Structure 3
    - Objective 3
    - How machines are taught 3
    - Factors contributing to the success of machine learning 4
    - Machine learning and artificial intelligence 4
    - Machine learning and deep learning 5
    - Machine learning and statistics 5
    - Machine learning and data mining 6
        - *What's the difference? 6*
    - Machine learning in finance 6
    - Importance of machine learning in finance 7
    - Robo-warning 7
    - How to utilize machine learning in finance 8
    - Utilize outsider machine learning arrangements 9
    - Development and combination 10
    - How is machine learning used today 10
        - *Financial services 10*
    - Conclusion 11
2. **Naive Bayes, Normal Distribution, and Automatic Clustering 13**
    - Introduction 13
    - Structure 15
    - Objective 15
    - Naive Bayes 15
        - *Bayesian classification 16*
        - *Strengths, weaknesses, and parameters of the Naive Bayes algorithm 16*

Applications of the Naive Bayes algorithm *17*
Recommendation systems *18*
Normal distribution 18
Automatic cluster detection in data mining 19
Searching for simplicity islands.. *20*
Light of the moon, bright star..*21*
Gaussian model.. *24*
Application of machine learning in cybersecurity............... 26
Spam detection.. *27*
Phishing page detection *27*
Malware detection... *29*
DoS and DDoS attack detection...*31*
Anomaly detection *32*
Conclusion 33

3. Machine Learning for Data Structuring.................................... 35
Introduction.. 35
Structure 35
Objective... 36
Data structuring 36
The future of big data... 36
Structured and unstructured data... 40
Conclusion 42

4. Parsing Data Using NLP 43
Introduction.. 43
Structure 44
Objective... 44
Uses of NLP 44
Key advantages of NLP 45
Data handling in NLP 45
NLP applications 46
Talent recruitment *46*

Voice assistants *46*
Health care *46*
Survey analysis *46*
Grammar checkers *47*
Email filtering *47*
Social media monitoring *47*
Online search autocomplete and autocorrect *47*
Conclusion 48

5. Computer Vision 49
Introduction 49
Structure 49
Objective 50
Computer vision application 50
Neural networks in computer vision 51
Activation function *51*
Data preprocessing *51*
Weight initialization *51*
Introduction to CNN *51*
Convolutional layers *52*
Pooling layers *52*
Fully connected layers *52*
Overview of computer vision 52
Image recognition 54
Biometric recognition 56
Software vulnerabilities 59
Conclusion 61

6. Neural Network, GBM, and Gradient Descent 63
Introduction 63
Structure 64
Objective 64
Working of neural networks 64

Types of neural networks in AI.. 77

Feed-forward neural networks.. *77*

Recurrent neural networks.. *78*

Convolutional neural networks (CNNs)............................. *78*

Modular neural networks.. *78*

Benefits of using artificial neural networks 79

Gradient boosting algorithms.. 79

GBM *79*

Light GBM *79*

XGBoost *80*

CatBoost *80*

Gradient descent *80*

Conclusion 81

7. Sequence Modeling 83

Introduction... 83

Structure... 83

Objective... 84

Word embedding 84

Feed-forward neural network algorithm 84

Convolutional neural network algorithm............................. 84

Recurrent neural networks (RNN) algorithm 85

Conditional random field (CRF) algorithm 85

Modeling procedure.. 86

Feature engineering and selecting a model........................... *86*

Training the model... *87*

Validating the model.. *87*

Predicting new observations... *89*

Conclusion 89

8. Reinforcement Learning for Financial Markets 91

Introduction... 91

Structure... 91

Objective.. 92
Problem types in machine learning 92
Identifying key predictors (data reduction).......................... 93
Learning from experience (reinforcement learning)........... 94
Fundamentals of reinforcement learning.............................. 96
Agent..97
Algorithms for control learning... 98
Applications of reinforcement learning 99
Reinforcement learning algorithms 99
Types of reinforcement learning... 100
Applications of reinforcement learning in real life........... 100
Conclusion 101

9. Finance Use Cases 103
Introduction... 103
Structure.. 103
Objective.. 104
Technology and finance 104
Automation.. 105
You might not hear the robots coming............................... 105
FinTech: Robocalypse comes to finance.............................. 106
The impact of FinTech.. 109
Guidelines to live by 113
Innovative technologies...114
AI/robotics including chatbot 114
Cognitive computing 115
Blockchain in the banking industry....................................115
What is Blockchain? 115
First blockchain user in the world: Australian Securities Exchange..116
Internet of things (IoT) in banking....................................117
Bank of America: Big data projects....................................117
Voice biometrics 118

Digital bank 118

AI as a strategy at the top level 119

Cost of investing in AI *120*

Crucial leadership involvement *120*

Inorganic growth *121*

Development status of different AI technologies 122

Risk management 124

Fraud detection and prevention 124

Improving the truth of financial rules and designs 124

Trading 125

AI in banking 125

Conclusion 126

10. Impact of Machine Learning on FinTech 127

Introduction 127

Structure 127

Objective 127

Overview of FinTech companies 128

Impact of technology 129

Challenges 129

Conclusion 130

11. Machine Learning in Finance 131

Introduction 131

Structure 131

Objective 132

Machine learning use cases in banking 132

Security 133

Guaranteeing and credit scoring 133

Algorithmic exchanging 134

Robo-advisors 134

Utilize outsider machine learning arrangements 135

Applications of machine learning 136

Current financial applications 136

Portfolio management *136*

Algorithm trading *137*

Detecting fraud *137*

Insurance or credit insurance...*137*

Machine learning and cryptocurrencies............................. 138

Day trading with machine learning *138*

Conclusion 139

12. eKYC and Anti-Fraud Policy.. 141

Introduction... 141

Structure 141

Objective... 142

Big data analytics: True Buzzword of today 142

How criminals obtain information for online banking..... 142

Common ways in which information can be stolen.......... 142

Retail stores or restaurants *143*

Online portals *143*

Hacked mail accounts *143*

ATMs 143

Pickpockets and thieves... *144*

Employee records *144*

Fake calls *144*

Social media *144*

Security measures 144

Some tips on card security *145*

Security tips for Internet banking *145*

Conclusion 146

13. Uses of Data Mining and Data Visualization 147

Introduction... 147

Structure 147

Objective.. 148

Data visualization.. 148

Data mining.. 148

Future health care.. 148

Education.. 149

Customer relationship management.. 149

Criminal investigation... 149

Fraud detection.. 149

Customer segmentation.. 150

Intrusion detection... 150

Lie detection.. 150

Conclusion... 151

14. Advantages and Disadvantages of Machine Learning............ 153

Introduction.. 153

Structure.. 153

Objective.. 153

Advantages... 154

Identifies patterns.. *154*

Improves efficiency.. *154*

Completes specific tasks... *154*

Helps machines adapt to the changing environment.......... *155*

Helps machines handle large data sets............................... *155*

Disadvantages.. 155

Concepts involved in machine learning............................. *156*

Statistics... *156*

Brain modeling... *156*

Adaptive control theory.. *156*

Psychological modeling... *156*

AI... *157*

Evolutionary models.. *157*

Conclusion... 157

15. Applications of Machine Learning in Other Industries....... 159

Introduction.. 159

Structure.. 159

Objective.. 159

General applications of machine learning 160

Conclusion 166

16. Ethical Considerations in Artificial Intelligence.................... 167

Introduction.. 167

Structure.. 168

Objective.. 168

Loss of jobs... 168

Inequality 169

Humanity... 169

Disinformation 170

Artificial intelligence and crime 171

Racist robots 171

Artificial intelligence vs. humans....................................... 172

Conclusion 172

17. Artificial Intelligence in Banking... 173

Introduction.. 173

Structure.. 173

Objective.. 174

Fraud detection 174

Cost cutting.. 174

Customer service 175

Risk management 175

Internet banking... 176

Conclusion 176

18. Common Machine Learning Algorithms 177

Introduction.. 177

Structure.. 177

Objective 178
Regression 178
Linear regression *182*
Logistic regression *183*
k-means clustering 184
Three steps of the k-means algorithm *185*
What does k mean? *187*
Distance and similarity *189*
KNN algorithm 190
How does KNN work? *190*
Preparing your training data for KNN *193*
Final remarks *194*
Principal component analysis (PCA) algorithm 194
Polynomial fitting and least squares algorithm 195
Forced linear regression algorithm 195
Support vector machine (SVM) algorithm 195
Conditional random fields (CRFs) algorithm 195
Decision tree algorithm 196
Conclusion 196

19. Frequently Asked Questions 197
Conclusion 206
Approaching a machine learning problem 206
Humans in the loop 206
Testing production systems 207
Next step 207
Machine learning packages 207
Where do we go from here? 208

Index .... 211 - 217

CHAPTER 1 Introduction

Introduction

Machine learning is a branch of artificial intelligence that allows machines to learn and act the same way humans do. This allows them to come up with different kinds of output on their own.

Normally, machines are programmed to act a specific way depending on the actions that the user performs. This means that the user can also dictate what the outputs should be. Basically, humans still guide computers throughout the process.

In the case of machine learning, however, there is no need for the machine to be programmed in a specific way. Humans do not have to direct a specific path for the machine to take.

All it takes is the right data set. The machine will study the patterns in the data set. This will allow the machine to make its own decisions based on those patterns, and most (if not all) of it will be done autonomously or without human intervention.

Machine language (ML) advances from the investigation of free data available and advanced calculations on information and to give us meaningful insight. It is so inescapable today that a significant number of us likely use it a few times each day without even knowing it.

In prior phases of advancement in machine learning, the organizations that most profited from the new field were data firms and online organizations that saw and took advantage of the huge amount of information available. The capacity to give genuinely necessary information and data spoke to an unmistakable first mover's bit of leeway for these organizations. While the first movers for quite a while were the huge victors, their favorable position won't last any longer as efficiency levels out. The development of Analytics 3.0 is a distinct advantage, because of the scope of business issues that savvy mechanization—a blend of artificial intelligence (AI) and machine learning—can unravel is expanding each day. At this stage, each firm in any industry can benefit from clever computerization. Organizations that invest promptly in AI can increase their long-haul profits. To press home these advantages, organizations must reevaluate how they can benefit from the information with regards to Analytics.

Enormous changes are hatching in the showcasing scene, and these movements are, to a great extent, down to make ML powerful. Such is its effect that 97% of pioneers accept the eventual fate of promoting will comprise of keen advertisers working in a joint effort with AI-based mechanization elements.

Machine learning methods are utilized to tackle a large group of different issues, and organizations continue to profit a lot as we veer towards a universe of hyper-joined information, channels, substance, and setting. For the advanced advertising group, machine learning is tied in with discovering bits of prescient information in the influxes of organized and unstructured data and utilizing them to further their potential benefit.

The ability to react rapidly and precisely to changes in client conduct is basic in this day and age, and hence the need for AI. In this chapter, we investigate the advancements in machine learning that are being utilized successfully, and its potential uses in different organizations. AI is known as man-made brainpower. It very well may be viewed as part of ML. The historical backdrop of AI can enable us to comprehend it better, so let us do a quick review.

Structure

In this chapter, we will cover the following topics:

- How machines are taught.
- Factors contributing to the success of machine learning.
- Machine learning and artificial intelligence.
- Machine learning and deep learning.
- Machine learning and statistics.
- Machine learning and data mining.
- Machine learning in finance.
- Importance of machine learning in finance.
- Robo-warning.
- How to utilize machine learning in finance.
- Utilize outsider machine learning arrangements.
- Development and combination.
- How is machine learning used today.

Objective

After studying this chapter, you should be able to do the following:

- Understand the process of how machines are taught, and the relation between machine learning, AI, deep learning, data mining, and statistics.
- Understand the application of machine learning in the finance domain.

How machines are taught

The entire process can be considered complex. There are also different approaches applied. However, for the sake of translating the basics and to give an overview of what happens during the process, here are the three basic parts of machine learning:

1. **Data input:** The information or data sets to be used are fed to the machine. These data can come in the form of SQL databases, text files, spreadsheets, or anything similar.
2. **Data abstraction:** At this stage, the data is labeled and represented as required. It is then analyzed using the

algorithm chosen for the process. This is where the basic learning process happens.

3. **Generalization:** Once the learning is completed, the machine starts to develop its own insights. From these insights, it comes up with an output. Not all machine learning processes require an output, though. In some cases, the goal is only to cluster the data together.

Note that, in this process, the goal is always to create a better version of the machine. After the process ends, it is expected that the machine becomes smarter regardless of whether there is an output or not.

Factors contributing to the success of machine learning

Although the computers used in the process are definitely more advanced than the regular ones, of course, there is still a margin of error to be considered. Because of this, there is a need to zero in on what could increase the chances of success.

These factors come to mind when it comes to ensuring success in machine learning:

- How well the generalization goes
- How well the machine can apply what it learned to practical use

When these two areas are done well, expect that the results will show a success. These are also the key elements in ensuring that a future course of action can be predicted and planned for.

Machine learning and artificial intelligence

Machine learning and AI and are closely related, but it is highly inappropriate to interchange the two terms. These are different concepts.

Artificial intelligence is a more general term that covers a number of applications. It involves the ability of machines to mimic the behavior of humans. This also includes the ability of machines to make intelligent decisions on their own.

Machine learning falls under artificial intelligence as it also gives the machine the ability to think. But where artificial intelligence covers all concepts involving machines acting and thinking the same way as humans do, machine learning focuses on a machine's ability to learn on its own. As mentioned in the earlier definition of the term, this ability comes without the need for specific programming.

Machine learning and deep learning

Just like artificial intelligence, deep learning is yet another concept that is closely related to machine learning but is still essentially a different application altogether.

Deep learning, in essence, involves the creation of artificial neural networks. These networks use algorithms to learn and make decisions on their own.

Machine learning and statistics

Statistics, as you probably know, deal with data coming from either an entire population or from samples drawn from that population. From there, you can carry out analyses and draw inferences.

Statistical techniques are used in a number of applications like conditional probability, regression, standard deviation, variance, and a lot more.

So how do statistics fit into machine learning?

Although machine learning is part of computer science and statistics is part of mathematics, they work hand in hand in delivering results for artificial intelligence.

One example is the way your emails are segregated in your inbox. Let's say you want to determine which emails are important and which ones should be recognized as spam. In this case, a machine learning algorithm called Naive Bayes will observe past spam emails to come up with a way to identify new emails coming in as spam.

Naive Bayes uses a form of statistical technique that is the basis for conditional probability. This technique will be discussed in a later chapter.

Machine learning and data mining

Again, it's the use of data in both machine learning and data mining that makes people think that these two concepts are the same or are closely related.

Basically, data mining is a term that describes the process of searching through data for specific information. Machine learning, on the other hand, is only concerned with one thing – completing the task it was asked to do using the algorithms applied.

What's the difference?

If someone is teaching you how to play the guitar, that's a process that describes machine learning. If someone asks you to look for the best guitar performances ever, then that's data mining.

Machine learning in finance

In finance, machine learning can do something amazing, even though there is no enchantment behind it (well, perhaps only a tad). In any case, the accomplishment of a machine learning undertaking depends on the structure of the foundation, gathering appropriate data sets, and applying the correct calculations.

Machine learning is making noteworthy advances in the finance-related administration industry. We should perceive any reason why budgetary organizations should keep in mind the advantage of machine learning, what arrangements they can actualize with machine learning and artificial intelligence, and how precisely they can apply this innovation. Most finance-related administration organizations are as yet not prepared to identify the genuine incentive of this innovation for the following reasons:

- Businesses have unrealistic desires and expectations towards *machine learning solutions* and their incentive for their associations.
- R&D in machine learning is expensive.
- The lack of data science/machine learning specialists is another significant concern.
- Financial savvy people are not deft enough with regards to refreshing the information.

Importance of machine learning in finance

Despite the difficulties, numerous budgetary organizations, as of now, utilize this innovation. They do it for a lot of valid justifications:

- Reduced operational costs because of procedure computerization.
- Increased incomes because of better profitability and upgraded client encounters.
- Better consistency and fortified security.

There is a wide scope of open-source *machine learning* algorithms and applications/ tools that fit extraordinarily with budgetary information. Also, established budgetary administration organizations have significant subsidies that they can stand to spend on cutting-edge registering machinery. Because of the enormous volumes of transactional information, *machine learning* can improve numerous parts of the budgetary environment. This is the reason such a significant number of financial organizations are investing vigorously in R&D on artificial intelligence. In the case of slowpokes, it can be expensive to disregard artificial intelligence and machine learning.

Robo-warning

Robo-advisors are currently in demand in the finance sector. As of now, there are two noteworthy uses of machine learning in the warning area.

Portfolio board is an online executives' administration tool that uses calculations and insights to assign, oversee, and streamline customers' advantages. Clients enter their present monetary resources and objectives, state, sparing a million dollars by the age of fifty. A robo-advisor at that point assigns the present resources crosswise over venture openings depending on hazard inclinations and ideal objectives.

Several online protection administrators use robo-advisors to prescribe customized protection plans to a specific client. Clients select robo-advisors over close-to-home financial advisors because of lower charges, just as customized and aligned proposals.

How to utilize machine learning in finance

Despite the considerable number of points of interest in machine learning and artificial intelligence, even organizations with deep pockets frequently experience serious difficulties extricating the genuine incentive from this innovation. Financial administrations need to use *machine learning* sensibly as they do not have a clear idea of how information science functions and how to utilize it.

Consistently, they experience difficulties such as the absence of business KPIs. This brings about ridiculous estimates and depletes spending plans. It isn't sufficient to have a reasonable programming foundation set up (although that would be a decent start). It takes an unmistakable vision, strong and specialized ability, and assurance to convey an important *machine learning* advancement venture.

When you have a decent comprehension of how this innovation will accomplish your business targets, continue with plan approval. This is an undertaking for information researchers. They research the plan and help you formulate reasonable KPIs and make sensible appraisals.

Note that you need all of the information gathered by now. Else, you would require an information specialist to gather and tidy up this information.

Contingent upon a specific use case and business conditions, financial organizations can pursue various ways to implement machine learning. How about we look at them?

Artificial intelligence came into spotlight on getting meaningful insight from raw data using natural language processing.

Frequently, financial organizations start their artificial intelligence ventures to acknowledge they need legitimate information building. *Max Nechepurenko*, a senior information researcher at N-iX, said the following:

"When building up a [data science] arrangement, I'd prompt utilizing the Occam's razor rule, which means not overcomplicating. Most organizations that go for artificial intelligence, in truth, need to concentrate on strong information designing, applying measurements to the collected information, and representation of that information."

Simply applying factual models to handle well-organized data would be sufficient for a bank to resolve different bottlenecks and wasteful aspects in its activities.

What are the instances of such bottlenecks? They could be long queues at a particular branch, dull undertakings that can be dispensed with, wasteful HR exercises, blemishes in the banking application, etc.

In addition, the greatest challenge of any data science venture is building an organized biological system of stages that gather siloed data from several sources like CRMs, websites, social platform, spreadsheets, and that's just the beginning.

Before applying any calculations, the data needs to be properly organized and tidied up. Then, that data can be transformed into experiences. ETL (extricating, changing, and stacking) and further cleaning of the data represents around 80% of the time for artificial intelligence undertaking.

Utilize outsider machine learning arrangements

Regardless of whether or not your organization chooses to use *machine learning* in its up and coming task, you don't need to develop new calculations and models. Most machine learning undertakings manage issues that have just been tended to. Tech goliaths like Google, Microsoft, Amazon, and IBM sell machine learning programming as an administration tool.

These out-of-the-crate arrangements are, as of now, prepared to explain different business errands. On the off chance that your task covers similar use cases, do you accept that your group can outflank calculations from these tech titans with goliath R&D focuses?

One genuine model is Google's Recommendation AI ecommerce tool. That product applies to different spaces, and it is recommended to check on whether it fits your business case.

A *machine learning* architect can execute the framework by focusing on your particular data and business space. The expert needs to extract the data from various sources, transform it for the specific framework, get the outcomes, and visualize the discoveries.

The trade-offs are the absence of authority over the external framework and constrained arrangement adaptability. Plus, AI calculations don't fit into each utilization case. *Ihar Rubanau*, a senior data researcher at N-iX, said the following:

"An all-inclusive machine learning algorithm doesn't exist, yet. Data researchers need to alter and tweak calculations before applying them to various business cases crosswise over various areas."

So, if a current arrangement from Google comprehends a particular assignment in your specific area, you ought to likely utilize it. If not, go for custom improvement and mix.

Development and combination

Building up a *machine learning* arrangement without any preparation is one of the most dangerous, costliest, and tedious alternatives. In any case, this might be the best way to apply *machine learning* innovation to some business cases.

Machine learning tasks focus on a special need in a specific specialty, and they require a top-to-bottom examination. In the event that there are no arrangements available to take care of those particular issues, external AI programming is probably going to deliver incorrect outcomes. You will most likely need to depend extensively on the open-source AI libraries from Google and the preferences. Current *machine learning* tasks are generally about applying the existing best in class libraries to a specific space and use case.

How is machine learning used today

Financial services

Experian is one of the, if not *the*, biggest credit reference agency in the world. They store an amazing amount of data about every individual in their records. They can convey any financial institution about your purchases, court cases, and other relevant information that could help banks figure out how your finances stand.

Knowing how much data their system goes through for every loan and mortgage application being processed, it isn't surprising to learn that they have machine learning helping them out. Basically, machine learning sorts through all that data and tells them whether a certain individual is a risky bet, or whether a loan application is worth approving.

American Express is another financial giant that takes advantage of machine learning. With over 110 million active AmEx cards, how do they manage to keep track of fraudulent activity?

You guessed it — through machine learning. By using the right data sets and algorithms, AmEx has the ability to notice discrepancies in spending habits among card users. This allows them to detect potential fraud in real-time.

Conclusion

In this chapter, we discussed the process of how machines are taught. We also discussed the relation between machine learning, artificial intelligence, statistics, deep learning, and data mining. We also discussed the application of machine learning in the finance domain.

In the next chapter, we will learn about normal distribution and automatic clustering.

CHAPTER 2

Naive Bayes, Normal Distribution, and Automatic Clustering

Introduction

The seemingly unstoppable interest in machine learning stems from the same variables that data mining and Bayesian analysis are applied more. The underlying factors contributing to this popularity are increasing amounts and varieties of data, cheaper and more effective computational processing, and cheap data storage. To get an idea of how important machine learning is in our daily lives, it's easier to pinpoint which part of our advanced way of life hasn't been affected by it. Every aspect of human life is influenced by smart machines that are designed to expand human capabilities and improve efficiency. Artificial intelligence and machine learning are central to the "fourth industrial revolution," which may cast doubt on our human thoughts.

All these factors imply that models that can analyze larger, more complicated data and deliver highly accurate results in a short time, produced quickly and automatically on a much larger scale. Businesses can easily identify potential growth opportunities or avoid unknown dangers by constructing desired machine learning models that meet their business requirements. Data runs through the veins of every company. Data-driven strategies are increasingly emerging as the distinguishing feature between the winner and the loser. Machine learning offers the magic of unlocking the importance of business and customer data to lead to actionable measures and decisions that can skyrocket a company's business and market share.

Machine learning has shown, in recent years, that many tasks that were once considered human-only activities can be automated, such as image recognition, word processing, and gaming. In 2014, machine learning and AI professionals thought it would take at least ten years for the machine to defeat its biggest player in the world in the board game Go. But they were mistaken; Google's DeepMind showed that machines can learn which movements to take into account, even in such a complicated game as Go. In the world of gaming, machines have seen several innovations, such as Dota Bot from the OpenAI team. Machine learning undoubtedly has enormous economic and social consequences in our daily lives. A full range of work activities and the entire industrial spectrum could potentially be automated, and the labor market will be transformed forever.

Machines can now learn and train on their own using previous calculations and underlying algorithms to produce high-quality, easily reproducible decisions and results. Machine learning has been around for a long time, but recent advances in machine learning algorithms have made it possible for machines to efficiently process and analyze large amounts of data. This is accomplished by using high speed and frequency automation to apply advanced and complex mathematical calculations to the machines. Today's sophisticated computers can quickly evaluate the massive amounts of data and deliver faster and more accurate results. Companies using machine learning algorithms have improved flexibility to tailor the training data set to their business requirements and train the machines accordingly. With these custom machine learning algorithms, companies can identify potential hazards and growth opportunities. Working with artificial intelligence and cognitive technologies, machine learning algorithms are used to develop computers that are highly effective and efficient at handling large amounts of data, or big data, and produce highly accurate results.

Hundreds and thousands of machine learning algorithms have already been generated as this research field continues to expand. Here are some of the most commonly used algorithms, categorized by machine learning type:

To refresh your memory, guided learning is driven by data scientists who teach the algorithm which conclusions to draw using a predefined training data set. Guided learning requires data about all possible outputs from the algorithm and training data set that are already labeled with expected or correct results.

Let's take a look at the two most famous guided learning algorithms used to develop machine learning models, in detail.

Structure

In this chapter, we will cover the following topics:

- Naive Bayes
- Normal distribution
- Automatic cluster detection in data mining
- Gaussian model
- Application of machine learning in cybersecurity

Objective

After studying this chapter, you should be able to do the following:

- Understand the technical aspects of Naive Bayes, Normal distribution, and automatic cluster detection.
- Understand the application of machine learning in the cybersecurity domain.

Naive Bayes

Naive Bayes is based on Bayes' rule. This classification approach assumes that different predictors are independent of each other, and that one feature has no relation at all with any other features present.

You may call a fruit an orange, for example, because of its orange color, round shape, and the texture of its skin. But this doesn't mean that the color, shape, and texture rely on each other to prove that the fruit being observed is an orange.

This model is exceptionally helpful if you have huge data sets to deal with. In fact, it has been proven to be the most efficient method when it comes to data classification.

Bayesian classification

Bayesian classification depends on the concept of conditional probability. Given that one condition is true, what is the probability that a second condition is also true? Think about this in the context of weather forecast. If you know that skies are overcast today, what is the probability of rain? If you know the skies are clear today, what is the probability of rain? It shouldn't be too hard to understand why the probability in the first case will be much higher than that in the second case.

In addition, for a Bayesian classifier, we need to be able to provide a general idea about how many items are expected to be in the groups after we classify them. In the case of weather, we might already know that, on average, 10% of our days will be rainy and 90% will be sunny (assuming we live in Arizona). These are called prior probabilities. This method is often used for building utilities like spam filters—if the word "Viagra" appears in an email, there's a high probability that it's spam. In general, system administrators will know what proportion of incoming emails are likely to be spam (prior probabilities). The more accurate estimates we can provide for our priors, the better (in general) the classifier should be.

In the below example, we observed four emails and classified them as spam or not spam (ham). In addition, we determined whether the word "Viagra" appeared in each of them, and built a training set from that data.

Bayes theorem, the cornerstone on which this method is based, can help us calculate the probability that an email is a spam if it contains the word "Viagra."

Therefore, $P = (1 \times 0.25/0.50) = 0.50 \text{ or } 50\%$.

As calculated earlier, the probability that an email is a spam if it contains the word "Viagra" is 50%. Although this result isn't that interesting, because it's easy to calculate manually, you can try it with larger data sets by adding more terms to the data frame at the beginning of this example.

Strengths, weaknesses, and parameters of the Naive Bayes algorithm

The following list explains the details:

- Both MultinomialNB and GaussianNB use an alpha parameter to control the complexity of the models. In this case, the algorithm adds to the alpha value many data points of positive values from all features. This results in a smoothing effect on the statistics eventually, and a large alpha value gives a more smoothing effect, thus reducing the model complexity.
- GaussianNB is most applicable with large dimensional data sets, while the other two types are used specifically with small data sets.
- Actually, Naive Bayes classifiers have strengths mostly similar to those of linear regression models. They result in a fast mode of training and prediction with reduced complexity, and hence are very easy to understand. In fact, Naive Bayes classifiers form an important baseline model for very large data sets, which may not be applicable to most linear models.
- It is a fast, easy way to predict a class of data set for testing. It can also be applied in the prediction of a multi-class problem.
- When the assumption of independence is true, the Naive Bayes classifier works very well compared to the other models. Much less training data is needed.
- It performs better when the input variables are categorical rather than numerical. In the case of numerical variables, we need to have an assumption that there is a normal distribution.
- For a categorical variable with a category that hasn't been observed in the data set, the model assigns a zero probability, making it unable to generate a prediction. This is called zero frequency. The smoothing technique can be used to solve this problem. Laplace estimation can be employed to solve it.
- Naive Bayes classifier is referred to as a bad estimator, i.e., its outputs are not considered reliable.
- The Naive Bayes classifier also assumes the predictors are independent. In a real-life situation, it's nearly impossible for us to get independent predictors.

Applications of the Naive Bayes algorithm

When used for text classification, the Naive Bayes algorithm gives a high accuracy compared to the other types of models. It is also used in spam filtering to identify spam emails and sentiment analysis (for example, in social media to determine positive as well as negative customer comments).

Recommendation systems

The Naive Bayes classifier together with collaborative filtering can be used to build a recommendation system that uses data mining and machine learning techniques for filtering any unseen data. It can also generate a prediction on whether a user will like or reject a resource.

Normal distribution

Normal distribution is also known as Gaussian distribution. This distribution is based on the probability of real-valued events that occur from different problem domains. A continuous random variable that has a normal distribution is known as normal or normally distributed (for example, the heights of people, test scores, and marks of students).

Normal distribution is defined by using two main parameters that are mean (*mu*) and variance (*sigma*^2). Standard deviation is the average deviation from the mean and is denoted by sigma as well. Visual representation of normal distribution can be seen in the graph below:

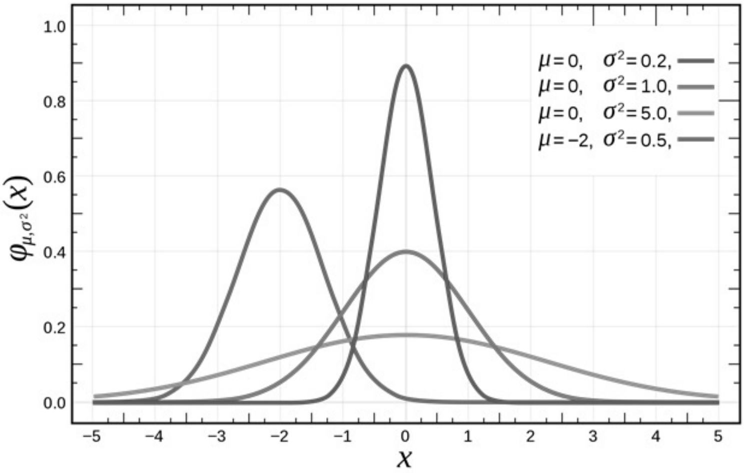


Figure 2.1

In this graph, we can see that Gaussian distribution or normal distribution can be identified from the distinctive bell-shaped curve.

This can be represented in the form of a mathematical formula as shown below:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu - x)^2}{2\sigma^2}}$$

Normal distribution is currently the most used probability distribution. It is used in various fields including marketing, psychology, finance, and economics. Because of its fine symmetry properties, normal distribution provides the probability of a variable being a given distance below the mean of distribution equal to the probability of it being the same distance above the mean. Owing to its reach of implementation and interpretation, normal distribution is widely used in the finance industry.

Some of the machine learning algorithms are based on distance and gradient descent measures such as k-means and k-nearest neighbor (KNN). These algorithms are quite sensitive to the scale of the provided numeric values, and for an algorithm to provide an exact solution, rescaling of the distribution is necessary. Rescaling usually mutates a range of values and can also affect variance. To perform rescaling, we can use statistical standardization (z-score normalization) and min-max transformation (normalization) techniques as well.

Python and R programming languages provide built-in functions for distribution transformation. Standardization can also be achieved by using the scale function (try help scale) in the R programming language. To perform the same operation in Python, we can consider the `scikit-learn` module preprocessing approach.

In machine learning, independent variables are also known as features. These are the input for a process that is being analyzed, whereas the dependent variables are the output of the process.

Automatic cluster detection in data mining

To find meaningful patterns in data, the data mining techniques described in this book are used. But sometimes there are no patterns or models to be found. Other times, the problem is not a shortage of patterns but an excess. The data can contain such complex structures that even the best techniques of data mining cannot coax out meaningful patterns. When mining such a database to answer some specific questions, competing explanations tend to cancel out each other. Too many competing signals add up to noise, as with radio reception.

Clustering provides a way to learn about the complex data structure, breaking up the cacophony of competing signals into its parts or components. When humans try to give meaning to complex questions, our natural or common tendency is to break the subject into small pieces, each of which can be explained more simply. If someone were asked to describe the color of trees in a forest, the answer would probably be different for winter, spring, summer, and fall. People may know enough about woodland flora to predict that the best factors to use to form clusters of trees that follow similar coloring rules are the hundreds of variables associated with a forest: season, foliage, etc., rather than height and age. Simple patterns can often be identified within each cluster after the proper clusters have been established. "Deciduous trees don't have leaves in winter, so the trees tend to be brown," or "Deciduous trees leaves change color in fall, usually to oranges, reds, and yellows." In many situations, a very noisy data set simply consists of a number of better-behaved clusters.

The question is, how can these find themselves? This is where automatic cluster detection strategies come in — to help see the forest without being lost in the woods.

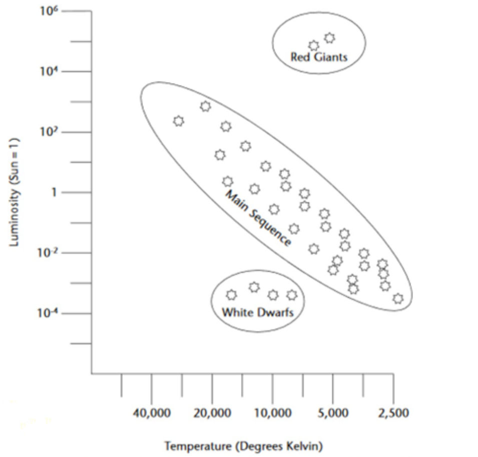
The geometric ideas used in k-means raise the more general question of similarity, relation, and distance measurements. These distance measurements are very sensitive to variations in the representation of data, so the next subject discussed is clustering data planning, with particular attention being paid to scaling and weighting. k-means is not the most widely used algorithm for automatic cluster detection.

Searching for simplicity islands

Data mining techniques are classified as direct or undirected, and automatic cluster detection is defined as a tool for the discovery of undirected data. That is valid in the technical sense since the automated cluster detection algorithms themselves simply find a structure that occurs in the data without regard to any unique target variable. Most data mining tasks begin with a pre-classified training set that is used to create a model that can score or classify previously unseen data. Clustering algorithms look for groups of records — clusters — composed of related records. These similarities are discovered by algorithms. It is up to the people who run the study to decide whether similar documents are of interest to the business, or something mysterious and maybe unimportant. However, clustering may be a focused practice in a wider context. Clusters developed for a business purpose are generally called segments in marketing, and clustering is a common method of customer segmentation. Automatic cluster detection is a technique of data mining that is rarely used in isolation, because cluster discovery is not always an end in itself. When clusters are observed, other approaches must be utilized to decide what the clusters represent. The results can be dramatic when clustering is successful. One famous early application of cluster detection led to our current understanding of stellar evolution.

Light of the moon, bright star

Astronomers attempting to explain the relationship between the luminosity (brightness) of stars and their temperatures in the early twentieth century rendered scatter plots like the one shown below. Two different astronomers, Denmark-based *Enjar Hertzsprung* and US-based *Norris Russell*, thought about doing this at around the same time. They both observed the stars dropping into three clusters in the resulting scatter map. This discovery led to further study and an understanding that, in very different phases of the stellar life cycle, these three clusters reflect stars. The following image illustrates the relationship between luminosity and temperature within each cluster:



The Hertzsprung-Russell diagram clusters stars by temperature and luminosity.

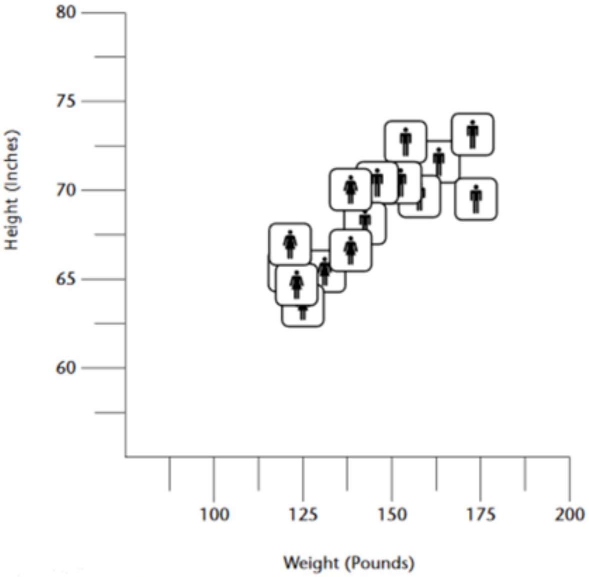
Figure 2.2

The relationship between luminosity and temperature within each cluster is similar, but the relationship between the clusters is different since the heat and light are produced by fundamentally different processes. By converting hydrogen to helium by nuclear fusion, 80% of stars that fall on the main sequence produce energy. This is how the majority of all stars spend their productive lives. The hydrogen is used up after several billions of years. The star then either starts to fuse helium, or the fusion ceases, depending on its mass. In the latter case, the star's core collapses, producing extreme heat in the process. At the same time, the outer gaseous layer extends away from the core, forming a red giant. The outer layer of gasses is gradually stripped away, and the remaining core begins to cool off. Now the star is a white dwarf.

A Google search for the term "Hertzsprung-Russell Diagram" returns thousands of pages of links to current astronomical research focused on this sort of cluster detection. Even today, HR-based clusters are used to search brown dwarfs (strong objects lacking sufficient mass to induce nuclear fusion) and to consider the pre-main sequence of stellar evolution.

A diagram of *Fitting the Troops The Hertzsprung-Russell* is a liable introductory example of clustering because it is simple to identify the clusters visually with only two variables (and, interestingly, it is a good example of the value of good visualizations of data). Also, in three dimensions, it is not too difficult to pick out clusters from a scatter plot cube by eye. If all of the problems had too few dimensions, automatic cluster detection algorithms would not be necessary. As the number of dimensions (independent variables) increases, cluster visualization becomes progressively difficult. Even our understanding of how close objects are to each other breaks down easily with more dimensions. To claim that there are many dimensions to a problem is an invitation to explore it geometrically.

Both the items that must be calculated separately in order to define something are one-dimensional. In other words, if there are N variables, imagine a space where each variable's value represents a distance in an N-dimensional space along the corresponding axis. For each of the N variables, a single record containing a value can be viewed as the vector that defines a particular point in that space. That is easily plotted because there are two dimensions. One such example was the HR diagram. The figure below is another example of how a group of adolescents maps their height and weight as points on a line. Note how boys and girls mix:



Heights and weights of a group of teenagers.

Figure 2.3

The above map provides a rough picture of the shapes of the people. But if the objective is to suit them for clothing, we need a few more measurements! In the 1990s, the US military conducted a study on how to redesign female soldiers' uniforms. The army's aim was to reduce the number of different uniform sizes (in terms of height) that must be kept in stock, while also supplying well-fitting uniforms to each soldier. As anyone who has ever shopped for women's apparel knows, there is already a multitude of classification structures (including measurements, odd measurements, plus sizes, junior, petite, etc.). None of these systems have been developed with US military's needs in mind. Researchers at Cornell University, *Susan Ashdown* and *Beatrix Paal*, went back to basics; they designed a new set of sizes based on the real shapes of women in the military.

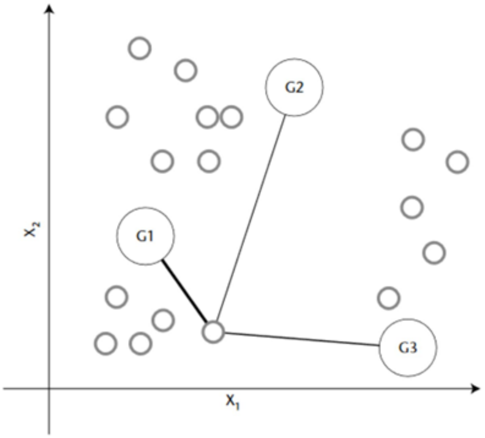
Unlike conventional clothing size structures, Ashdown and Paal did not come up with an organized set of graduated sizes where all sizes expand together. Rather, they have come up with sizes that suit various types of bodies. Each type of body corresponds to a cluster of records in a body measurement database. One cluster may consist of short-legged, small-waisted, big-busted women with long torsos, medium bodies, wide shoulders, and thin necks, while other clusters capture other measurement constellations. The database comprised over hundred measurements for each of almost three thousand women. The technique used for clustering was the k-means algorithm, defined in the next section. It took only a few of the more than hundred measurements to classify the clusters. A further advantage of the clustering method was discovering this reduced number of variables.

Gaussian model

As mentioned, the k-means method has some drawbacks:

- It doesn't do well with cluster overlaps.
- The outliers quickly drag the clusters off center.
- Each record is either within or outside of a given cluster.

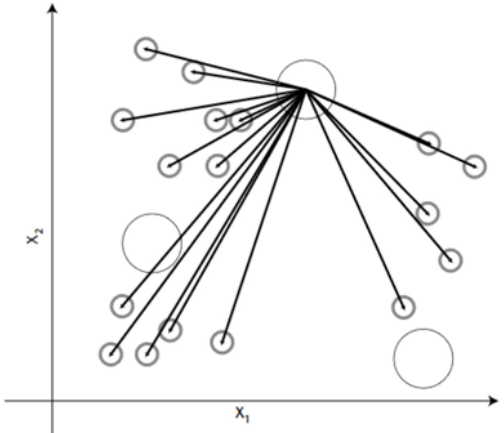
Models of Gaussian mixture are a probabilistic version of k-means. The name derives from Gaussian distribution, distribution of probabilities frequently presumed for problems of large dimensions. The Gaussian distribution generalizes to variable with the normal distribution. As before, k seeds were selected to start with the algorithm. However, this time the seeds are considered the source of Gaussian distributions. The algorithm goes by iterating over two steps, called the step of estimation and the step of maximization. The estimation stage calculates the liability for each data point kept by each Gaussian. Take a look at the following image:



In the estimation step, each Gaussian is assigned some responsibility for each point. Thicker lines indicate greater responsibility.

Figure 2.4

Every Gaussian has a strong responsibility for points close to their mean and a weak responsibility for distant points. In the next step, the roles will be used as weights. A new centroid for each cluster is determined in the maximization stage, taking into account the newly calculated responsibilities. The centroid for a given Gaussian, as shown in *Figure 2.5*, is determined by averaging all the points weighted for that Gaussian by the answerability:



Each Gaussian mean is moved to the centroid of all the data points weighted by its responsibilities for each point. Thicker arrows indicate higher weights.

Figure 2.5

These steps are repeated until the Gaussians do not move anymore. The Gaussians themselves can both change form and travel. Every Gaussian is constrained, however, so if it shows a very high responsibility for points close to its average, it means a sharp drop in responsibilities. If the Gaussian encompasses a broader spectrum of values, then it has smaller obligations for points nearby. Given that the distribution must always be integrated into one, Gaussians are always getting weaker as they grow larger. The reason this is called a mixture model is that the likelihood is the sum of a mixture of many distributions at each data point. Every point at the end of the process is bound to the different clusters with higher or lower probability. This is often called soft clustering since a single cluster does not classify points uniquely. One indication of this approach is that certain points in more than one cluster can have a high probability. Other points may only have very low odds of being in any cluster. Each point can be assigned to the cluster where it has the highest probability, making this soft clustering a hard clustering.

Automatic cluster detection technique is the undirected data mining tool, which can be used to learn about complex database structures. By splitting complex data sets into simpler clusters, it is possible to use automated clustering to boost the efficiency of more focused techniques.

Automatic clustering can be applied to almost any form of data by choosing different distance measurements. Clusters can be found just as easily in news story collections or insurance reports, as in astronomical or financial details. Clustering algorithms can rely on some form of similarity metric to determine whether two records are near or distant. A geometric representation of distance is sometimes used, but there are other possibilities, some of which are more acceptable when non-numeric data in the records is to be clustered. One of the most common automatic cluster detection algorithms is the k-means.

The k-means algorithm is an iterative approach to remote-based identification of k clusters. Gaussian mixture models are a variation on the concept of k-means that allows clusters to overlap. Divisive clustering creates a tree of clusters by splitting an initial large cluster in succession. The agglomerative clustering begins with several small clusters, and combines them progressively until there is just one cluster remaining. Divisive and agglomerative approaches allow the data miner to use external parameters to assess which cluster tree level is most useful for a specific application.

Application of machine learning in cybersecurity

Machine learning techniques can be applied to a variety of cybersecurity problems:

- Spam mail and phishing page detection.
- Malware detection and identification.
- DoS and DDoS attack detection.
- Anomaly detection.
- Biometric recognition.
- User identification and authentication.
- Detection of identity theft.
- Social media analytics.

- Detection of information leakage.
- Detection of advanced persistent threats.
- Detection of hidden channels.
- Detection of software vulnerabilities.

In the following sections, we will briefly discuss some of these cybersecurity problems and solution approaches with machine learning techniques.

Spam detection

Spam is the electronic equivalent of the junk mail that arrives in our mailbox. Spam may be defined as an unsolicited commercial email or any unsolicited bulk email. Spam mails are not only just annoying but can be dangerous. The content of spam can be an advertisement of goods and services, pornographic material, financial advertisements, information on illegal copies of software, fraudulent advertisements, fraudulent messages to solicit money, links to malware, phishing websites, etc. Spam may also be used for launching denial-of-service (DoS) attacks.

Spam filtering can be performed on the basis of the textual content of email messages. This can be seen as a special case of text categorization, with the categories being spam and non-spam. Text classification techniques such as TF-IDF, Naive Bayes, SVM, n-gram, boosting, etc., can be applied for spam filtering. Some of the limitations of these techniques are the requirement of a large amount of training data, processing power, etc.

Other machine learning techniques involve converting emails into feature vectors, where features include token of the email, size, presence of attachment, IP, and the number of recipients. We can use machine learning techniques such as CVM, decision trees, neural networks, and so on.

Phishing page detection

With phishing, people are psychologically influenced and the contents of the malicious website are made believable, thereby making users enter personal details such as username, password, bank account number, credit card details, and so on. Generally, this is achieved by spoofing the websites of reputed organizations, so that the user never doubts an illegitimate activity prior to entering their personal information. These days, phishing sites are not only used for sniffing user credentials, but also for spreading malware such as cookie stealers, keyloggers, etc. Such malware can steal cookies from the system or capture user keystrokes.

The different kinds of features used in existing machine-learning-based detection algorithms can be grouped as follows:

- URL-based features.
- Domain-based features.
- Page-based features.
- Content-based features.

Some of the URL-based features include the number of digits in the URL, total length of URL, number of sub-domains in URL, misspelled domain names, TLD used, and so on. Some of the domain-based features include the presence of domain name or its IP address in well-known blacklisted services, age of the domain, category of the domain, availability of the registrant name, and so on. Some of the page-based features include global and country PageRank, position in the list of Alexa top 1 million sites, estimated number of visits for the domain on a daily, weekly, or monthly basis, average page views per visit, average visit duration, web traffic shape per country, number of references from social networks, presence of similar websites, and so on. Some content-based features include page titles, meta-tags, hidden text, text in the body, images, videos, and so on.

Thus, the common features that are used for phishing page detection include its URL, PageRank, SSL certificate, Google indexing, domain, and features based on the web page source code. However, these features have limitations. Web pages that are hosted on free web hosting services would never include some of the relevant features possessed by a legitimate web page such as PageRank, age of the domain, SSL certificate, and Google indexing. On the other hand, it is possible for a hacker to promote his phishing page by utilizing Black Hat SEO techniques such as keyword stuffing to get higher indices for the phishing page. Further, incoming traffic to the website can be increased with the help of third-party tools such as the web clicker application, which generates thousands of page views within minutes.

Phishing page detection is a supervised classification problem. We need labeled data that has samples of phishing pages and legitimate pages in the training phase. The training data set that we use is a very important component in building a successful detection mechanism. So, we have to use training samples whose classes are precisely known. Phish Tank is a well-known public data set of phishing websites. Site reputation services can be used for collecting legitimate sites. Machine learning techniques such as SVM decision tree and Naive Bayes have been used for phishing page detection.

Malware detection

Malware (malicious software) is any software intentionally designed to cause damage to computer systems and networks. Malware can be divided into the following classes:

- **Virus:** It is a type of malware that is loaded and launched without the permission of the user and reproduces itself as well as infects other software.
- **Worm:** This type of malware is very similar to a virus, but unlike a virus, it can spread over the network and replicate to other machines on its own.
- **Adware:** This type of malware displays advertisements on the computer.
- **Spyware:** This type of malware performs espionage.
- **Rootkit:** This type of malware enables the attacker to access the user's system and network with root access.
- **Backdoor:** This type of malware provides an additional secret entry point to the system for the attacker.
- **Keylogger:** This type of malware logs all the keys pressed by the user and captures all sensitive data including passwords, credit card numbers, and so on.

Existing malware detection mechanisms are classified as static, dynamic, or hybrid (combination of static and dynamic). The static analysis uses features of the source code of the software such as signatures, API calls, function calls, and permissions for detecting its malicious nature without executing the software. On the other hand, dynamic analysis is conducted on the software while it is being executed ideally in a virtual environment. Malware can easily bypass static detection mechanisms using encrypted or obfuscated source code or bypass the permissions. Most of the existing malware detection mechanisms use machine learning algorithms where permissions, API calls, system calls, system call frequencies, densities, co-occurrence matric, and Markov chain state transition probability matrix are used as features. The static analysis includes various techniques:

- String extraction for the examination of the output.
- Scanning with anti-malware scanners.
- Disassembling, i.e., reversing the machine code to assembly language and interfering with the software logic and intentions.

In dynamic analysis, the behavior of the software is monitored while it is executed in a virtual environment (sandbox), and the properties and intentions of the software are inferred from that data.

Machine learning techniques can be applied in both malware detection and classification of malware into families. Signature-based malware detectors can perform well with known signatures. However, they may not detect polymorphic malware that has the ability to change signatures, as well as new malware with different signatures. In malware classification, unknown malware types are clustered into several groups on the basis of certain features identified by the machine learning algorithm.

Behavior-based malware detection in android systems based on crowdsourcing collects the system calls from the users and sends it to a centralized server for analysis. The server will preprocess the data by creating the system call vector for each interaction of the users within their applications.

Statistical mining technique is another approach for detecting malware applications in smartphones. In this case, features such as CPU usage, memory usage, network traffic, and battery usage are collected every five s and then stored in a database.

First, a data set with permissions in well-known goodware and malware applications are created. Then, feature selection algorithms are used to select relevant features from this data set. Then, the k-means clustering algorithm is used to distinguish the malware from the goodware applications. Finally, it classifies the malware in the cluster using machine learning algorithms such as J48, random forest, into various types such as Trojan and data stealer.

Hidden Markov model (HMM) can also be employed for malware detection. This method is based on the observation that every smartphone application involves a series of interactions between the user and the device. Thus, the system call is correlated with user operations and can be used to detect malicious activities. It uses process state (set of system calls invoked by the application) transitions and user operational patterns. The application is classified as malware if the likelihood is less than a threshold.

HMM can also be built with system call sequence as observations and the keypress sequence as hidden states. The keypress and the system call sequence generated by the application are collected, and the hidden keypress sequence is obtained using the Viterbi algorithm.

Maline is a tool used to detect android malware applications using system call dependencies. A feature data set is built from several malware and goodware applications. It then uses a machine learning algorithm to determine whether the given application is malware depending on these features. One limitation of this approach is that it needs a large training data set.

DoS and DDoS attack detection

A denial-of-service (DoS) or distributed denial-of-service (DDoS) attack is a malicious attempt to disrupt the normal traffic of a targeted server, service, or network by overwhelming the target or the infrastructure with a flood of the Internet traffic. A DoS attack exploits a victim machine or network by exhausting its bandwidth, memory, and processing capacity. A DoS attack targets the victim by generating numerous malicious packets to prevent legitimate users from using different services. A DDoS attack is a DoS attack involving more than one computer to target a victim in a coordinated manner.

The detection of a DoS attack is mainly carried out in one of the following two ways of intrusion detection:

- Signature-based intrusion detection.
- Anomaly-based intrusion detection.

Signature-based intrusion detection works by comparing the network with known network attack patterns stored in the database. Although this method has fewer false positive alarms, it may not detect zero-day attacks. Anomaly-based intrusion detection uses the statistics of network traffic such as packet header information, packet size, and packet rates to detect various intrusions in the network. Although anomaly-based intrusion detection mechanisms do not require databases of anomalies to compare (resulting in lesser memory requirement and maintenance), it has difficulty in detecting malicious traffic that is similar to normal traffic, such as low-rate DDoS attacks, and has no guarantee of detecting unknown attacks.

Some of the features used by machine learning models are entropy, number of data bytes transferred from source to destination, number of connections to a host, source, and destination IP addresses, byte rate, packet rate, TCP flag ratio, SYN packet statistics and flow statistics, SYN flag presence, classification fields and protocol fields, TCP SYN occurrence, destination port entropy, entropy of source port, UDP protocol occurrence and packer volume, and so on. Machine learning techniques such as SVM, KNN, Naive Bayes, and k-means have been used for DoS/DDoS detection.

Anomaly detection

Anomalies may not correspond to cyberattacks all the time but can be new behavior. Anomalies or abnormal or unexpected behavior can be detected and categorized as discussed below.

In k-means clustering, the objects are grouped into k disjoint clusters on the basis of the feature vector of the objects to be grouped. A network data mining (NDM) approach along with the k-means clustering algorithm can be used to separate time intervals with normal anomalous traffic in a training data set.

In k-medoids, each cluster is represented by the most appropriate center object (medoid) in the cluster, rather than by the centroid that may not belong to the cluster. k-medoids can be used to detect network anomalies that contain unknown intrusions, and is found to have more accuracy than the k-means algorithm.

The new mean of the cluster is computed on the basis of weight measures. EM-based anomaly detection mechanisms are found to outperform both k-mean- and k-medoids-based methods.

Unsupervised and supervised learning algorithms can be combined to form hybrid approaches for anomaly detection. In this manner, the efficiency and accuracy of anomaly detection can be improved. A combination of k-means and ID3 decision tree is known for the classification of anomalous and normal activities in computer address resolution protocol (ARP) traffic.

Some machine learning approaches for intrusion detection also use artificial neural networks and support vector machines. A hybrid approach by combining the entropy of network features and SVM can overcome the drawbacks of individual approaches, resulting in anomaly detection with high accuracy.

Thus, different machine-learning -based anomaly detection approaches exist. A model is built on the basis of normal and abnormal behavior of the system. Different approaches can be adopted depending on the type of anomaly. When the model for the system is available, it is tested with the observed traffic. If the deviation is found to be more/less than a predefined threshold, then the traffic is identified as anomalous.

Conclusion

In this chapter, we discussed the technical algorithmic aspects of Naive Bayes, normal distribution, and automatic cluster detection with the Gaussian process. We also discussed machine learning techniques that can be applied to a variety of cybersecurity problems such as spam detection, phishing page detection, malware detection, and anomaly detection.

In the next chapter, we will be discussing how to structure the data.

Chapter 3
Machine Learning for Data Structuring

Introduction

Text mining is one of the most complex technologies in the machine learning field. This is because text mining involves unstructured data. There are no well-defined observations and variables. Therefore, in order to perform some sort of analytics, you must first transform this unstructured data set into a structured data set, and then continue using the usual modeling framework. What is the difference between structured and unstructured data? What will be the future of big data and data science? You will find the answers to these questions in this chapter, so let's get started.

Structure

In this chapter, we will cover the following topics:

- Data structuring
- Future of big data
- Structured and unstructured data

Objective

After studying this chapter, you should be able to do the following:

- Understand the difference between structured and unstructured data.
- Understand the future of big data.

Data structuring

There are many different forms of data, but at its highest level, data is mainly categorized in three ways:

- **Structured:** This type of data is incredibly organized. It is found in databases, CSV files (as values separated by commas), or other repositories. The data format ensures it is right for computation and inquiries, using SQL, or other structured languages.
- **Semi-structured:** This type of data doesn't follow the way data models are structured when associated with data tables like relational databases, but does have markers and tags that keep the semantic elements apart and enforce hierarchy in the data fields and records.
- **Unstructured:** This type of data has no real structure, such as natural language text or audio streams.

Clearly, the most useful data is structured data because it is ready for immediate manipulation. As a rule, only around 20% of the total data is represented by structured data, while the remaining 80% is a combination of unstructured and semi-structured data.

Beware, though; most of what we call unstructured data does have some kind of structure. Take an article for publishing on the Internet, for example. While it is classified as unstructured, it does have a structure of sorts by way of tags and metadata for the article content. It is classified as unstructured because the content itself has no real structure that is usable straightaway.

The future of big data

We are in the midst of a seismic shift in the data science world. This realization is beginning to manifest itself in the marketplace, with the newer analytics tools focusing more on deriving insight from data than running reports and dashboards. Which is, of course, the correct way to look at this problem or any problem—focus on the end goal. Don't focus on the "how"; keep your eye on the prize.

However, the tools available today are primarily old-school business intelligence tools that are being retrofitted as fast as possible to try to ride this wave. Whether or not they will actually succeed or a new generation of tools will need to emerge is still an open question.

With that in mind, here are five developmental directions to watch:

- **Your data will tell you what is interesting:** Current business intelligence does exactly the opposite; you tell your tool what you're interested in. The coming generation will tell you what's interesting.

  Generating these types of insights will become much easier in the coming years, and personalization will have a lot to do with this. Think about how revolutionary Pandora was for music when you first tried it—it actually learns what you find interesting! This is the direction analytics is heading in.

  The systems themselves will also get better at automatic learning and pattern recognition, about what you want to see. This will enable them to spot outliers and changes in trends and other interesting data. Today, they must be spotted by a human being.

- **Wild visualizations:** Charts and graphs have been around for a long time. And there's a reason for that: they're good at conveying meaning in an easy-to-grasp way.

  But again, they were developed as a way to understand much smaller amounts of data than we're looking at now. New ways of looking at data—and especially interacting with it—are being developed. Right now, we're seeing the first wave of this, where designers are trying out different techniques, and some work while some don't. Things like infographics, interactive web pages, and some of the newer intelligence tools are examples of trying new techniques of looking at data.

  We expect to see this trend accelerate. And as the field becomes more popular and more vital to the world economy, the brains and time spent on this problem will lead to breakthroughs in new ways to look at and understand data.

- **Self-service intelligence:** The ability to use analytics will no longer be a specialized skill, and we already see it happening in the case of data discovery tools. It will be a common and everyday thing for executives down the line when they use it as part of their daily and hourly workflow. We've seen how this is playing out through IBM's Watson, and it will be especially pervasive in the knowledge industry.

  This is slightly counterintuitive; the amount of data being used is growing by the day, and yet the tools will be simpler? Yes, that's right.

  For example, the days of calling for a vital KPI and having to wait ages, if not days, for the IT department to dig out the report are gone. Self-service intelligence places the details in the control of the customer.

  Data can be viewed on the fly for real-time analyses and instantly actionable perspectives, offering the staff a strategic edge.

  As the amount of data increases, it is forcing a complete reevaluation of the current intelligence tools. The complex user interfaces with miles of toolbars and byzantine menu systems are no longer adequate—they were built for a world with much less data.

  Which brings us to the new types of interfaces that are needed: Sisense, Zoho Reports, Microsoft BI, Tableau Desktop, Tibco Spotfire, Alteryx Analytics, and Salesforce Einstein Analytics.

- **Natural and intuitive interaction with data:** Business intelligence is one of the last bastions of enterprise-y software simply because of its complexity. You know you're using business intelligence software if half the screen is taken up by menus and a passerby would have no idea what you're looking at.

  Natural interfaces such as touch, voice, and gestures abstract away the complexity. The more intuitive a functionality is, the more productive a user can be because they won't have to learn how to use it. When you can literally ask the question "What demographic should I focus on to increase sales?" you can find valuable information without training or expert help.

Apple did the world a great service by training people to use touch and coming up with intuitive—and, now, universally accepted—touch gestures. Apple's Siri is important because it is training people to use voice. Microsoft's Kinect is important because it's training people to use body gestures. All of these modes of input combined create an interactive environment that lets you play with your data—explore it and interact with it.

You will think with your eyes and follow your nose to the data you're looking for, even if you're not sure of the right question to ask.

Data will be fun.

- **Collaborative:** A few years ago, you couldn't escape "collaboration" as a buzzword. The idea was that, by enabling people to work together, you freed them to pool their creative and mental resources, hopefully, of course, resulting in a synergistic explosion of ideas and good decisions.

  The collaboration was, in a sense, a way to deal with a lot of data—big data—before big data got really big. It was crowdsourced report hunting, basically. Unfortunately, the amount of data that needs to be evaluated is simply too big even for large numbers of people to efficiently looking at.

  The new types of interfaces (as explained above) will alleviate that problem. But there was a core value in collaboration that doesn't go away just because it's no longer "the new black." Two heads are still better than one if they're in the right place at the right time.

  When tools allow people to use intuitive interfaces to not only explore their data and discover from it, but also to allow more than one person to do this at the same time, the results will be nothing short of magical. Imagine being able to look at data in interesting ways, exploring it, and sharing what you're looking at in real time with colleagues from around the world: igniting conversations. Enabling people to work on the same thing but giving them the freedom to explore independently gives you the best of both worlds.

  But when?

  Most of these changes will happen within the next three to five years. Some of them are happening already. But all signs point to a type of inevitability for them—they are at the intersection of two or larger trends driving our society towards a new type of information consumption, one that's very different from what we have today.

Structured and unstructured data

Data contained in documents, databases, data warehouses, emails, and other data files can be categorized either as unstructured or structured data. Structured data is organized data. The data always follows a consistent order and is always easy to search. This data can be accessed easily and understood by any person or machine.

A classic example of structured data is that found in an Excel sheet. Each data set has labeled columns, and the data is always consistent. The column headers provide a brief description of the data in the columns. Using this information, you can determine the type of data in the columns. If you have a column in the data set called Images, you know there will only be images in that column. This consistency always ensures that structured data is amenable to automated data management.

Structured data is often stored in databases and data warehouses in well-defined schemas. The data is often stored in a tabular form with columns and rows that define the exact attributes of the data. Unstructured data is of free form. It is often dispersed and never found in a tabular form. This type of data is often not easily retrievable, and such data takes a lot of time to be understood. Numerous emails, web pages, documents, and files all in scattered locations are classic examples of unstructured data.

It is impossible to categorize such data since it is often created in an unstructured manner. This makes it difficult to understand the different attributes of the data, which is why it takes a lot of time to group this data. The content obtained from unstructured data is often hard to work with or even make sense of using programs. Most computer programs find it difficult to analyze or generate reports on such data without being able to process it since the data lacks structure. Newer technologies are being developed that make sense of unstructured data.

There is indeed more unstructured data in the world when compared to structured data. Since it takes more work to make unstructured data usable, it tends to grasp the most attention and takes more time. It is no wonder that the promise of a machine or a program being able to process and analyze unstructured data is a huge selling point for any predictive analytics model. You must remember never to underestimate unstructured data and how powerful it is to your analysis. It is always more efficient to analyze structured data as opposed to analyzing unstructured data since the latter is costlier to process when you are building a predictive analytics model. The selection of the data, its relevance to your analysis, cleansing, and the subsequent transformations are often lengthy and tedious tasks. It is only after the data has been through all these processes that it becomes worth using in a predictive analytics model.

It is often easier to get results using structured data as opposed to unstructured data since you will never have to worry about there being a delay in the scrubbing process. Tagging documents and text analytics are two ways to structure unstructured text documents, grouping, summarizing the data, and linking their content. This will also help the user uncover any hidden patterns in the data. It is also important to note that most search engine platforms often provide tools to index data and make it searchable.

Unstructured data does not often lack structure; it just takes a little more effort to understand the structure. Every text in any digital file has some structure associated with it, and it often shows up in the metadata. This applies to every other form of structured data. It may be worthwhile to conduct a thorough analysis of the structural components of the data. This will allow you to estimate the potential value of any analysis conducted on unstructured data. When the analysis provides a little insight into the data, you should try to determine the resources you will need to allocate for analyzing that portion of unstructured data.

The idea is that you can always find some order in the data you have collected. You will need to do a lot of digging into the data to understand it better. The content in a bunch of emails between two people may stray away from the original subject of the emails, although the subject line stays the same. There are times when the subject line may not even make sense.

The line that separates the two types of data is not always clear since there are some aspects of unstructured data that are those of structured data. Whether or not that structure reflects the content of the data is unclear on most occasions. For that matter, structured data can also hold some unstructured data within it. For instance, in a web form, users may have to give feedback by choosing answers from multiple-choice questions and provided with a comment box where they can provide feedback in text form. This comment box contains unstructured data because the information is in free form. These cases are often considered a mixture of both structured and unstructured data.

There will always be some exceptions in the field of defining data since the definitions of the two types of data could be blurry. There is always a way to make the distinction between the data.

Conclusion

In this chapter, we discussed the difference between structure and unstructured data. The current advanced tools help businesses make quicker decisions. We also discussed how machine learning tools help businesses put data in order so that it can be processed properly and in a more orderly manner. In the next chapter, we will discuss parsing.

Chapter 4
Parsing Data Using NLP

Introduction

The development of natural language processing (NLP) is based on the theories and models of artificial intelligence. NLP is the ability of a machine learning model to understand human language and is a major part of artificial intelligence. This approach has helped companies and businesses to communicate with people from all over the world in different languages. Human speech is not always precise, and every single one of us has a unique linguistic structure. The process of translation is dependent on several major factors like regional dialects, variables, and social context.

Moreover, NLP is a way for computers to understand, analyze, and derive accurate meaning from human language in a useful and smart way. Machine learning and artificial intelligence models are designed to perform multiple tasks such as translation, relationship extraction, automatic summarization, topic segmentation, and sentiment analysis. NLP is an effective approach to analyze text and allows machines to understand how actually humans speak. This approach is commonly used for machine translation, text mining, and automated question answering.

Parsing means to extract information and meaning from the text data in line with laws of grammar through NLP. In other words, parsing is to break down a sentence in order to clarify each aspect of the sentence.

Structure

In this chapter we will cover the following topics:

- Uses of NLP.
- Key advantages of NLP.
- Data handling in NLP.
- NLP applications.

Objective

After studying this chapter, you should be able to do the following:

- Understand the use and advantages of NLP.
- Understand the applications of NLP.

Uses of NLP

Natural language processing (NLP) algorithms are used for a variety of purposes. Basically, they help developers develop software that understands human language. Owing to its complexity, NLP is difficult to implement and learn correctly. With the passage of time, data scientists and machine learning engineers are working to improve this process and bring a better turnover rate.

NLP algorithms are used to summarize blocks of text. To complete this operation, a summarizer is implemented into the machine learning model that extracts useful ideas and ignores any irrelevant information. Further, the NLP models accurately generate keyword tags from the content through AutoTag, which is an essential technique to discover the topics briefed in the content. Furthermore, the approach of sentiment analysis identifies the sentiment of a string from the given text in a positive manner.

The current approaches to NLP are dependent on deep learning concepts. Deep learning is a type of AI that examines and utilizes patterns in data to improve the understanding abilities of a program. The earlier NLP models were based on the rule-based approach, which only had the capability to search for a phrase in the text and generate simple responses. Today, AI and machine learning have greatly customized how NLP models work and deliver outstanding results. Tools for NLP models are Genism, Intel NLP Architect, and NLTK.

Python programming language comes with built-in libraries that help developers to create high-end machine learning and AI models. These libraries include SciPy, NumPy, and Pandas. Research is being done by data scientists to improve search results so that users can query data sets through a question and get accurate translation in real time. Machine learning models have the capability to understand the vital elements of the human language and its sentence structure.

Key advantages of NLP

NLP models provide unlimited advantages. The models provide the capability to improve the efficiency and accuracy of a document, along with the opportunity to create a readable text summary. Moreover, this approach is the main driving force behind personal assistants like Alexa and Siri, through which users can perform multiple tasks with ease. NLP provides long-term benefits to companies as they can communicate with their clients effectively and solve queries in real time through chatbots.

Conferring to industry figures, about 21% of the accessible data is in a structured form. Data is created as we express, as we tweet, as we send messages to WhatsApp, and other activities. Most of this data is in a word-based form, which is extremely unstructured in nature. NLP helps in extracting valuable data from such a format of data.

Few noteworthy examples include social media tweets and posts, user-to-user messages, articles, blogs, news, product or service reports, and patient information in the health-care field. Others include chatbots and other voice-driven bots.

Data handling in NLP

NLP systems are designed to understand conversations of different styles and languages. Data that is generated from declarations or conversations are a part of unstructured data. As compared to structured data, machine learning models find it difficult to comprehend unstructured data because there are no predefined relationships and labels. NLP is a field of AI that enables automated machines to learn, understand, and derive accurate meaning from human languages. This is only possible if the models are trained with labeled data so that they can easily comprehend multiple languages with a high accuracy rate.

NLP applications

NLP is one of the hottest subjects in the world of computer science. Companies in finance, supply chain, marketing, and logistics domain are pouring lots of capital into this area of science. All are seeking to understand NLP and its implications for a career around it. Every company out there seems to want to somehow incorporate it into their company.

Talent recruitment

Human resource departments in various companies use NLP models to review resumes and conduct candidate interviews. In the search and selection process of talent recruitment, interviewers are able to identify the skills and abilities of potential hires and also review the latest trends in the job market.

Voice assistants

NLP tools such as Siri and Alexa work with AI models to respond to vocal prompts. Particularly, NLP models can do anything from finding a specific shop to scheduling an appointment for a given date. Voice assistants learn through structured data and need a lot of training data to develop an understanding of each language.

Health care

Health-care and medical departments have had great breakthroughs thanks to the artificial intelligence and machine learning technologies. With the help of NLP tools, doctors and health specialists can explore more health conditions.

Survey analysis

Surveys are an effective way to measure the success of an organization. Companies are running a variety of surveys to get inputs from consumers on different items. NLP can be very helpful in identifying the weaknesses and encouraging businesses to develop their goods.

Grammar checkers

This is one amongst the most widely used applications of linguistic communication processing. Grammar checking tools like Grammarly provide plenty of features that help people write better content. NLP will change any ordinary piece of text into beautiful literature. If you would like to jot down an email to your boss, or if you're writing a report, there's no denying the fact that you simply need these helpful friends.

Email filtering

I am sure you have already noticed that every email you get gets classified into spam and inbox. Isn't it amazing?

Emails are sorted using a NLP technique. Yes, as you may have guessed, text Classification is the method of classifying a piece of text into predefined groups.

Another excellent example of text sorting is the grouping of news stories into separate groups.

Social media monitoring

Today, NLP is being used by businesses to evaluate social media content and to know what consumers think of their goods. Companies often use social media analysis to consider the challenges and difficulties that their consumers face when purchasing their goods. Not only businesses but also governments use it to detect possible threats to the security of the country.

Online search autocomplete and autocorrect

NLP is something everybody uses every day, but they never pay any attention to it. Autocomplete and autocorrect both allow us to find reliable answers very quickly, and they are a perfect example of how NLP impacts millions of people around the world, including you and me. Several other businesses have since begun using this feature on their websites, such as Facebook.

Conclusion

In this chapter, we discussed NLP as a vital tool for converting or translating large amount of information into data that humans can comprehend. We also discussed NLP application in recruitment, voice assistance, health care, survey analysis, grammar check, email filtering, social media monitoring, and online search.

In the next chapter, we will be learning computer vision.

CHAPTER 5 Computer Vision

Introduction

Computer vision is one of the most modern fields of study in data science. It has also been part of our lives. We all use a range of features that have machine vision strategies working on the backend. For example, we use it for face unlock on our smartphones.

Computer vision is a division of artificial intelligence that aims to emulate the strong capabilities of human vision. Computer vision operates through visual recognition strategies such as image processing, optical character recognition, image captioning, object tracking, image segmentation, and object identification.

Structure

In this chapter, we will cover the following topics:

- Computer vision application.
- Neural networks in computer vision.
- Overview of computer vision.

- Image recognition.
- Biometric recognition.
- Software vulnerabilities.

Objective

After studying this chapter, you should be able to do the following:

- Understand the application of computer vision.
- Understand the architecture of convolutional neural networks (CNNS) used for computer vision.

Computer vision application

We already know that human vision sense is incredibly advanced. As soon as you open your eyes, you see every detail and object around you within a fraction of a second. We can instantly detect and identify objects in our field of view unconsciously without any hesitation or previous thought.

We are also able to name every object that is around us, as well as perceive the depth of these objects with great accuracy, and we can perfectly distinguish objects' contours and detect and identify various objects in the background. Our eyes take in some raw voxels of various color data, and our brain transforms this data into some relevant and meaningful information. So, color information will be perceived as different lines, shapes, curves, etc., which may indicate what we are looking at, like objects, people, building, etc.

Programming machines for computer vision to replace human in various fields such as surveillance, immigration check, candidate screening etc. Therefore, many major companies are used deep learning techniques and CNNS for enhanced performance in facial recognition and object detection, which are part of the computer vision field. Some social network platforms also use this technique for automatic facial recognition, like Facebook.

These techniques are also incorporated into self-driving cars that are the future. However, we already know the challenges in computer vision, and they might be very difficult to solve. Humans automatically see objects, contours, lines, and shapes, but computers, on the other hand, just see a great collection of matrices and numbers.

To solve this problem, and to learn more complex image features from raw pixels and their values, CNNs are used. Moreover, a great place to start working on computer image is the programming language Python, providing you everything you will need on this journey.

Neural networks in computer vision

Activation function

Let's learn some essential concepts such as ReLU activation functions, data preprocessing, weight initialization, and introduction to CNNs in CNN architecture.

ReLU is the most widely used activation function for CNNs. It returns 0 if the input is negative. It is computationally very effective as easy clipping is required. Empirically, it is found to be faster than sigmoid or tanh.

Data preprocessing

All images are transformed into similar sized square shapes. The mean value of each image can be deduced from each pixel.

Weight initialization

The weights can be adjusted to the Gaussian distribution of 0 mean and standard deviation (0.1 to 1e-5). This works for shallow networks, for example, networks with five hidden layers. In the case of deep neural networks, small weights make the outputs small, and as you step in towards the end, the values are much lower. As a result, the gradients also become thin, resulting in gradient-killing in the end.

Introduction to CNN

The first layer will attempt to distinguish edges and create templates for the detection of edges. Subsequent layers may then attempt to merge them into simpler forms and finally into models of various object positions and scales. The final layers align the input picture of all the patterns, and the final forecast is like a weighted sum of all the templates. So, deep CNNs are capable of modeling complex variations and behavior that offer very precise forecasts. A CNN contains three types of layers: convolution layer, pooling layer, and fully connected layer.

Convolutional layers

They are the main structural blocks used in CNN architecture.

Pooling layers

They are used to pad the convolution layer so that the image size remains the same. So, pooling layers are used to decrease the magnitude of the image. They work through each layer by means of filters.

Fully connected layers

Networks typically use fully connected layers in which each pixel is known to be a separate neuron, much like a normal neural network. A fully connected layer will include as many neurons as the number of groups to be forecast.

Overview of computer vision

Python is one of the most widely used programming languages, and it will be the perfect destination for CNNS and tasks for computer vision. CNNs are behind some great successes in the field of computer vision and image and speech recognition. We already know that CNNs use two fundamental structures: feature maps and filters, which are also called feature detectors.

These fundamental structures can be presented in the form of certain groups containing various neurons that will enable us to build different networks. So, for instance, we can say that we have one image, and our main goal is to detect both vertical and horizontal edges. To achieve this, we have to create an individual convolutional filter.

A convolutional filter is, in fact, a very small matrix that represents a certain feature that is relevant to us and that we want to find within the original image. The convolutional filter that is placed on top, in fact, tries to identify and detect various parts of that original image with some vertical lines. On the other hand, the convolutional filter that is placed at the bottom tries to detect and identify various parts of the image that contain some horizontal lines.

The actual process of detection will work by taking the convolution of a certain filter with the original image. When we perform the convolution in Python, outputs of that convolution will locate various positions of different features within the original image, which will represent our features maps. To convert the feature map filters into some concrete models in a CNN, we will use a hierarchy scheme.

In this scheme, a feed-forward neural network and all its layers containing neurons will represent both the feature map and the original image. In contrast, the convolutional filters represent a specific combination of different neural network relationships. These relationships are replicated throughout the complete input.

You have to initialize every relationship in that group with identical weights, and you should always average the certain weight updates of that collection before applying them at the end of every iteration.

To do more meaningful tasks, you can also create a feature map at some layer that will detect various objects or faces. For instance, if you have faces showing nose, mouth, and two eyes, you will, in fact, need the data from all of these three different feature maps that will be represented at the CNN layers.

In other words, you can create filters that depend on the volume of data, and they still might traverse in various feature maps within a single layer. These kinds of relationships can also be captured if you are using a full-fledged neural network, which would do just that.

When you are moving deeper into your convolutional network, you may need to sharpen the data contained within your feature map. This is mainly because when a feature is presented in a previous layer, it may result in the feature map containing a hotspot surrounded by a halo. Therefore, the feature map often may be weakly or not completely detected, so you might need to sharpen the data included in the layer.

To achieve this, you will use a technique called max pooling. Using this technique, you will divide your feature map into disjoint squares or titles, and then, you will take an action that is a maximum of all neurons that are situated at each tile. Max pooling for the feature maps is great when it comes to the detection process that will be clearer by removing some irrelevant and uninformative halos.

Max pooling also reduces the total number of parameters in a CNN. Max pooling is also used for combating some potential issues in overfitting. If you put all these concepts together, you are ready to start working on some more complex computer vision problems. For instance, let's imagine that we have a blood smear from a patient, and our main goal is to detect the potential for malaria invasion, as well as to diagnose what stage of infection is present in that patient. The stage of infection may range from early to late stage.

There is also the possibility that there is no infection present in the patient. You can create CNNs that will contain four layers; this would be followed by classic dense feed-forward networks that end in a three-way SoftMax, providing confidence in all of these three possible outcomes: early stage, late stage, and no infection at all.

Just like a traditional convolutional feed-forward network, your CNN will be trained using stochastic gradient descent and a dropout in your dense layer in order to prevent overfitting. Therefore, you will further tune your CNN for the infection stage, and preliminary results will indicate that this neural network performs much better than the classic approaches to machine learning, like Bayesian classification techniques or vector machines.

In addition, some of the most recent works done by Baidu and Google indicate that CNNs are have great accuracies, and at some tasks, they are even better than humans. It really seems that CNNs can be of great influence on some of the futuristic technology that is just around the corner.

Image recognition

One of the applications of machine learning models is sorting and classifying data. This model can even be used for the classification of images. Search engines use this kind of algorithm to identify photos, and social media sites now use facial recognition to identify a person in a photo before the photo is even tagged. They do this by learning from data composed from other photos. If your social media account can recognize your face in a new photo, it's because it created models with data from all the other photos in your account.

Image recognition techniques require in-depth learning models. These models are created with an artificial neural network, which will be discussed in more detail later in this book. Deep learning is the most complex type of machine learning where data is filtered through several hidden layers of nodes. They are called hidden layers because the models are unattended, which means that the features identified by the model are not preselected by the data scientist. Usually, the features are patterns that the model identifies on its own. Functions identified in neural networks can be quite complicated; the more complex the task, the more layers the model will have. Image sorting models may have only two or three layers, while self-driving cars have between one and two hundred hidden layers.

We have made great strides in this area in recent years thanks to the increased availability of computing power. Imagine the computing power needed to run thousands of data points through hundreds of stacked nodes at once. Deep learning and artificial neural networks have become more feasible over the past decade with the improvement of computers and the reduction of costs for processing large amounts of data. Certainly, with the advent of the cloud, data scientists have access to enormous amounts of data without using physical storage space. **ImageNet** is a website that is a great resource for data scientists interested in photo classification and neural networks. It is a database of images that is publicly accessible for use in machine learning. The idea is that, by making it publicly available, data scientists from around the world can collaborate to improve machine learning techniques.

The ImageNet database has approximately 14 million photos, with over 21,000 possible class groups. This offers a world of opportunities for data scientists to access and classify photos for learning and experimenting with neural networks.

Every year, ImageNet hosts a competition for data scientists worldwide to create new image classification models. The competition gets tougher every year. Now they are starting to move to classify videos rather than images, which means that the complexity and required processing power will continue to grow exponentially. Using the millions of photos in the database, the ImageNet competition has made groundbreaking advances in image recognition in recent years.

Modern photo classification models require methods that can work very specifically. Even if two images have to be placed in the same category, they can look very different. How do you make a model that can distinguish them? Take a look at the following image:

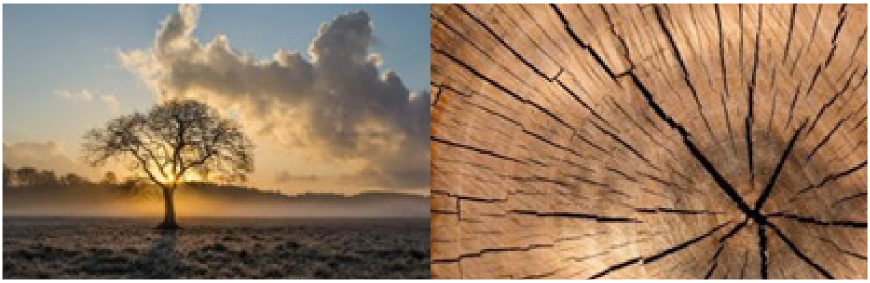


Figure 5.1

These are two different photos of trees. Ideally, if you were to create a neural network model that classified images of trees, you would want your model to categorize both as photos of trees. A person can recognize that these are both photos of trees, but the characteristics of the photo would make it very difficult to classify them with a machine learning model.

The fewer differences the variables have, the easier it will be to classify them. If all your photos of trees looked like the image on the left, with the tree in full view with all its features, the model would be easier to make. Unfortunately, this would lead to overfitting, and when the model is presented with photos like the one on the right, your model may not classify it correctly. We want our model to be able to classify our data even if they are not that easy to classify.

Incredibly, ImageNet has been able to create models that classify data with many variables and very similar data. Recently, they have created image recognition that can even identify and categorize photos with different dog breeds. Imagine all the variables and similarities that the model would need to recognize in order to properly see the difference between dog breeds.

The challenge of identifying similarities between classes is known as intra-class variability. If we have an image of a tree stump and a photo of a tree outlined in a field, we are dealing with variability within the class. This problem is how variables within the same class can differ, making it more difficult for our model to predict which category they will fall into. Most importantly, a lot of data is needed over time to improve the model and make it accurate.

To have an accurate model despite high levels of variability within the class, we will have to use additional techniques with our neural network models to find patterns between images. One method involves the use of CNNs. Instead of just having one model or algorithm, data is fed through several models stacked on top of each other. The neural networks convert image features into numerical values to sort them.

Biometric recognition

As today's technology has sneaked into every nook and corner of modern living, the protection of personal data has become more crucial. Biometric technology uses our body as a natural identification system through the application of statistical analysis to physiological or behavioral data. Now we are in the age of a technological revolution in the field of biometrics with a wide range of research and product developments taking place to utilize the complete benefits of this exciting technology in its entirety.

Biometrics deals with measuring the physiological or behavioral information to verify an individual's identity, and hence it is accurate and reliable.

The two different phases involved in any biometric system are enrollment (registration) and verification (authentication). If a positive match is established, the user will be provided with the privileges or access to the system or service. In the case of personal identification, the entire database needs to be searched against the query biometric template. A good example of a personal identification system is an automated fingerprint identification service (AFIS), which is used by many law enforcement agencies to identify and track known criminals.

The biological characteristics used for identification or authentication should be quantifiable or measurable, as only quantifiable characteristics can be compared to obtain a Boolean result (match or non-match). The different components in a generic biometric system are sensors or data acquisition modules, preprocessing and enhancement module, feature extraction module, matching module, and the decision-making module. Capturing the user's biometric trait for authentication is performed by the data acquisition module, and most of the time the captured data will be in the form of an image. The quality of the acquired biometric data needs to be improved for better matching performance in the preprocessing stage. The salient features are then extracted from the enhanced biometric data in the feature extraction stage. The resulting feature template is stored in the database, and it is used for future comparison with query biometric templates in the matching stage. The final decision of the comparison is taken by the decision module based on the match score obtained.

Biometric-based authentication or person identification is currently used in various applications, and its main advantages include the following:

- Biometric systems are based on who a person is or what a person does, not on what a person knows (password, PIN) or what a person has (token, smart card).
- Biometric systems use a physiological or behavioral characteristic for authentication. It is unique and accurate as the duplication of a person's biological characteristic is rather difficult.

- Stealing of the biometric data and its re-usage is difficult. Users are relieved from the need to remember passwords, and forgery can be minimized as biometric features cannot be shared.

There are some important characteristics of biometric traits that need to be analyzed before fixing the appropriate biometric trait to be used in a given application. The different characteristics of a biometric trait that need to be taken under consideration are as follows:

- **Universality:** The selected trait should be possessed by almost all individuals.
- **Inter-intra-class performance:** There should be sufficient distinctive features between the inter-class templates (templates of two different individuals). Intra-class templates (templates of the same individual) should only possess a minimum amount of distinctiveness.
- **Collectability:** It should be easy to collect the biometric template of the selected trait from users.
- **Acceptability:** The target population should be willing to reveal the selected biometric template to the biometric system, and the user interface of the system should be as simple as possible.
- **Cost:** The infrastructure and the maintenance cost of the system that can process the selected trait should be minimal.

Fingerprint and iris are the commonly used biometric traits in most of the biometric systems. This is mainly because of its user convenience (high for fingerprint-based systems) and accuracy (high for iris-based systems). Aadhaar (Aam Aadmi ka Adhikar) recognized as the world's largest universal civil ID program and biometric database is currently used by the Government of India to provide social services to the citizens. Both fingerprint and iris biometric traits are taken during Aadhaar enrolment:

Features and ML techniques		
Modality	**Features**	**ML techniques**
Face	Distance between eyes, DCT, Fourier transform, Ratio of the distance between eyes and nose, Principal components	PCA, LDA, Kernel PCA, Kernel LDA, SVM, Deep neural network

Iris	DCT, Fourier transform, Wavelet transform, Principal components, Texture features	PCA, LDA
Fingerprint	Delta, Core points, Ridge ending, Island, Bifurcation, Minutiae, FFT	Artificial neural networks, Support vector machine, Genetic algorithms, Bayesian training, Probabilistic models
Finger vein	LBP, Minutiae, Bifurcation and endpoints, Pixel information	SVM, Deep learning
Palm print	Shape, Texture, Palm lines, PCA, LDA coefficients, DCT	Naive Bayes, KNN, HMM
Palm vein	LBP, Minutiae, Bifurcation and endpoints, Pixel information	SVM, Deep learning
Voice	Linear prediction coefficient (LPC), Cepstral coefficient (CC), MFCC features	Gaussian mixture models, HMM, ANN, SVM deep learning

Table 5.1

Machine learning has played a major role in improving the performance of biometric systems. One-to-one or one-to-many matching tasks can be done automatically and seamlessly in biometric systems with machine-learning-based algorithms.

Software vulnerabilities

Vulnerability refers to a flaw in a system that can leave it open to attacks. A software vulnerability is a flaw in a software system that can cause the software or system to crash or produce invalid output or to behave in an unintended way. Some of the common software vulnerabilities are as follows:

- Buffer overflows.
- Numeric over- and underflows.
- Errors in type conversion.
- Operator misuse.

- Bugs in pointer arithmetic.
- Evaluation order logic errors.
- Structure padding errors.
- Procedure errors.
- Errors in macros and preprocessors.
- String and meta-character vulnerabilities.
- Privilege problems.
- File permission issues.
- Race conditions.
- Errors in processes, IPC, and threads.
- Environments and signaling issues.
- SQL injection vulnerability.
- Cross-site scripting (XSS) vulnerability.
- Vulnerabilities in file inclusion and access.
- Vulnerabilities in shell invocation, configuration.
- Access control and authorization flaws.

Software vulnerability detection is the process of confirming whether a software system contains flaws that could be leveraged by an attacker to compromise the security of the software system or that of the platform on which the software system runs. Code injection attacks allow an attacker to execute a (malicious) code within the privileges of the vulnerable program. Identifying the vulnerabilities and fixing them are very important measures to evaluate and improve the security of the software systems and the platforms on which they are running.

The machine learning approaches for vulnerability detection methods can be classified as follows:

- Anomaly detection methods.
- Pattern recognition methods.

Anomaly detection methods use features such as API usage patterns, missing checks, lack of input validation, and lack of access controls. Machine learning techniques such as KNN have been used for classification. Pattern recognition methods identify vulnerable lines of code along with keywords (specific to programming languages) using machine learning algorithms. The features include system calls and API calls invoked by an application, syntax trees, etc. Machine learning techniques such as logistic regression, multilayer perceptron (MLP), random forests, neural networks, BLSTM, etc., have been used for the classification of software vulnerabilities.

Conclusion

Computer vision plays a huge role in bridging the gap between human and machine. With the groundbreaking technological advancements today, this will be a very helpful tool in automating verification, especially in the finance sector (for example, bank accounts). In this chapter, we discussed the technical architecture of CNNs used in computer vision. We also had a walkthrough of the application of computer vision in image recognition, biometric recognition, and software vulnerabilities.

Chapter 6
Neural Network, GBM, and Gradient Descent

Introduction

Neural networks enable AI to solve complex problems. Since these networks are designed similar to the human brain and nervous system, data scientists and machine learning engineers can develop high performing AI systems that work similar to a human brain. Neural networks, deep learning, and AI represent powerful and exciting machine-learning-based techniques that are best suited for solving real-world problems. Although no computer could ever achieve the level of human intelligence, AI models based on decision-making, inference, and reasoning can deliver great outcomes and accurate predictions.

Billions of neurons with trillions of have connections between them—that's what drives our intelligence and our ability to learn and adapt when we encounter something new.

Back in the 1950s, scientists first developed the concept of an artificial neuron that simulated, very simplistically, the behavior of biological neurons. Later, the idea of linking several artificial neurons together, to create a neural network, was proposed as a way to represent complex problems. However, it has only been in the last few years that the necessary computing capabilities have become readily available to allow their full potential to be realized. Today, advanced neural network models are at the cutting edge of AI and machine learning research.

To show how a neural network works, let's start with the basic building block of a neural network: the neuron.

Structure

In this chapter we will cover the following topics:

- Working of neural networks.
- Types of neural networks in AI.
- Benefits of using artificial neural networks.
- Gradient-boosting algorithms.
- Gradient descent.

Objective

After studying this chapter, you should be able to do the following:

- Understand the concept of neural network, GBM, and gradient descent.
- Understand how neural networks work, their different types, and their benefits.

Working of neural networks

Similar to a human brain, neural networks have the capability to make decisions and give predictions that are best suited for the given data set. They interpret sensory data by using machine learning perceptions and clustering of raw input. Further, the patterns recognized by neural networks are contained in vectors and are usually numerical. Input data for neural networks can include real-world data such as text files, images, audio, video, or graphical files. Each type of data is translated into a format that is suitable for the neural networks so that relevant operations could be done without any hassle.

Moreover, neural networks are based on classification and clustering models, which make it easier for machine learning models to achieve better data insights. Neural networks help manage unlabeled data by reviewing the similarities within the inputs. After the review is done, the model classifies the data and gives a labeled data set to train on. For example, if we want to develop a network that identifies cats, the initial training should include a series of pictures of cats from each angle. As each input is given with a matching identification, such as "animals" or "not animals," the model can differentiate and deliver accurate predictions without any hassle.

To determine and define rules, the decision of each node is sent to the next tier; thus the rules depend upon the inputs from the previous tier. Neural networks are based on genetic algorithms, gradient-based training, Bayesian methods, and fuzzy logic theorems. Refer to the following image:

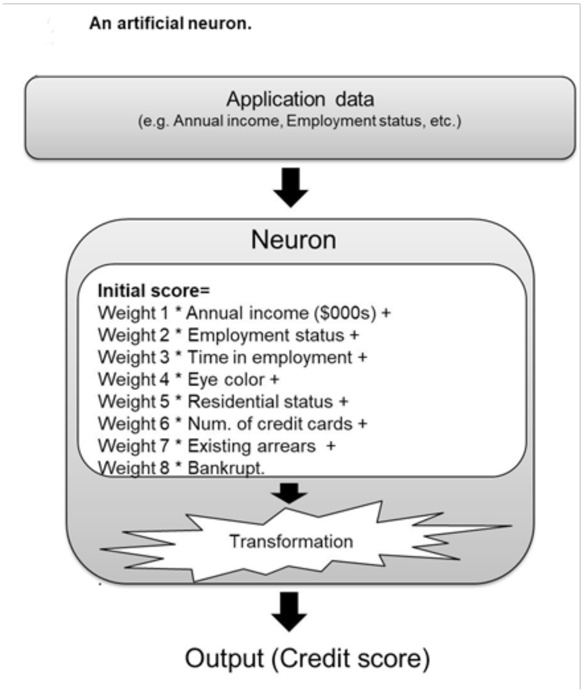

An artificial neuron.
Application data
(e.g. Annual income, Employment status, etc.)
Neuron
Initial score=
Weight 1 * Annual income ($000s) +
Weight 2 * Employment status +
Weight 3 * Time in employment +
Weight 4 * Eye color +
Weight 5 * Residential status +
Weight 6 * Num. of credit cards +
Weight 7 * Existing arrears +
Weight 8 * Bankrupt.
Transformation
Output (Credit score)

Figure 6.1

The working of a neuron can be explained as follows:

1. Application data provides the inputs to the neuron. This is just like in scorecard and decision tree models.
2. Each input is multiplied by a weight.
3. The resulting values are added together to get an initial score.

4. The initial score is transformed to lie within a certain range, often between 0 and 1. This is so that when several neurons are combined to produce a neural network, all the neurons produce values in the same range.
5. The transformed version of the initial score is the output produced by the neuron; i.e. the credit score.

An artificial neuron isn't mysterious or complex. It isn't really much different from the scorecard model. The main difference is that the neuron score lies in a fixed range. To produce a neural network, several neurons are connected together, as shown below:

Figure 6. A neural network model.

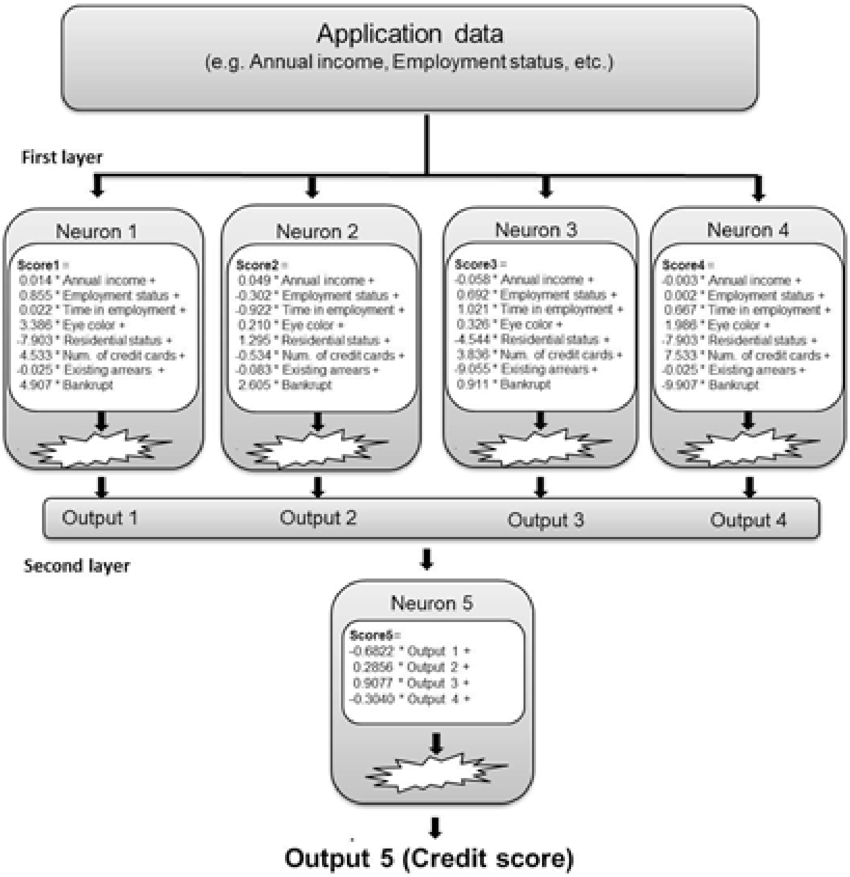

Application data
(e.g. Annual income, Employment status, etc.)
First layer
Neuron 1
Score1 =
0.014 * Annual income +
0.855 * Employment status +
0.022 * Time in employment +
3.386 * Eye color +
-7.903 * Residential status +
4.533 * Num. of credit cards +
-0.025 * Existing arrears +
4.907 * Bankrupt
Neuron 2
Score2 =
0.049 * Annual income +
-0.302 * Employment status +
-0.922 * Time in employment +
0.210 * Eye color +
1.295 * Residential status +
-0.534 * Num. of credit cards +
-0.083 * Existing arrears +
2.605 * Bankrupt
Neuron 3
Score3 =
-0.058 * Annual income +
0.692 * Employment status +
1.021 * Time in employment +
0.326 * Eye color +
-4.544 * Residential status +
3.836 * Num. of credit cards +
-9.055 * Existing arrears +
0.911 * Bankrupt
Neuron 4
Score4 =
-0.003 * Annual income +
0.002 * Employment status +
0.667 * Time in employment +
1.986 * Eye color +
-7.903 * Residential status +
7.533 * Num. of credit cards +
-0.025 * Existing arrears +
-9.907 * Bankrupt
Output 1
Output 2
Output 3
Output 4
Second layer
Neuron 5
Score5 =
-0.6822 * Output 1 +
0.2856 * Output 2 +
0.9077 * Output 3 +
-0.3040 * Output 4 +
Output 5 (Credit score)

Figure 6.2

The credit score (Output 5) from the network is calculated as follows:

1. The application data is supplied separately to each of the four neurons in the first layer.
2. Each neuron has its own weights. The weights, when combined with the application data, create scores (Score1, Score2, etc.).
3. The scores are transformed to produce four outputs.
4. The four outputs provide the inputs to the single neuron in the second layer (Neuron 5).
5. Neuron 5 combines the inputs with a further set of weights to create Score5.
6. Score5 is transformed to create the final credit score.

The collection of weights in the network represents the patterns identified by the training algorithm. If you want a biological analogy, the weights can be thought of as a memory that can be recalled whenever you want to make another credit scoring decision.

How does the training algorithm determine what the weights should be? Many different algorithms can be used to find the weights in a neural network, but they all tend to adopt the following principles:

1. Assign each weight a random or zero value.
2. Calculate the scores generated by the network for all cases in the training data.
3. Assess the model's accuracy, e.g., for the credit scoring model, how well it assigns high scores to good-paying customers and low scores to defaulters.
4. Adjust the weights to improve the model's accuracy. For the credit scoring example, good payers get higher scores and bad payers get lower scores.
5. Repeat steps 1 to 4 until no further significant improvement in model accuracy is observed.

Complex math occurs in step 4. This process of adjusting the weights is called training. A simple training approach is to randomly try different values and see what works best. However, this is very inefficient. Even the most powerful computer could run for years and still not find a good model using this approach. In practice, neural network training algorithms are cleverer than this. They adopt different weight adjustment strategies based on the model's performance between each iteration of the algorithm. The algorithm terminates when further adjustments deliver no significant improvement in model accuracy.

One reason neural networks are popular is that they are often better than scorecards and decision trees at spotting subtle patterns in the training data. Their main drawback is that their outputs are not intuitive. You may know what all the weights are and understand how the final score is calculated. For example, if I asked you which of the inputs in the image above contribute most to the final score, then that's far less obvious than with the scorecard model. This can be a problem if there is a business or legal requirement to explain how the model score was arrived at.

Deep learning is the latest evolution of neural networks. The network shown earlier has two layers of neurons, but there is no reason why there can't be more layers, as illustrated below:

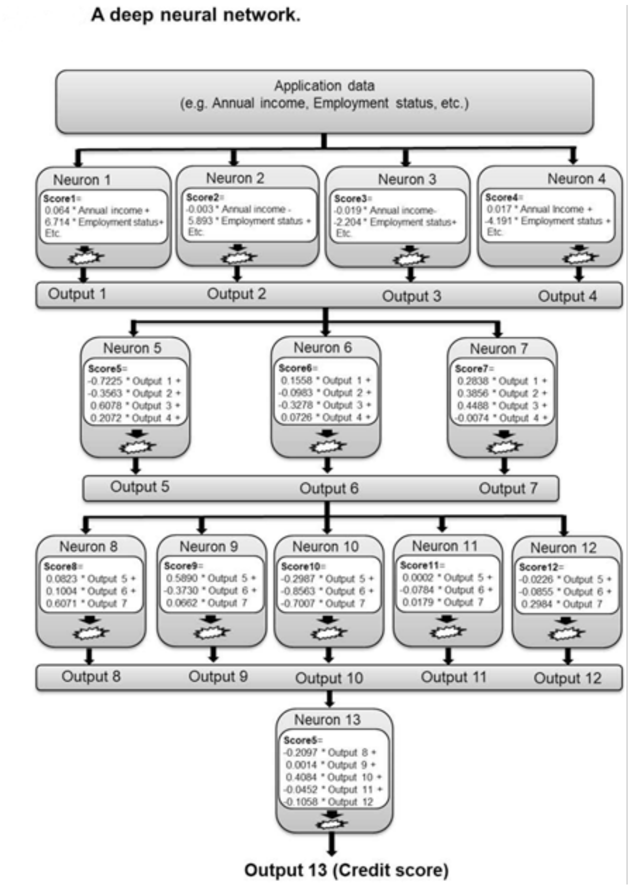


Figure 6.3

The two models take exactly the same input data and deliver the same type of output. However, the network in *Figure 6.3* has two extra layers of neurons. What you tend to find is that as more layers are added, the ability of the network to identify complex patterns increases. The more layers, the deeper is the network.

The networks in the above figures have a single neuron in the final layer that generates a single credit score. Another big strength of neural networks is that they can be structured to have more than one output. This is important for tasks where there can be thousands of options to choose from, each requiring a separate output.

Returning to our object recognition task, let's think about how a neural network might be structured to identify objects in a picture. As before, we need to gather lots of environmental information for the training algorithm to use. In this case, let's assume that the training data contains several thousand pictures of cats, cakes, and potatoes as follows:

- **Input data:** Each pixel in a picture is represented by four pieces of data—a red, blue, and yellow component to indicate the color of the pixel, plus a value to represent the intensity (brightness) of the pixel.
- **Category data:** Each image is labeled to indicate if it contains a picture of a cat, a cake, or a potato.

The objective is to use the input data to predict the category of the object in the picture (cat, cake, or potato). A neural network model like this could have millions of inputs, one for each pixel element, thousands upon thousands of neurons spread across a dozen or more layers, and three outputs as illustrated below:

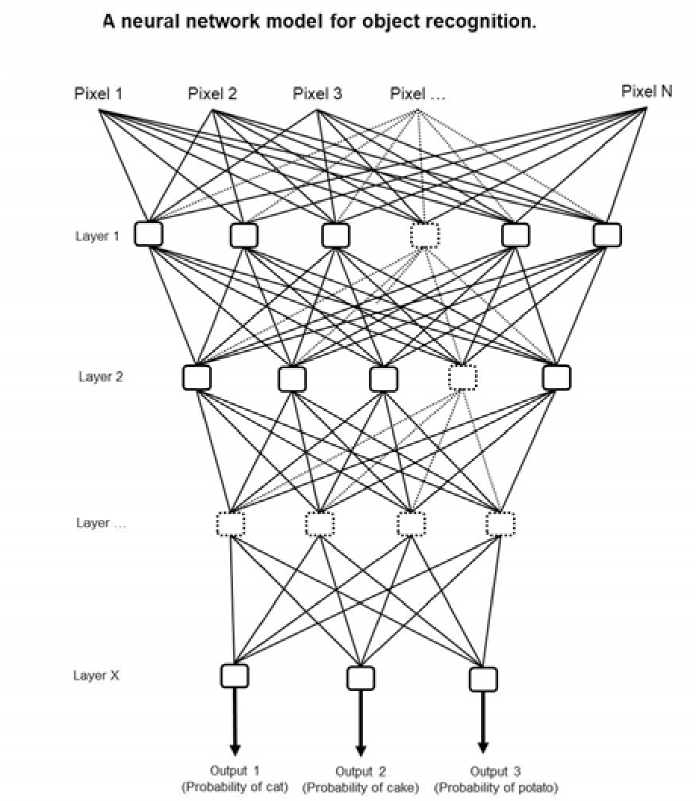


Figure 6.4

In the figure above, each output represents the probability that the image is either a cat, cake, or potato. After training, when you present the network with a completely new image, the pixels representing the image are processed through the network. Then, you simply compare the three outputs and select the one that has the highest probability. If output 1 gives a 5% chance that the image is a cat, output 2 a 15% chance it's a cake, and output 3 an 80% chance it's a potato, then the conclusion is it's a picture of a potato.

If we want a more general system that can identify many more everyday objects, then the same principles apply. It's just a case of having enough images to train a suitably structured network with.

Other avenues of research associated with deep learning consider how the neurons in the network are connected. This makes it possible to incorporate a representation of time or event order. Text prediction (like you get on your phone) is one such example. The word order is a key factor in predicting what word comes next. Another approach is to create sparse networks (convoluted networks) that contain relatively few connections between the neurons in different layers. This can help reduce the computing power required for certain types of problems.

One recent advancement in deep neural networks is general adversarial networks or GANs, where two neural networks compete with each other to generate new and original content.

Let's say that we want to produce a neural network that can create artificial pictures of people. They look just like real people, but the people don't actually exist. Why would you want to do this? Maybe it's just for a bit of fun, but maybe you want to use the pictures in advertising or on virtual catwalks so that you don't need to get the permission of real people to use their images and so on.

We start with two untrained neural network models. For argument's sake, let's call the first neural network the "Judge" and the second the "Student."

The Judge assesses the images it's presented with, and decides if the images are of real people or something else. The Judge has a single output that represents the probability that the image is real. In principle, the Judge works just like any of the other models. If the model output is more than 50%, then we assume that image is real; otherwise we assume it's artificial.

Now, let's think about the Student. The Student's goal is to create realistic pictures of people. The Student is doing its job well if it can fool the Judge; i.e., it creates fake pictures that the Judge assigns a 50%+ probability of it being a real person.

The Student model works back to front. With a normal object recognition system, the inputs to the model are a set of pixels and the output a set of probabilities. With the Student, however, the outputs from the model are a set of pixels and the inputs are a set of random numbers.

First, we train the Judge using images that are labeled to indicate if they are of real people or something else. Once this initial training is complete, then the Judge is pretty good at identifying images of people.

Returning to the Student, the model weights are initially chosen at random. The result? A set of meaningless images of random dots. These are presented to the Judge to assess, which has no problem identifying all the images as fake; i.e., it assigns them very low probabilities of being real.

The probabilities generated by the Judge are fed back to the Student, and the training algorithm adjusts the model weights so as to produce a better set of images next time round. The algorithm used to train the Judge is also rerun. However, this time, the training data also includes the images generated by the Student.

This process is repeated many times, with each network continually learning from the outputs of the other. Eventually, a status quo is reached. The Student produces images of people that look real. As far as the Judge can tell, they are real, even though they are completely artificial.

Creative opportunities for GANs are vast. Amazon, for example, is reported to be using GANs to design garments by using images of fashionable clothing to train it to produce new designs that have similar stylistic features but which are completely original. Likewise for music and other creative works.

GANs show huge promise, but the potential for misuse is also extensive. In particular, it's driving many of the "Deepfake" stories being reported in the press to create fake media recordings of politicians and other famous people.

Why are neural network models so good at what they do? How do they manage to capture the nuances of complex problems leading to better judgments than the best human experts? One way to think about this is that learning to play a game like chess or Go is analogous to mapping an unknown landscape that the training algorithm explores as it goes. The model weights capture the features of the landscape and how one gets from one part of the landscape to another using different actions (game moves).

In mapping the features of a game, the training algorithms find parts of the landscape that people haven't explored before, as well as new (better or more efficient) routes for traveling between two points. In effect, they have discovered new styles of play that have never been seen by human players before.

It has been argued that, in finding these new styles of play, neural networks have displayed non-human, i.e., alien, intelligence.

However, another perspective is that it's not a new way of thinking, but rather the training algorithms have found new features and patterns that humans have just not got around to thinking about yet. It doesn't necessarily mean that the network is displaying a different type of thinking altogether, but instead, the training algorithm has found a new area of the landscape to explore that people have not yet reached.

This principle of mapping and exploration also applies in other application domains, not just games. One example is supporting scientific discovery, particularly in areas of research where huge numbers of possibilities have to be explored to find useful solutions—the needle-in--haystack-type problems.

A classic example of this type of problem is predicting the 3D structures of complex molecules such as proteins. There are trillions of possible protein molecules, but only a handful have any practical applications. Being able to determine the likely shape of a molecule from its component parts means it's much easier to identify which molecules are likely to be useful and which are not.

What's amazing about machine learning models, and deep neural network models in particular, is that they can be trained to predict almost any outcome imaginable to a degree that equals or exceeds the best human decision-makers if good-quality information (training data) is available. Want to predict the outcome of baseball or soccer games? Then feed the training algorithm with data about historic games and a well-designed model will beat the best human pundits. Want a tool that decides which stocks to buy? Then better to trust a machine learning model than a stockbroker. Want to optimize the layout of a factory or supermarket? Then there are neural network algorithms that can do that better than any human could.

All of the different ways of building models that we've discussed so far have assumed that there is a database containing data about the environment, which includes details about the thing you want to infer (predict). Let's summarize the tasks we've considered so far:

- When building the decision tree model to predict heart disease, patients' medical records were flagged to indicate which one developed heart disease.
- When creating the scorecard and neural network models for credit scoring, each loan application was matched to data about loan repayment.

- When talking about using a deep neural network to identify images, each image in the training data was tagged to indicate if the image was of a cat, cake, or potato.

For all of these tasks, the training data was labeled to indicate what the outcome was in each case. This type of machine learning, using labeled training data, is called supervised learning.

Most real-world AI-based applications, such as target marketing, voice recognition, and employee vetting, are examples of supervised learning. However, there are times when there isn't any labeled data for the training algorithm to use. In these situations, a different set of techniques, referred to as unsupervised learning, can be applied.

The most common type of unsupervised learning in use today is clustering. Clustering algorithms are based on the principle of minimizing the distance between cases in the training data. This doesn't necessarily mean physical distance, but how similar cases are in terms of specific data items. The distance between two people aged twenty-three and twenty-five is less than that between people aged seventeen and seventy-six. If we are talking about smoking, then two smokers have a distance of zero, whereas a smoker and a non-smoker don't, and so on.

With clustering, a model isn't produced at the end of the process, which means it can't be used to generate predictions about how individual cases are going to behave. There is just an identifier to say which cluster an observation belongs to. For example, you are in cluster 9 while I've been assigned to cluster 4.

Customer profiling to group similar people together is one of the best-known applications of clustering, but clustering approaches are also being applied successfully to many other tasks such as document clustering. In fields such as law and academic research, there is a requirement to regularly trawl through the ever-growing pile of published literature to find information relating to certain cases or types of research. In medicine, for example, almost a million academic papers are published each year. Documents are grouped (clustered) together on the basis of how similar they are in terms of subject matter, writing style, word count, and a host of other features.

Document clustering can also be applied to tweets, news feeds, blog posts, and other rapidly changing media in real time. Pictures and videos can also be categorized in a similar way. News organizations use these approaches to automatically flag new posts about specific topics as they appear. They can then include them in their own media publications almost immediately. Social media companies and governments use similar methods (as well as predictive models) to identify illegal or undesirable content.

Supervised learning works well when there is a large amount of labeled data. If there are no labels available, then unsupervised learning, such as clustering, can sometimes prove useful, albeit in a somewhat different context. There are, however, situations where there may be no data initially, but the learning process is able to assess its performance on a case-by-case basis as it goes. The model is adjusted each time, on the basis of some measure of success or reward that is calculated each time a task is attempted. This type of machine learning is called reinforcement learning.

Data scientists often talk about three distinct types of machine learning: supervised, unsupervised, and reinforcement learning. However, reinforcement learning shares many features with supervised learning. In particular, it delivers a model typically based on some form of neural network. The key difference is that it generates training data as it goes, rather than using pre-existing training data to derive the final model.

A great example of the difference between supervised and reinforcement learning is how model training occurs to create a chess-playing program. A supervised approach would take thousands of game moves (or sequence of moves) from previously played games as observation data, with the labeled output data providing an indication of whether the move was a good one or not. The scores produced by the model are used to indicate which piece to move and to where. The algorithm then finds the model weights that result in the overall best set of moves, measured against the moves contained in the development sample.

With reinforcement learning, no data is provided initially—none at all! The model scores still indicate which move to make just like the supervised approach. However, initially, these will be more or less random, given that there isn’t any training data. Each time a move is made, the status of the board (the new state of being) is reevaluated. The algorithm then adjusts the weights in the model on the basis of how successful its move was deemed to be.

Evaluating the success of a move in chess is complex and will often incorporate probable future states of being, as well as the current one.

However, for the purposes of this example, a simple success criterion is the difference in the value of each player's pieces remaining on the board after a move has been made. If your move results in the taking of a high-value piece, but not losing one yourself, then that's a strong success. Making a move and then having one of your pieces taken is a failure, yielding a low measure of success. The ultimate success or failure is losing one's king—checkmate. In this way, by assessing each move and adjusting the model weights accordingly, the program learns by itself without needing to be supplied with any prior information.

An advantage of reinforcement learning is that there's no limit to the set of moves that can be explored as the training algorithm progresses. With supervised learning, you are limited to the labeled examples in the training data. For a game like chess, even a huge amount of training data, from millions of games, will contain only a tiny proportion of all possible moves. This was demonstrated very effectively when Google's DeepMind AI team used two reinforcement algorithms to play against each other. Not only did the resulting model outperform the best existing chess program at the time, but during the process, the algorithm discovered completely new strategies of play, previously unknown to human grandmasters.

Reinforcement learning has potential, but it does have its weaknesses. One issue is that the training algorithm is bound by the speed of the trial-and-error process. If a reinforcement algorithm is being trained to make mortgage lending decisions, then the time between an action being taken and assessing how successful it was could be years. Consequently, the training process will take far too long to be of practical use. DeepMind's chess-playing program needed to play 68 million games to become as good as it did. It could only do this by playing against another computer, allowing games to be played in a fraction of a second.

OK, so being superhuman at chess is all well and good, but what are the real-world applications of reinforcement learning? I've done quite a lot of research, and my conclusion is that, although there are lots of interesting articles about how great reinforcement learning is, the number of practical solutions based on reinforcement learning in use today are pretty small compared to those based on supervised learning.

One area where reinforcement learning is having an impact is the improvement in the efficiency of control systems in uncertain or chaotic environments. Complex systems such as power plants, heating systems, and server farms have lots of controls that are adjusted to manage different parts of the system. The relationship between the controls and system performance isn't always obvious. It's not a simple case of one control impacting just one parameter; everything is interconnected. Tweaking the water pressure in the cooling system to improve turbine performance has a negative impact on the efficiency of a transformer further down the line. Just like the chess problem, where there are almost infinite number of possible games that can be played, in a power plant, there are almost infinite number of combinations of control settings. Reinforcement learning finds better ways of setting the controls through trial and error, to optimize the overall efficiency of the system. This mirrors the way an experienced engineer uses their knowledge and intuition, learnt over many years of practice, as to what the system-wide effects of tweaking different controls are likely to be.

Another area where reinforcement learning shows promise is robotics. It can be used to train robots to carry out complex manual tasks that could previously only be undertaken by a trained person. A robotic device is given the task of doing something like pulling pints, sorting parts, or stacking shelves. Through a process of trial and error, they can potentially learn to do this very effectively.

Types of neural networks in AI

As discussed before, the main components of a neural network are input layers, output layers, and hidden layers. Hidden layers are present between the input and output layers that define the relationships between the data and also help machine learning models to generate data insights and predictions in a better way. Neural networks are associated with deep learning because of the forward and backward propagation of data within the tiers.

Here is a brief description of the major types of neural networks.

Feed-forward neural networks

The feed-forward neural networks are the base version of a neural network, and they have the responsibility to pass data through different input nodes. For the data to reach the output node , the process is repeated multiple times and it has the power to process large amounts of data without any interruption. Further, feed-forward neural networks are used in the development of major AI technologies such as computer vision and facial recognition. As compared to other types of neural networks, feed-forward neural networks are simple to operate and yield effective results.

Recurrent neural networks

Recurrent neural networks have the capability to save the output from processing nodes and return the results back to the model. Since they are complex in comparison to feed-forward networks, data scientists and machine learning engineers use this type of neural network to solve complex problems. Each node in the recurrent neural network model is considered as a memory cell and is responsible for completing the implementation and continuity of operations.

The process of data handling includes the reuse of outputs, and in case the network is unable to deliver correct inputs, the system starts to learn on its own and continues to provide accurate predictions during the backpropagation.

Convolutional neural networks (CNNs)

CNNs use different perceptrons and have more than one layer connected with each other. Further, these layers are responsible for creating future maps on information related to data so that it could be further broken down for non-linear processing. CNNs are mostly used for the development of advanced AI applications such as text digitization, facial recognition, and NLP.

Deconvolutional networks work completely opposite of CNNs. They are supposed to find lost signals or features that are ranked as unimportant by a CNN.

Modular neural networks

Modular neural networks have multiple neural networks that work separately from each other. Moreover, the networks do not communicate in any case and never interfere in the activities of other neural networks.

Benefits of using artificial neural networks

Artificial neural networks are best suited to store information through the entire network and have parallel processing abilities as well. This makes it easier for artificial neural networks to perform multiple tasks simultaneously without compromising on performance and efficiency. Moreover, they have the ability to model complex relationships and learn from nonlinear data sets by defining relationships between the input and output data. Artificial neural networks do not have any restrictions for input variables, and they can be distributed in different patterns as well.

Artificial neural networks are widely implemented in the development of various AI and machine learning models. Popular neural network applications are language generation, translation, and NLP systems. Chatbots, stock market prediction, route planning, and optimizing systems are also developed using neural networks.

Gradient boosting algorithms

Gradient boosting relies on one premise—a weak learner can improve. As Michael Kearns said in his "Hypothesis Boosting Problem," it is the pursuit of turning a poor hypothesis into a very good one.

Gradient boosting algorithms come in different forms.

GBM

GBM is used when there is a huge amount of data to deal with. Boosting produces a number of learning algorithms made up of predictions from different base estimators. These, in turn, will trigger improvements on a single estimator. It puts together a mix of weak to average predictors in an effort to create a strong predictor.

Light GBM

Light GBM utilizes tree-based learning algorithms. It is highly efficient and exhibits faster training speed than other models. It also uses an impressively low amount of memory while delivering better accuracy. It is supported by GPU and parallel learning, and can easily handle huge amounts of data.

Light GBM was developed using Microsoft's Distributed Machine Learning Toolkit Project. It's perfect for classification and ranking tasks, among others.

XGBoost

XGBoost is yet another gradient boosting algorithm that dictates a win or a loss in Kaggle competitions. It has amazing predictive powers, giving it the advantage in terms of accuracy. This is because it includes both the tree learning algorithm and a linear model. This makes it ten times faster than any other gradient boosting technique.

How is XGBoost applied? It can actually be used across a number of applications including classification, ranking, and regression.

This algorithm is also known as a regularized boosting technique. It's perfect for reducing overfit modeling. It also works great for different languages like C++, Java, Julia, Python, R, and Scala.

CatBoost

CatBoost is fairly new in the game and was developed by Yandex. It is an open-sourced algorithm and can be easily integrated into different deep learning frameworks like Core ML from Apple and Tensor Flow from Google.

Where other machine learning models require extensive data training, it's a different story for CatBoost. It is still pretty robust, though, and can be used across many different data formats.

One huge advantage of CatBoost is that it can help you focus on fine-tuning your entire model instead of wasting time trying to fix small errors. This is because CatBoost automatically runs through the variable while discounting type conversion errors.

Gradient descent

Gradient descent occurs when there are one or more inputs; it is possible to optimize the values of the coefficients by reducing the error of the model on the training data iteratively. This is done by starting with random values for every coefficient. For every pair of input and output values, the sum of the squared errors is calculated. By reducing the error, a learning rate is used as a scale factor, and the coefficients are updated. The process is performed repeatedly until a minimum sum squared error is obtained, or no further improvement is possible.

When performing gradient descent, it is essential to choose a learning rate (alpha) parameter that ascertains the size of the improvement step to be applied to every iteration of the procedure. Gradient descent is often implemented using a linear regression model, and it is relatively straightforward to understand. In practice, it is used when a large-scale data set contains a massive number of rows or columns that may not fit into memory.

Conclusion

User experience will be widely improved with the development and implementation of advanced neural networks. In this chapter, we discussed how they work, their different types, and their benefits. We also had a walkthrough of different types of gradient boosting algorithms, and discussed the advantages of gradient descent in neural networks.

In the next chapter, we will try to understand sequence modeling.

CHAPTER 7
Sequence Modeling

Introduction

The ability to predict what will come following in a series is interesting. Interestingly, human brains are very good at it, but that's not the case for computers. In the face of a shadowy plot in a movie, the human brain begins to yield effects. But how do you train robots to do something similar?

Thanks to deep learning, we can do a lot more now than we might have done a few years before. The ability to work with sequence details, such as song lyrics, sentence translation, DNA sequence analysis, speech recognition—all this is now possible owing to sequence modeling.

Structure

In this chapter, we will cover the following topics:

- Word embedding.
- Feed-forwarding neural network algorithm.
- Convolutional neural network algorithm.

- Recurrent neural network (RNN) algorithm.
- Conditional random field (CRF) algorithm.
- Modeling procedure.

Objective

- Understand the different types of sequence modeling techniques.
- Understand the modeling procedure.

Word embedding

Word embedding help one generalize well when dealing with word representations. Suppose you are performing a named entity recognition task and only have rare observations in the training set. In such a scenario, you can either take pre-trained word embeddings or build your own embeddings. These embeddings may provide functionality for all the words in the vocabulary. They have numerous NLP applications such as named entity recognition, machine translation, and sentiment analysis.

Feed-forward neural network algorithm

Essentially, these are the classifiers for multi-level logistic regression. Nonlinearities (tanh, sigmoid, SoftMax + ReLU, and SELU) activation function work in different level of numerical scales. These are often referred to as multilayer perceptrons. FFNNs may be used as autoencoders for the description of "learning without an instructor." FFNNs can be used as autoencoders to train a classifier or to retrieve functions.

Convolutional neural network algorithm

With the advent of CNNs, practically all advanced successes in the field of machine learning have been realized. They are used for the recognition of pictures, identification of artifacts, and also segmentation of images. Invented in the early 1990s by Jan

Lekun, CNNs possess convolutional layers that act as extractors of hierarchical objects. You may use them for text work (and also for graphics work).

Recurrent neural networks (RNN) algorithm

RNNs create sequences by repetitively adding the same range of weights to the aggregator state at time t and data at time t. Only RNNs are seldom used today. However, in most series modeling issues, its analogs such as GRU and LSTM are among the most up-to-date. LSTM is used in pure RNNs instead of just a thin base. RNNs are used for text classification, time series forecasting, computer translation, and computational linguistics.

Conditional random field (CRF) algorithm

These algorithms are intended to represent an RNN-like series and can be used in combination with an RNN. These could also be used, for instance, in image segmentation and other organized prediction activities. CRF models each part in the series (say, a sentence), such that the neighbors influence the item mark in the series and not the individual labels. CRFs are used for sequences (image, text, time-series data, and DNA).

CRFs are a type of discriminative classifier, and they model the decision boundary between diverse groups. Owing to their ability to classify sequential data, CRFs are mostly applied in NLP. One of the applications is parts-of-speech (POS) tagging.

Section of sentence speech are based on previous words, and they take benefit of this through features, we can use CRFs to learn how to differentiate which sentence terms refer to which POS. An alternative related application is named entity recognition (NER) or extracting nouns from sentences. CRFs can be used to forecast any sequence in which numerous variables rely on each other. Some applications include visual and gene estimates.

Modeling procedure

The process of modeling comprises four steps:

1. Feature engineering and selecting a model.
2. Training the model.
3. Validating the model.
4. Testing the model on new data.

The first three steps are usually repeated because you most likely will not build an optimal model for your project on the first try. As such, you will be building several models and then select the one that performs the best on the testing data set (which is unseen data).

In addition, the testing step of the modeling process is not always performed because, in some cases, the goal of the data science project is root cause analysis (basically explanation) instead of predictions. For example, the goal of your project might be to determine the reason why some species are going extinct (explanation) instead of predicting which species is most likely to go extent (a prediction).

Another trick commonly used in machine learning is chaining. The way chaining works is that, when several models are chained, the output of one model is used as input by another model and so on. When chaining different models together, each of them needs to be trained independently of the others, and only their results are combined together. In machine learning terms, this technique is known as ensemble learning.

Feature engineering and selecting a model

In feature engineering, you basically create feasible predictors for the data model. This is the very first and also one of the most crucial steps in the modeling process because the model will be able to give predictions by combining these features, and the accuracy of the predictions depends on how good the predictors are. In this step, it is advised to consult machine learning experts as it is very difficult for beginners to create good predictors.

In textbook exercises, the data sets provided already include some features as variables; however, in practice, you will have to find the features all by yourself from different data sets. You may even find that these features are usually scattered between several different data sets. For example, in one of our data science projects, we had to procure a total of twenty data sets, after which we finally got the raw data we needed. This is a very detailed topic, and only touching the surface of feature engineering would not be helpful. Hence, it is recommended that you consult a machine learning expert or a dedicated machine learning book to learn about feature engineering. Once the features have been created, we need to choose the respective model to use with it.

Training the model

After we have created the right predictors and have selected the appropriate modeling technique for the project, we can now proceed to train the model on a training data set. A training data set is a data sample that you select for the model to learn to perform actions from. Popular modeling techniques are available to be implemented in almost any programming language that you may choose, including Python. These techniques essentially allow you to train your model by using a simple set of lines of codes. Advanced data science techniques require the scientist to be capable of using heavy mathematical calculations and then implement modern data science techniques to use with these calculations.

Once we have trained the model, the next thing to do is check whether it works as we intended it to.

Validating the model

In the field of data science, you will come across a lot of different modeling techniques, and the question here is, which modeling technique is best suited for your project? Well, there is no specific answer to this, but every good model has the same two features:

- Splendid predictive power.
- Generalizes fairly well to data that is new to it.

To create such a model, we need error measure in addition to a modeling technique. Usually, this error measure is paired with a validation strategy to properly validate the model.

By now, we already know that the most common uses of machine learning in data science are regression and classification. As such, the popular error measures for both of these cases are mean squared error and classification error rate, respectively. In the classification error rate, we receive an error report that indicates how your model performed on the test data, i.e., how many observations have been mislabeled (lower percentage is better). The mean squared error tells us the average percentage of the model's predictions error. However, it has one drawback. Faulty predictions can be in two directions, and squaring the error cannot cancel out the wrong prediction in one direction with another wrong prediction in the other direction. For instance, a model's prediction overestimates the next month's turnover by a value of 5,000. This wrong prediction cannot be canceled out by the model's prediction underestimating the next month's turnover by the same value of 5,000. Further, squaring the errors has one more problem, and that is large errors become even larger, although small errors can shrink (if they are less than one) or remain at the same level.

There are several validations techniques available, but the most common techniques are as follows:

- **Dividing your data into a training set with X% of the observations and keeping the rest as a holdout dataset:** This technique is the most popular.
- **k-folds cross-validation:** This validation technique divides the original data set into a defined number of parts. The model then uses each part as a testing data set, while the rest of the parts are used as a training data set. The advantage of this technique is that you put the entire data in the data set to good use.
- **Leave one out:** This validation technique is strikingly similar to the k-folds cross-validation technique, but with a slight change, that only a single observation is left out of the training data set, which is later used as test data. This technique is commonly used with data sets that are small in size and are more feasible in laboratory experiments than for analyzing big data.

We will now discuss regularization in machine learning. Regularization involves analyzing the number of variables that are used to construct the model and inducing a penalty for each extra variable that is used. There are two types of regularization techniques in machine learning: L1 regularization and L2 regularization. These regulate the structure of the machine learning model. In L1 regularization, the model we get is one with as few predictors as possible. The reason for such a model is that it improves robustness. After all, when we have simple solutions, they are applicable to a larger number of situations as compared to complex solutions. In L2 regularization, the main point of concern is coefficient and predictors, and the regulators keep the level of variance between these two elements to a minimum. This is because if the variance between these two overlaps, then we cannot clearly understand each of the predictor's impact. By keeping the level of this overlapping variance to a minimum, we are essentially increasing the clarity. In other words, the main function of regularization is to keep the model from overfitting (due to the use of too many features).

Validations are very important in the machine learning process because these determine whether our model can work with real-life problems or not. In other words, validation determines the worth of your model, as a model that cannot solve problems or perform tasks it is built to is not worth anything.

Predicting new observations

If we have performed the first three steps of the machine learning process successfully, then we will end up with a model that can generalize well to new data. Model scoring is defined as using a machine learning model on unseen data.

Model scoring is done in two steps:

1. Preparing a data set with features that are defined by the model.
2. Using the model on this prepared data set.

Conclusion

While developing any program, or any system for that matter, there is always a sequence to be followed so that the data is captured efficiently and interpreted properly, and this was what this chapter was all about. We also had a walkthrough of the modeling procedure, which consists of feature engineering and selecting a model, training the model, validating the model, and testing the model on new data.

In the next chapter, we will learn how reinforcement is related to financial markets.

Chapter 8

Reinforcement Learning for Financial Markets

Introduction

Reinforcement learning is an interesting and dynamic field. Reinforcement learning is a goal-oriented learning process focused on engagement with the environment. Reinforcement learning is stated to be the hope of real AI. And understandably so, for the promise reinforcement learning carries is huge.

Reinforcement learning is gaining popularity quickly, creating a wide range of learning algorithms for diverse applications. Therefore, it is also important to be acquainted with Reinforcement learning techniques.

Structure

In this chapter, we will cover the following topics:

- Problem types in machine learning.
- Identifying key predictors (data reduction).
- Learning from experience (reinforcement learning).

- Reinforcement learning algorithms.
- Types of reinforcement learning.
- Applications of reinforcement learning in real life.

Objective

After studying this chapter, you should be able to do the following:

- Understand business problems wherein machine learning can be applied.
- Understand the data reduction techniques.
- Understand the concepts, types, and applications of reinforcement learning.

Problem types in machine learning

Machine learning algorithms that look for patterns in massive amounts of data are used to build intelligent systems and AI applications. As a subfield of AI, hundreds of machine learning algorithms have been developed over the past few decades, each of which can be used to directly or indirectly generate new business knowledge. Part of the challenge is to determine which algorithm, or family of algorithms, is appropriate for your business task.

In general, there are five main things you can do:

- Find patterns with unsupervised learning.
- Identify key predictors through data reduction techniques.
- Identify irregularities or anomalies.
- Learn from examples with supervised learning.
- Learn from experience with reinforcement learning.

For example, consider credit card fraud detection. Because fraud patterns change over time, using machine learning algorithms to automatically detect emerging patterns can enhance a bank's capabilities to defend its customers and respond faster than if humans had to monitor all the transactions manually.

You can conceptualize past transaction data as a spreadsheet with millions of rows, where each row (an observation) represents a transaction. The columns might contain information like the date, time, latitude and longitude of the transaction, the amount, the merchant, and perhaps even what was purchased. There is another column that gives us the answer to the question of fraud, telling us whether the transaction is legitimate or fraudulent. This column exists because the bank has had enough time to verify the validity of questionable transactions with the account holders. An unsupervised approach might cluster the transactions into two groups to see if you can separate the legitimate from the fraudulent. In a supervised or semi-supervised approach, the algorithm looks for the similarities and differences between legitimate and fraudulent transactions, because it knows (at least to some extent) which are which.

Identifying key predictors (data reduction)

Sometimes, the data you have to process is too unwieldy and overwhelming. Imagine the same credit card transaction data set considered earlier, but instead of just information about the transaction amount, location, and characteristics, there are also a few hundred other variables. Which variables should you use to determine whether a transaction is valid or fraudulent?

If you use all the predictors, there are two risks. First, the model-building process may be computationally intensive, meaning that it may take more time or more processing power to build the model than you have available. Second, by building a model with too many predictors, you run the risk of modeling the noise instead of the signal, an outcome called overfitting. An overfit model has fantastic predictive power on the data you used to train it, but it won’t do as well on new data it hasn’t seen before.

Common methods for data reduction (also called dimensionality reduction) include principal component analysis (PCA), linear discriminant analysis (LDA), and autoencoder-type neural networks. Although they are not expressly adaptive, reactive, or proactive (and thus are more appropriately classified as statistical models than machine learning models), applying them to commercial problems tends to require significant processing power. They are often applied before sending data through unsupervised or supervised machine learning algorithms, so are sometimes included in machine learning texts.

Both computational performance and the accuracy of the resulting machine learning model are improved when you use the right features and reduce their number so that only the best predictors are used. Reducing dimensionality can increase the ultimate power of your model.

Learning from experience (reinforcement learning)

Reinforcement learning (RL) is a type of unsupervised learning that defines potential rewards for making different choices instead of labeling each observation with a defining characteristic. While supervised learning builds a model based on known information, reinforcement learning dynamically explores an environment to discover it. Reinforcement learning can be used to determine the best ways for an intelligent agent to interact with its environment.

For example, you may want to find an optimal path from a location in a building to the nearest fire exit. If you associate each outdoor location with a high reward, areas near exits with a small reward, and areas deep within the building with zero reward, reinforcement learning can identify the optimal paths that generate the highest reward.

Dowling and Cahill (2004) claim that R reinforcement learning L may be the most useful and applicable technique for solving problems in industrial environments. Here are some examples:

- RL can be used to learn customer preferences by observing behavior, which can provide information that is critical for marketing departments. *Halperin* (2017) demonstrated how the method can be used to design marketing strategies for new products and services, and devise pricing strategies that are tuned to the competitive environment.
- Students learn better when concepts are presented in logical ways. *West et al.* (2019) used reinforcement learning to find optimal learning paths for a curriculum, which could improve higher education and training outcomes while enhancing student satisfaction.
- Improving the quality of a sound source, especially in a noisy environment, is important for hearing aid manufacturing. *Koizumi et al.* (2017) explored this problem using reinforcement

learning by defining reward as an increase in the perceived quality of the source.

- Challenging medical conditions like sepsis require doctors to quickly identify treatment policies. *Raghu et al.* (2018) used reinforcement learning (supplemented by several other supervised and supervised approaches) to identify how clinicians could use models to substantially improve mortality.

Reinforcement learning involves building a model by letting the algorithm explore the system of observations on the basis of the defined rewards, make mistakes, and try over and over again. The approach is very similar to what organizations do when they set policies to support innovation, even more so because reinforcement learning seeks to maximize rewards over the long term. Although it requires large amount of data to be effective, reinforcement learning has been used to develop many wildly successful game-playing Ais, and it addresses a different type of the problem than the other machine learning approaches.

Reinforcement learning involves learning the data used to give feedback that will determine how the model adjusts to the prevailing conditions in order to achieve the target. The system on its own is able to evaluate the model performance on the basis of the feedback provided and then make necessary adjustments. For instance, the concept is aptly applied in self-driving cars (AI) and also in chess master algorithms.

Reinforcement learning as indicated above occurs in an interactive environment. In other words, the learner is in no way taught what to expect but finds out on their own according to the consequences of their actions, a concept known as goal-oriented learning. The major reinforcement learning components that we will cover in this section are listed below:

- Fundamentals of reinforcement learning.
- Reinforcement learning algorithm.
- Reinforcement learning environment.
- Various reinforcement learning platforms.
- Applications of reinforcement learning.

Fundamentals of reinforcement learning

Assume we want to train a dog how to catch an object thrown into the air. The dog may not necessarily understand all the concepts and theories involved here. The best way to accomplish this is to throw the object in the air, and every time the dog catches it, be ready to reward it. If the dog fails, then there is no reward. The dog will eventually analyze what actions transpired until it was rewarded. The reward will motivate the dog to repeat the same actions for more and more rewards.

Likewise, in the reinforcement learning environment, you may be able to teach the model what to do, but based on positive rewards, the model will tend to repeat certain actions. The agent or the model is more responsive to positive rewards and forgets actions that led to negative rewards.

Sometimes, we have scenarios of delayed rewards or rewards that are only given out when the task is done. There might be rewards or no rewards at each step just to confirm if there are mistakes.

Assume you are to teach a robot on how to navigate without hitting a mountain:

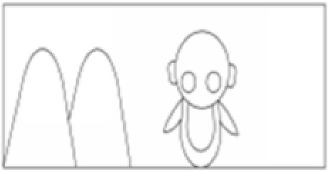


Figure 8.1

One rule would be to take away 10 points when the robot hits a mountain and gets stuck. This way, the robot will understand that, whenever it hits a mountain, there is a negative reward and therefore will not hit the mountain again:

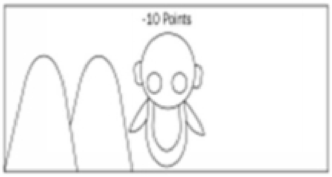


Figure 8.2

The other rule would be to give the robot 20 points each time it navigates in the right direction. Of course, the robot must try to maximize the rewards by remaining on the right path:

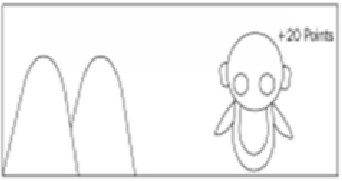


Figure 8.3

The model, or the robot, can apply two basic concepts of reinforcement learning to navigate smoothly and get more rewards. The reinforcement learning model can decide to explore different possibilities so as to get more rewards. And in this case, there is a higher possibility that the agent will poorly perform since it might utilize the wrong cautions. The other way is to exploit only the previous actions that resulted in positive rewards. When the model follows the path of exploiting only the known previous actions, there is a greater possibility of it missing out on the best actions though it will receive rewards. And unfortunately, it is impossible to perform both exploration and exploitation at the same time.

Agent

An agent basically is a software program that is capable of making intelligent decisions and is a learner in reinforcement learning. They make decisions and act according to the environment of interaction, plus rewards that follow:

- **Agent–environment interface:** Being intelligent software, the agent is able to perform a function at time, *t*, and then move from state *At* to the next state labeled *At+1*. Again, according to the action, the agent is given reward R.
- **Model:** This represents the agent's learning environment. The learning here can be either through model-based or model-free learning. For model-based learning, the agent is free to exploit all the previously learned information before making a decision. In model-free learning, on the other hand, the agent entirely relies on trial and error to perform the required action. For example, when you want to move to college from home, you either use the foreknowledge of the routes to college or just try different routes and finally choose the fastest method.

- **Policy function:** A policy basically is what controls how the agent behaves in the environment. The behavioral nature of the agent depends on the type of policy it is operating under. From our previous example about getting to college from home, the various routes represent the different possible policies. Some routes may be very short, while others are long. The routes are called policies since they are the guidelines towards achieving your ultimate goal. The symbol is always used to denote a policy.

  Most policies are always represented in a lookup table or as a complex search process.
- **Value function:** There is always a value function for each and every state of the agent. This function indicates how profitable it is for the agent to be located in a particular state. The function is dependent on the policy and is denoted by v(s). The value function represents the total expected rewards received by the agent right from the initial state. The various types of value functions are as follows:
  - **Optimal value function:** Gives the highest value for all states when compared to the other value functions. And thus, it is the optimal policy that has the optimal value function.
  - **Stationary policy:** This kind of policy involves action-distribution returns that are entirely dependent on the last two states visited by the agent.
  - **Deterministic stationary policy:** This is a policy that can deterministically select an action of the agent on the basis of its current state.

Algorithms for control learning

Various reinforcement learning algorithms are discussed below:

- **Criterion for optimality:** A learning agent is expected to make mistakes during its first moments of getting into the environment. So, the agent is supposed to receive discounted rewards if it is to achieve optimal goals. The policy set here should be one that expects a minimum return from the agent itself. There are three policies that are totally involved in obtaining optimality: optimal policy, deterministic stationary policy, and stationary policy.

- **Brute force:** This reinforcement learning algorithm includes the following processes:
    - o For any possible policy there are sample returns accrued to it.
    - o The policy with the highest expected returns is selected.

  There are various issues with this algorithm. The number of policies can be extremely large and therefore difficult to manipulate. Another possible issue would be a difficulty in dealing with a policy of varied returns. Imagine a policy with such a large variation on its returns. Getting the average value of returns per policy may be tedious at times.
- **Value function approaches:** This kind of algorithm tries to get the maximum returns by maintaining a set of estimated returns for some particular policies, either current policy or optimal policy. A policy that achieves maximum returns from the initial state may be deemed best for application.
- **Direct policy search:** Another possible way of determining the best policy involves searching directly from a set of policy space. The two methods used here are gradient-based and gradient-free methods. The gradient-based search involves mapping from finite-dimensional parameters to specifics where the policy space is located.

Applications of reinforcement learning

Among the vast areas of application for reinforcement learning, some are still under investigation and yet some are already implemented:

- Large-scale empirical evaluation.
- Predictive state representation.
- Lifelong learning.
- Dopamine-based learning in the brain.
- Part of the model for human skill learning.

Reinforcement learning algorithms

There are three techniques for the implementation of a reinforcement learning algorithm:

- **Value-based:** Here, the agent is anticipating a long-term return of the prevailing states under the policy, and so you ought to maximize the value function.

- **Policy-based:** Under this reinforcement learning scheme, you endeavor to find a policy such that the action executed in each state leads to maximal reward in the future. The policy-based method is further classified into deterministic, where the policy produces the same action for any state, and stochastic, where every action has a definite probability determined by the stochastic policy. The stochastic policy is *n{a\s) = P\A, = a\S, =S].*
- **Model-based:** In this case, you are expected to generate a virtual model for every environment, where the agent learns how to perform in that very environment.

Types of reinforcement learning

Two types of widely used reinforcement learning are as follows:

- **Positive:** This is an event triggered by specific behavior. It positively influences the action taken by the agent. This happens by enhancing the frequency and strength of the behavior. This method helps you capitalize on performance and sustain change for a longer period. Even so, you have to be careful as over-reinforcement may cause state over-optimization and impinge on the results.
- **Negative:** It involves strengthening behavior prompted by a negative condition that should have been dodged or stopped. Although it helps define the least-stand performance, this method set the minimum behavior, which is a drawback.

Applications of reinforcement learning in real life

- Data processing and machine learning.
- Planning of business strategy.
- Robotics for industrial computerization.
- Aircraft and robot motion management.
- Creation of customized training systems for students.

Conclusion

With the technological advancements in the financial market, learning about the system as a whole will always be a constant process. We discussed the concept, types, and applications of reinforcement learning. Reinforcement learning is a machine learning method with three algorithms: (1) value-based, (2) policy-based, and (3) model-based. The two types of reinforcement learning are (1) positive and (2) negative. Reinforcement learning should not be used for problem-solving when you have less amount of data. A major drawback of reinforcement learning is that the learning speed may be affected by algorithms parameters. We discussed a business problem wherein machine learning was applied, gave an overview of data reduction techniques.

In the next chapter, we will review some use cases in finance.

CHAPTER 9
Finance Use Cases

Introduction

We are living in the age of technology. When was the last spell you went into a store that did not have Paytm? These emerging innovations have rapidly become a significant fragment of our everyday lives.

And not only at the personal level, these new innovations are at the heart of every financial organization. Money processing or conversion of funds has been very smooth thanks to several choices such as POS machines, UPI, Internet banking, credit cards, ATMs having unfailing systems running.

Structure

In this chapter, we will cover the following topics:

- Technology and finance.
- Automation.
- The impact of FinTech.
- Guidelines to live by.

- Innovative technologies.
- AI as a strategy at the top level.
- Development status of different AI technologies.
- Risk management.
- Fraud detection and prevention.
- Improving the truth of financial rules and designs.
- Trading.
- AI in banking.

Objective

After studying this chapter, you should be able to do the following:

- Understand different use cases of machine learning in the finance industry.

Technology and finance

Finance has always been a field at the forefront of technology. That's because that's where the edge is—that's where the money is.

If you are going to allocate technology to industries with high degrees of ROI, finance would be the right place, because the business of money is at the position of fulcrum when it comes to monetizing information.

Of course, I'm not talking about material nonpublic information or anything like that. I" talking about performing market analysis, identifying data relationships, exploiting correlations, and using technology to build a better mousetrap. Because a little edge can turn into a lot of money when the business of finance is involved.

This is why technologies that we will see in the future of finance are already here. Many of them are accelerating the pace and ease of transactions, reducing costs, and increasing access.

Overall, the main takeaway for the future of finance is that any new innovations are likely to be adopted more rapidly by finance than by almost any other industry. Because something that works is likely to generate big financial returns. Further, even though the trends of the next decade and beyond have already been set in motion for the future of finance, there are likely to be new innovations that pop up. And finance is likely to take up the mantle quickly.

It was a lesson that banks had to learn this cycle.

FinTech is the digital barbarian at the gates, threatening to erode margin, and the banks and traditional financial services firms are finally onboard and pushing innovation forward. This is why many of the big banks started technology accelerators and incubators several years ago. After all, FinTech solutions pose a threat to traditional financial services. And if FinTech was going to eat financial services' lunch, then financial services might as well be pushing FinTech forward. As the saying goes, if you aren't at the table, you're on the menu.

This means that the future of finance is likely to continue to involve traditional financial services firms using their accelerators and incubators to foster innovation. And it means that there is no going back. The future of finance is FinTech in all of its permutations, and it is driven by three critical levers.

Automation

Yes, the robots are coming for low-skill, low-income, and low-education jobs. But they are coming for other jobs too. I learned several years ago that they are coming for mine. And if you work in finance, they may be coming for your job too.

You might not hear the robots coming

The first time I heard the word FinTech, I was at the Atlanta Federal Reserve Bank's Financial Markets Conference on Amelia Island in May 2016. At this annual meeting, which I have attended nine times, about 150 of the world's top economists are invited to join regional Fed bank presidents, government regulators, academics, and often the chairman of the Federal Reserve to discuss the hottest economic, monetary policy, and fiscal policy issues of the day.

Against a backdrop of this prestigious conference, a Fed reporter, whom I have known for years, and I skipped out of some sessions to enjoy the beautiful early May Florida weather. My friend was with another reporter whose specialty was FinTech. At the time, I had not yet heard of FinTech. So I asked naively, "What's that?" The reporter told me FinTech was “like Bitcoin and stuff like that.” I knew Bitcoin was a digital currency, so that was that.

I didn’t think too much of this conversation until several months later when I tried to hire a salesperson for Prestige Economics. I had difficulty finding candidates. Highly qualified people were exiting the space in droves, and I didn’t know why. Finally, one senior salesperson told me that everyone was getting out of financial market research because of FinTech. Essentially, robots were disrupting the research business. After being told that FinTech was disrupting my own business, I decided to learn as much as possible about it by taking a FinTech course at MIT. In short, the robots had been coming for me, and I hadn’t even known it.

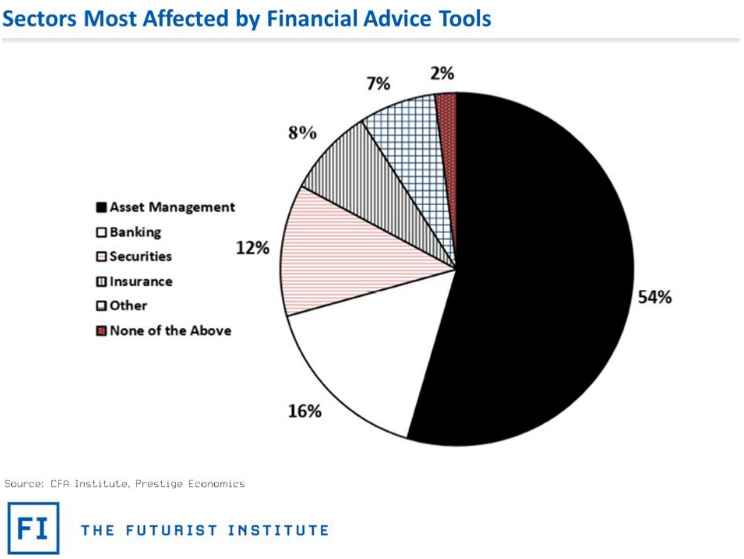
FinTech: Robocalypse comes to finance

FinTech is a buzzword for financial technology, which represents a host of businesses that are designed to disrupt (and eat the lunch of) traditional financial institutions. FinTech companies generally reduce costs, reduce complexity, or increase ease of use for transactions that had previously been the domain of banks.

FinTech is affecting financial services, and awareness has been spreading. Asset management has long been dominated by computers, statistical analyses, and programming. And FinTech has been disrupting asset management, often with passive trading strategies. Some of these strategies are known as robo-advising, owing to their automated (i.e., robot-like) nature. And the result? Asset managers are losing their jobs, and the disruption potential for asset management is very high.

In the movie “Wall Street,” Gordon Gekko asks Bud Fox, “Ever wonder why fund managers can’t beat the S&P 500?” Well, with the advent of exchange-traded funds (ETFs), fund managers and retail investors can just buy the S&P (and other indices), which is what they have done. A number of these ETFs are very liquid.

The impact of FinTech on financial advice is as shown below:



Source: CFA Institute, Prestige Economics

Figure 9.1

Passive asset management techniques and robo-advising are often easier and cheaper to administer than active asset management. These strategies can be implemented at significantly lower costs than active asset management strategies because they no longer require human asset managers. Plus, there is an economy of scale when computer programs do all the strategy work, analysis, and planning, as well as all of the buying and selling of securities. And passive asset management has also been adopted by finance and trading because these fields have historically embraced technology, with many firms using a black box model, algorithms, and technical trading strategies for many years.

The number of ETFs is as shown below:

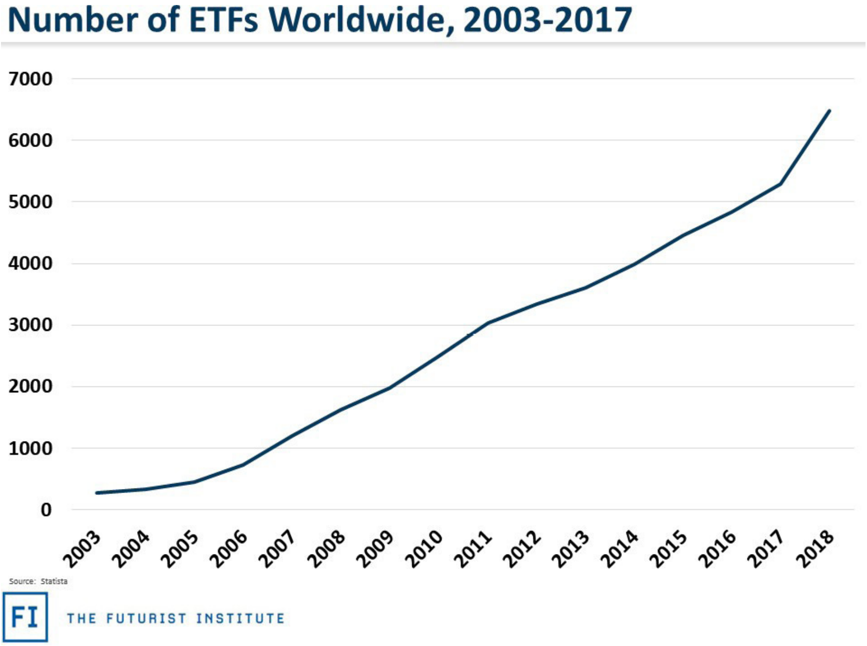


Figure 9.2

Expensive items (like market research) would also no longer be part of the budget since decisions are made by computers. After all, trading programs do not read words. But they really like lines—especially lines above (or below) which the price of a traded security has consistently stayed for a long time.

Technical trading has become more important, so analysts have been trying to add value by knowing what lines and supports matter most for the computers in different markets. This is why there has been a significant uptick in the number of financial professionals pursuing the Chartered Market Technician®, or CMT®, program. I completed the CMT® in 2016, and it focuses exclusively on these kinds of technical trading dynamics. Essentially, you are looking for computers in the market. I incorporated these technical trading dynamics into my forecasting long ago, which is why I have long expected that these kinds of trading dynamics will become increasingly important as passive asset management and robo-advising continue to grow.

The impact of FinTech

According to the PwC global report, FinTech's most powerful weapon is the disintermediation of the system. Cutting-edge innovations and strategies in FinTech demolish the need for a middleman. Until now, in a conventional system, banks and financial services did the tough work of managing your money and assets. All the transactions used to occur via banks and monetary institutions. However, FinTech lets in customers to deal with the money themselves, peer to peer. There's no needs to pay expenses and provider fees to an intermediary entity like banks. Clients can directly have interaction with any business and service, and pay them in advance through the latest FinTech like Bitcoin or similar cryptocurrencies.

So, what is the future of banks and financial institutions? What will happen to the existing financial products and services? It is now clear that, with the emergence of FinTech and virtual currencies, disruption in customer banking and fee sectors is imminent. Insurance and asset management sectors are also on the list of endangered species. Studies show that up to 28% of businesses may be at risk by 2020, and they all are banking and payment offerings. In the PwC survey, a senior executive at a global banking organization said, "We thought we knew our customers, but FinTech's really known our customers."

Customer banking and capital markets are going to be affected by FinTech the most. The emergence of transparent online platforms enables people and organizations to lend and borrow cash at once from each other. Innovative ideas on lending strategies and credit risk modeling algorithms deliver the customers powerful alternative credit models and risk calculation strategies. FinTech's made the complete technique more customer-centric, and due to the fact, there are no intermediaries and the operating fee is very low.

In the latest years, FinTech has created many online and mobile apps for payment methods. Formerly, even transferring credit score from one account to another required paying the middleman. However, now, virtual wallet apps and secure payment techniques have allowed alternative ways for transactions. Increasing the usage of virtual and digital money via smartphones is speeding up the entire technique. Additionally, we see that Internet penetration is rising globally, and people are buying products online more often. Their trust in FinTech is on the rise. Within the UK and many parts of the EU, there are shops that be given nothing but bitcoins for currency.

FinTech has significantly modified the fee infrastructure. As actual-time payments are very famous these days, digital wallets are becoming more beneficial than conventional credit cards. Cryptocurrencies like bitcoins are more convenient than credit cards, and there are not part of the banks. On top of that, the underlying Blockchain generation can track the movement of money and assets more effectively and reliably than the conventional method, which takes days to verify. In a volatile market where your cash is not safe, Blockchain allows you the safe keeping of your cash and offers you overall transparency, free of feasible fraud or hacking.

In many of the underdeveloped and developing areas of the finance sector, FinTech in payment is helping human beings drastically. Mobile technology and FinTech have created this new connected economy, or a form of shared economic system, that has made the economic system more organized. With the help of FinTech, you can join others who've money and individuals who need cash directly through peer-to-peer lending and crowdfunding. This allows one to invest cash in a low-price but surprisingly visible and transparent way.

It absolutely seems like FinTech is literally swiping the economic offerings industry. It is remodeling how corporations and customers interact with each other. Traditional banking models had been following the same system for more than forty years, and their methods have not been modified that much—tons of papers have been replaced by Excel sheets. William Lynne, director of the broking company Hybridan, concurs that the system is largely outdated, but it is very difficult to trade because of the global nature of the commercial enterprise and all its systems.

FinTech is now seeking to sort out a number of the troubles that the traditional systems cannot resolve, and it is beginning to make enhancements. On a traditional system, the processes to complete transactions continue to be identical. Where it took days, or maybe weeks, to confirm and verify a big transaction, the present-day FinTech strategies most effectively take a few minutes. FinTech appears to pose a risk to conventional business models. In the coming years, SME banking and client banking in Russia are predicted to be affected heavily with the aid of FinTech.

The wealth and asset management offerings enterprise, which concerned many huge firms previously, is now dealing with extinction because of the automation of asset allocation and innovation in alternative digitalized brokerage services. The consumer base for such styles of FinTech offerings is getting larger and larger each day as more and more people use online offerings or mobile apps to do these kind of tasks themselves. PwC research found that 74% of all of the insurance corporations are expecting that their services will be disrupted by FinTech over the following five years, and 51% of asset and wealth managers said their offerings might be disrupted because of today's user-friendly wealth management apps and online services.

Thanks to FinTech, the entire process is now self-directed, and anybody can, without problems, confirm profiles via secure encryption methods. But we additionally see developing investments on FinTech in regard to how customers invest cash and buy coverage. The existing system is already being reinvented. Over the past four years, investments in FinTech within the insurance quarter have extended beyond $3.4 billion, according to Denovo.

FinTech is said to develop through evolution, no longer through revolution. Current corporations and business models aren't going to exchange overnight—not even organizations like Apple or Google, who are understood to be the trendsetters of business models. The opportunities and influences of FinTech are large, and there will usually be a few winners and some losers within the technique of exchange. As it's for a new emerging generation that's developing new markets every day, the possibility of some groups and startups failing at some point in the system may be very excessive. Any kind of surprising growth, whether it's technological or financial (in this example it is both!), creates inevitable pitfalls, and there are a few agencies that are certain to fail. One such organization is Monetise, who started off on a very note, gaining masses of exhilaration; however at the end of 2015, their share fee fell approximately 90%!

On the other end of the spectrum is PwC. They are very quiet, so you no longer hear a great deal about them inside the marketplace, but they are doing quite well. They are a profitable organization that makes use of FinTech to control risks, money laundering, or verifies a customer's identity. By 2015, their share rate increased by 60%. There are organizations that are getting the most out of FinTech. Mi-Pay, who developed a mobile payment system, created a superb anti-fraud software program and is doing nicely. Additionally, there's Vipera PwC, who already exceeded its own projection and expectancies of the marketplace.

Another phase of FinTech is the emergence of effective data analysis algorithms. The automation of the entire system makes it easier for buyers to look for particularly tailor-made merchandise and make more dependable and powerful decisions. Also, the trend of crowdfunding and improvements in lending and equity transfer makes it easier for individual investors to inspect possible opportunities that have been formerly inaccessible to them, like a business property.

One of the honestly huge impacts of FinTech is the upward thrust of robo-advisors. Don't fear—it's not a few robot revolting or AI like Skynet taking over the world. Robo-advisors are an automatic machine or investment service, and they are gaining popularity day by day. Robo-advisors are doing what registered investment advisors (RIAs) used to do. They're backed by the latest behavioral algorithms and tracking systems that predict the behavior of the market. Robo-advisors are less expensive than RIAs and are very smooth and easy to deal with. Besides, many robo-advisor user interfaces make the entire process of investing fun and interesting.

Currently, there are large players—Betterment, Wealth Front, and the six hundred-million-dollar agency future guide, which was later acquired by way of the wealth management giant BlackRock—who are thoroughly exploiting the FinTech industry and flourishing by turning innovative ideas into money. Such huge investments in FinTech are pushing conventional monetary advisors to rely more on young traders and create innovations on their own. That's why Vanguard and Schwab launched their very own robo-consultant platforms that can interact with the consumer. A combination of automation and statistics can do what a conventional RIA does, so now it seems the role of an RIA is out of date. Formerly, only the handiest wealthy humans could find the money for such wealth control services, but now, even a middle-profits family can find the money for such online services and robo-advisors to manipulate their property and wealth more sophisticatedly than ever before.

The disruptive power of FinTech is nowhere more apparent than in the realm of global finance. This is both a good and a bad thing. It is good because there is considerable potential for development here, but it also poses an inherent risk to many jobs. In fact, one report by Citigroup estimated that almost 30% of employees in the global financial sector are in danger of having their jobs eliminated by FinTech. Why is that? Well, FinTech is about interconnectivity. As such, it is easy to imagine that many of the jobs are based solely on facilitated communication between various global markets that just do not need to exist. These jobs can be easily overtaken by technology that is not only more accurate but also much faster than having humans communicate with one another.

FinTech offers accessibility to everyone with a reliable Internet connection—no sophisticated infrastructure necessary. In a rare turn of global events, it is not the banks of the US and North America who are dominating the development in this sector. China has embraced investment in these technologies at even higher levels. Banks in Africa and South East Asia are well aware of the potential for making lucrative profits in these markets. It is an uncharted territory and fair game for everyone.

Guidelines to live by

One of the troubles with FinTech is that the app needs to function in a way that verifies the identity of the operator. Otherwise, how can people trust in the veracity of the service they are operating if they cannot be sure they are exchanging money with legitimate individuals who have solid credit backgrounds? This principle is called "know your customer (KYC)." The problem with the KYC approach is that it takes a lot of effort to correctly determine a person's identity electronically.

Here are some tips for simplifying the overwhelming task of getting to know your customers:

- **Don't ask for all the information at once:** This is time-consuming, and people will be suspicious of your service if you go for it all at once. The best way to collect customer information is to start with the bare minimum and then allowing them to complete the first transaction. After trust in your service has been established, you are free to get the rest of the necessary information.
- **Make it easy:** Of course, you could ask your customers to scan and upload copies of their identity documents, but that is a huge hassle for them. Partnering with other identity servicing companies to gather this information reduces the burden on your customers. Using the information they have provided, these services will attempt to match photos from their database. In the event that this fails to work out, you can always ask the customer for a scan.

- **Give them reasons:** People are naturally skeptical about offering up personal information, especially when it seems hard to understand what their gender could possibly have to do with their finances. This is why it is best to always explain why you are asking for certain information. These exchanges have to be based on trust, and the best practice for establishing trust is transparency.
- **But don't make it too easy:** Even Venmo has come under fire for their security failures. These technologies have to be simple to interact with, but not so simple that they can be easily hacked. Customers have to feel safe using your app, which is why simple features such a security pins are essential. This is where FinTech user experience differs from that of other applications. Unlike photo-sharing apps, messaging, etc., there is an element of security in the design that must be present, which is not required for other platforms.

Innovative technologies

While the whole world was focused on the Internet, mobiles, social media, and the rise of FinTech, new innovations have shaken up the entire banking industry one more time. While in 1970 to 1990 there was a big debate whether computers and the Internet shall make manual workers obsolete, now there is a larger debate whether robots will make Internet workers and manual workers obsolete? Debates are endless and they shall continue, but let us study the best breed of new technologies such as AI, machine learning, Blockchain, and big data in this chapter.

AI/robotics including chatbot

Let us now focus on the use of new innovations such as AI/robotics and chatbot in banking. Many times, banks struggle to identify a genuine customer because of the huge amount of data at their disposal. In such cases, AI is very useful. For example, previously robots could answer questions such as "What is my current balance in the account?" "When will the check book be delivered to me?" "What is the status of my fund transfer request?" However, new-age robots are becoming more analytical, and can now answer questions like "What is the 360° view of all my family accounts?" "How is my investment pattern?" "Where should I invest now?"

Cognitive computing

Cognitive computing systems are designed to solve problems the way humans do. So how is this different from AI? In cognitive computing, the system provides information to help the expert decide, whereas, in AI, the system guides the expert on the best course of action to be taken.

Blockchain in the banking industry

Nowadays, we hear a lot about the Blockchain concept using cryptography technology. What is Blockchain? In simple words, it can be explained as data entry done in multiple databases at the same time. For example, when a seller raises an invoice on a buyer, that invoice data can be stored in the database of the buyer, seller, buyer's bank, seller's bank, and any related government agencies, and all of this can be done in milliseconds. Moreover, this transaction is safe, transparent, and can be encrypted, so that no third party can intercept it. In banking, this can be very well used for various trade finance transactions, especially when there are multiple parties involved at the same time, and it brings in more trust and transparency among the multiple stakeholders.

What is Blockchain?

Blockchain is a new technology where a digitized, decentralized, public ledger of all transactions is maintained using cryptography. Each node (a computer connected to the network) gets a copy of the blockchain, which is downloaded automatically.

Too difficult to absorb? Ok, let's make our life simpler and consider a case study of a typical international buying-selling (called trade finance) transaction involving multiple parties or stakeholders:

- Seller (exporter, call him/her X).
- Buyer (importer, call him/her Y).
- Seller's bank.
- Buyer's bank.
- Central bank of exporter's country (e.g., Reserve Bank or Federal Bank).
- Central bank of importer's country (e.g., Reserve Bank or Federal Bank).

- SWIFT (Society for Worldwide Interbank Financial Telecommunication), which records every international banking transaction.

Assuming an invoice has been raised by X on Y of US$100,000. Assume that a copy of the invoice should be available with different parties. Can you imagine online entry of US$100,000 in the books or computerized systems of seven different stakeholders from different countries at the same time by a computerized system? Can a standard computerized program and infrastructure pull this off? Absolutely not. Even if we use Internet-based e-commerce, this may not be easily possible. But the Blockchain technology makes it possible.

And the advantages are huge:

- Extremely safe.
- Transparent.
- Trustworthy.
- Very easy record maintenance.
- Low cost of maintenance.
- No intermediaries necessary in the whole operation as parties directly communicate with each other.

In fact, recently there was a huge banking fraud in India by a famous diamond exporter, Nirav Modi, and many finance experts feel that the fraud could have been avoided simply if Blockchain had been used. Many leading institutions and banks have started experimenting with Blockchain in their operation. For example, Australia's stock exchange will replace its registry, settlement, and clearing system with Blockchain technology to cut costs for customers.

Cryptography was used in Bitcoin when the new virtual currency was created. Bitcoin has got good as well as a bad name in the international market, but people are amazed by the Blockchain technology used in Bitcoin. Therefore, the world may see huge investments in Blockchain technology in near future, if not in Bitcoin.

First blockchain user in the world: Australian Securities Exchange

To give some different industry examples of Blockchain, Australian Securities Exchange (called ASX) has become the first major stock exchange to announce the adoption of Blockchain technology for its stock settlement activity. ASX will use cryptography to record shareholdings and manage the clearing and settlement of equity transactions. This will enable huge savings to the cost of operations, bring more transparency, and reduce the need for intermediaries.

Internet of things (IoT) in banking

Let's discuss the Internet of things (IoT) for banking. But first, what is the Internet of things? In layman's terms, it's a machine-to-machine or device-to-device connectivity (e.g., connected cars, connected homes, connected malls, etc.). In other words, all the computers, whether they are at home or in the mall or in the bank or say vehicle, shall talk to each other and help the customer. For instance, a homemaker will not have to worry whether, in the current week, he or she has sufficient groceries at home. In fact, home computers will detect any shortage and inform the mall for the supply of that grocery while connecting with the bank to pay on the customer's behalf. Everything is truly automated. And here, banks will play a major role. Not only can they arrange for payment, but also based on the ready data, they can give targeted offers, advice, and rewards to their customers.

Bank of America: Big data projects

So, Bank of America is using big data projects in various areas, e.g., improving customer retention, improving time and accuracy of financial forecasting, etc.

Voice biometrics

Voice biometrics is playing a major role in the life of people as well as banks.

"Voice biometrics" for a good cause -
Proof of life for pensioner's using phone call

Photo credit (of a lady) : -
Artist: - Jackie Weisberg
Website: -
https://www.flickr.com/photos/jackiew/4399678775/meta/

Figure 9.3

Nuance Corporation developed a voice biometrics application for BBVA Bancomer recently, to strengthen their pension system by improving the process of Proof of Life verification. This means a pensioner can identify him- or herself with just a simple phone call, and once the computer recognizes the voice, it will be considered as evidence that the pensioner is alive. How cool is that for our grand-pops and grannies, isn't it?

Digital bank

By now, we have seen many innovations in banking such as desktop/ mobile Internet, FinTech, social media in banking, AI, robotics, machine learning, Blockchain, and big data. However, is it enough for banks to concentrate only on these fancy names? Is that what makes a bank fully digital? Absolutely not. the CIO of a bank has a huge task of enabling the entire digital landscape at the front end as well as backend. A typical midsize bank in the US may have anywhere between five hundred to nine hundred different IT applications, and the amount of interactivity with the changing technology landscape is a mammoth task for IT operations. In the next segment, we shall take a glimpse at the illustrative areas that may impact our digital world.

AI as a strategy at the top level

Worldwide, executives are increasingly looking to AI to create new sources of business value and revenue streams, or opportunities to save costs. This applies especially to top AI adopters, those who have invested in AI initiatives and have achieved impressive results. This is a small group of firms, but they are doubling their AI investments.

Banking as a sector has competitors not only amongst its peers but also from FinTech, mobile industry, and tech industry. With the way technology is evolving so rapidly, we are not really sure which competitor will be added to the list and when.

Banks can save from automation by retiring old technology and investing in innovations. For any bank to survive and grow, the most important criterion is increasing the bottom line. Apart from gaining market share, banks also look at cutting costs especially for tasks that are mundane and repetitive. Squandering human resources on such tasks just adds to the costs of running the business. If we can compare the banking industry to investment banking and asset management companies, they are currently in a similar predicament of generating alpha and reducing costs. It's a bitter truth for banks that the quaint banking days are over. AI-based technology is not a fad that will pass, and banks can choose to ignore or look the other way for a few years more, only to realize that they are fading away or they should have opened up their doors sooner to AI.

Visionary leaders from the banking and financial services industry can foresee the future and steer their institutions in that direction. The exponential growth of AI has not gone unnoticed. These leaders already know that, in a few years, AI will be an integrated part of their systems. However, to succeed, the AI strategy should be a top-down approach, although for deciding the best use cases, it should have a bottom-up approach. This can happen when the workforce remains assured that AI is being adopted as an enabler and not to replace the human workforce. This communication is equally important to get the organization as a whole to work towards making the AI journey worthwhile.

But how many are ready to take the leap? Are the CEOs who are already way ahead in adapting AI reaping benefits from it? These are the most important questions at the moment.

Like any business decision, AI should have a board-level strategy. It calls for the strategy from top-level leadership. AI is not one of the technologies banks usually adopt, wherein there is no executive-level involvement. If executed well, it can result in cost savings for the banking industry. As per the Autonomous report on AI, by 2030, traditional financial institutions can save 22% in costs. The report states that banks can save $490 billion by adopting AI in the front office.

With pressure, building upon executives to reduce costs and increase revenue by investing in AI makes a strong case for it. It requires executive-level strategy, and in absence of that, all AI investment will fall flat on its face.

Why we need executive-level indulgence in AI is because of the following factors.

Cost of investing in AI

Undoubtedly, financial institutions will reap the benefits of AI, but the cost of investing in AI is very high and there are no two ways about it. If costs are huge, the stake is high. Half-baked execution will not only dent the bank's image, it will also cost a fortune. AI journey should not be undertaken like injecting a suite in the current IT system. Just infusing AI without a strategy will not yield the desired results.

Crucial leadership involvement

According to the recent PwC US supplement to the 21st Annual Global CEO Survey 2018, US CEOs' fears over the past five years of losing a technological edge have risen sharper than other types of threats. The fear is rational considering the leap technological innovations make every day. Banks cannot afford to lag here. However, if fear is rational, then rational solutions can help overcome this impediment. This clearly calls for a more proactive role from CEOs rather than just assigning AI journeys to the CIO or CDOs. It is very important to chart out the AI journey rather than just joining the bandwagon. The journey for each bank or financial institution will be different. There is no fit-for-all solution in the AI journey. And such enterprise-specific roadmaps are what will actually make the journey profitable for the bank.

So, how do banks adopt AI? Some banks have set up an in-house AI division. Singapore-based OCBC Bank (Overseas Chinese Banking Corporation Limited) set up an in-house AI division in early 2018, while some are collaborating with AI companies who develop products for banks.

Inorganic growth

Another route for banks to enter in the AI domain is to invest in AI startups. There is now a growing trend amongst banks to acquire FinTech startups to keep up in the ever-changing landscape of technology. In 2017, JPMorgan Chase acquired WePay, a US-based small-business-focused payments company to provide its four million small business clients with WePay's payments technology. In 2016, Ally Financial, a US financial services company, acquired Trade King, a digital wealth management company. In early 2018, Canada-based TD bank bought its first technology firm, an AI-based startup, Layer 6. The main focus of the bank was to provide customers with personalized services.

Banks have to compete with other industries that are providing personalized customer experiences as well as financial services. Customers expect the bank to anticipate their needs and preferences. As banks move forward in the digital bank journey, there will be a vast amount of digital data available with banks. AI can transform isolated customer interactions into a seamless thread of customer experience.

The banking industry, which is facing competition from FinTech, eCommerce companies, tech giants, and telecom companies, is willing to explore and implement AI with two underlying motives: one is reducing the cost of serving customers, and second is improving customer experience through personalization.

Banks can adopt AI either in partnerships, acquisitions, or in-house development. There is no one strategy that fits all financial institutions.

Development status of different AI technologies

Take a look at the following image:

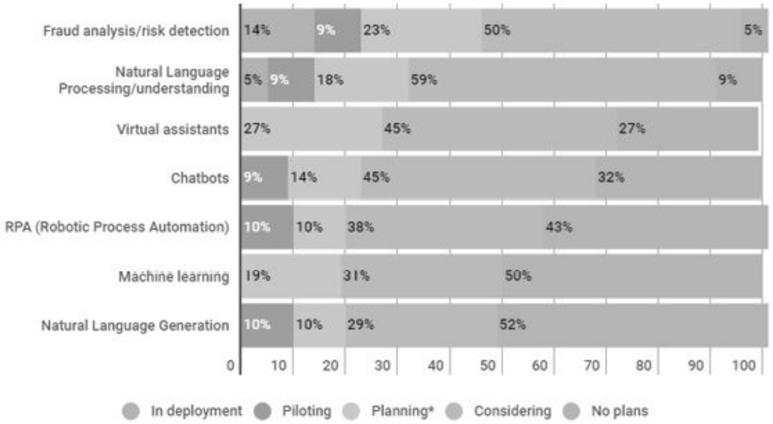


Figure 9.4: *Understanding the Investment into AI in Banking, 2017 (Source: Celent)*

"Planning" in the above chart implies that a bank has a funded project for implementation within the next twelve months. As per some of the recent trends, what we can observe is that banks do not have a clear roadmap when it comes to investing and adopting AI. Supervised machine learning can be the first stepping stone for novice banks who want to embark on the journey. It is the most common type of machine learning. Banks can also explore collaborative methods.

One way is for banks to look at established use cases for AI adoption. Each bank has to initiate its own journey based on its business strategy. Where is the bank standing on its technology threshold? Does the bank already have data analytics in place? If not, those are essential stepping stones to start the AI journey.

The average gain of implementers of "must do" use cases over low implementers is shown here:

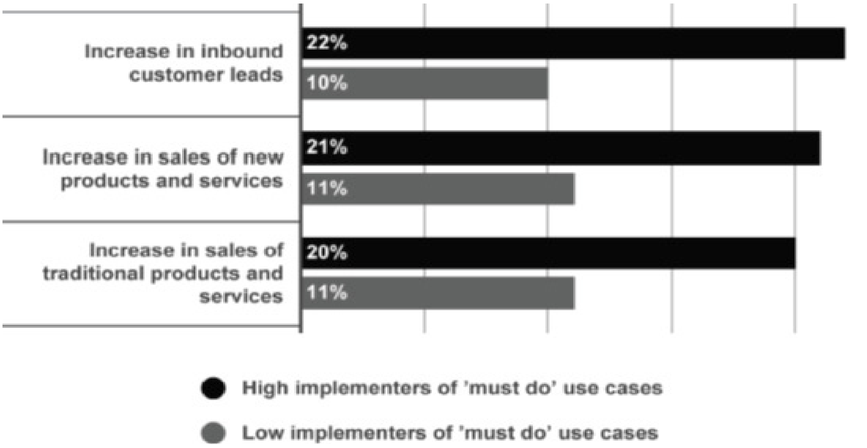


Figure 9.5: *(Source: Capgemini)*

A recent survey by Capgemini clearly shows that the "low hanging fruit" approach is showing gains for banks especially in sales and customer acquisition.

Banks need to chart their own path, pace, and destination on the basis of their resources and business plans. If a bank wants to start an in-house AI, then it will need vision, hire skill sets, technology incubation, and significant time from development to deployment. Another option is external help from AI startups who understand the bank's culture and needs and can accordingly look at developing the necessary AI model. Banks are also investing in AI startups, and this makes the startups part of bank's enterprise, thereby reducing the risk associated with third-party vendors while sharing sensitive bank data. Also, it brings in a culture of innovation in banks. And for the AI startup, on their part, they will now consider the bank as a part of their organization rather than viewing it as a client.

So, the bottom line is that banks need to see the cost benefit of AI. They are skeptical about investing huge amounts in AI without reaping benefits either in customer retention, cost-cutting, or customer onboarding. Banks do need to realize that, whether they want to adopt or not, AI is the technology of the future and all its competitors are adopting it. And when we say competitors, we are talking about FinTech companies who are way ahead in experimenting and adopting AI. It is just a matter of time when AI becomes a way of banking.

Banks at the top level need to have an understanding of various AI technologies—machine learning, predictive analytics, NLP, robotic process automation, and smart virtual assistants—and where these can be applied in bank processes, ranging from customer service to risk management. Banks need to understand applied AI; using these technologies in applied solutions will need a strong foundation of data analytics.

Risk management

It often becomes difficult to handle financial tasks owing to high security and performance requirements. To begin with, while implementing AI in the financial sector, data scientists and machine learning engineers have to perform in-depth research and evaluation. Financial information systems have enormous amounts of data, and the system should be capable of handling tons of transactions without any delay. Algorithms designed for the financial sector are highly reliable and are tested to deliver 100% results in any scenario. Moreover, they can also manage both structured and unstructured data, along with quick identification of potential failure causes.

Fraud detection and prevention

AI and machine learning models have the capability to detect fraudulent activities and stop unwanted operations in real time. In recent years, banks and financial institutions have implemented AI models into their systems and databases. Fraud detection systems have the capability to analyze the client's location, behavior, and credit history in real time.

Other fraudulent activities like money laundering can also be avoided through AI models. Machines are designed to put an end to alleged money laundering schemes and protect financial and government institutions from any kind of unwanted loss.

Improving the truth of financial rules and designs

Machine learning has a substantial effect on the finance industry. Some of the advantages of machine learning include trading, loan underwriting, and portfolio management fraud detection. According to the report "The Future of Underwriting," machine learning facilitates data evaluation for assessing and discovering nuances and anomalies. This assists in improving the accuracy of principles and units.

Trading

AI and machine learning models are based on the concepts and theories of statistics. Trading is greatly expanding in the stock markets all over the world, and traders are getting fast and the best predictions and recommendations using AI tools. Intelligent trading systems have the capability to monitor structured and unstructured databases. In a structured database, we can have data from spreadsheets or other databases, whereas, in the unstructured category, the data is taken from sources like social media or news.

Trading algorithms are based on AI models and make predictions and calculations depending on the market conditions. Machine learning algorithms are trained on past data and trade history, which allows them to deliver accurate insights and predictions. Moreover, the technology has greatly improved the validation process, which has helped traders get information regarding the trade without any hassle.

In the banking sector, AI powers smart chatbots that provide real-time answers and solutions to customers. Voice control assistants have now been introduced that have self-education features and are getting smarter each day. There are different AI apps and tools that offer personalized financial advice and support individuals in any financial problem. These intelligent tools can track income, recurring expenses, and calculate spending habits as well. On the other hand, AI finance tools can also be used to obtain an optimized plan and get financial support.

AI in banking

Tech giants tend to hog most of the limelight when it comes to cutting-edge technological advancements. But the financial sector, including the stodgy banking incumbents, are showing increasing interest and signs of adoption of AI. The banking or finance industry has a profound impact on virtually all consumers and businesses with a direct effect on the country's economy, so seeking insights and keeping up to date with the convergence of financial technology and AI is critical for every business. The banking industry is putting its money on AI-based solutions to address several traditional banking problems such as providing quality customer service to their millions of customers, fraud prevention, mobile technology, and data security, among others.

Here are some popular AI-based trends in banking:

In the 1960s, in an effort to coordinate the bookings made by travel agents across the US, the travel industry pioneered large-scale computer-based airline booking systems. Most travel companies are now collecting data from users, and with the right AI tools, they are able to provide customers with relevant future travel programs based on their past searches. The use of AI pricing tools that can autonomously adjust flight prices depending on the market demand, weather, and other determining factors is helping airlines optimize their business.

When prospective travelers browse, shop, book, fly, or even change their travel plans, companies collect this data, which is analyzed by AI-powered solutions, to gather actionable insights and tailor their services to the customer's needs. The competition in the travel industry is fierce, with the savvy travelers always comparing and looking for prices of flights, hotels, buses, trains, and car rentals to get the best out of their buck. AI has immense potential to alter the travel experience with personalized services and efficient travel experience.

Conclusion

Finance is at the forefront of technology, and we discussed why. We also discussed why machine learning should be adopted in finance. We also had a walkthrough of machine learning impact on the finance industry. For establishments employed in the finance industry, it has become more and more critical to keep up with the opposition and improve their standing in terms of technology.

In the next chapter, we will learn the impact of machine learning on FinTech.

Chapter 10

Impact of Machine Learning on FinTech

Introduction

Although it is clear that the traditionally conservative financial organization was not at the forefront of the growth in machine learning, FinTech machine learning is now a popular word. It delivers a new standard of operation for financial analysis, customer care, and data management.

Structure

In this chapter, we will cover the following topics:

- Overview of FinTech companies.
- Impact of technology.
- Challenges.

Objective

After studying this chapter, you should be able to do the following:

- Understand the impact of technology on FinTech companies.

- Understand the challenges faced by FinTech companies.

Overview of FinTech companies

Financial technology, or FinTech, comprises businesses that use technology to ensure efficiency in financial services delivery. The companies are usually startups that disrupt existing financial systems and challenge established financial institutions that do not rely on technology.

FinTech solutions can fall under any of the five areas: banking or insurance sector, business processes support, targeted customer segment, business-to-business interaction, or market position solutions.

There is an increase in global investment in this type of industry since 2008. In fact, in 2014, it reached at least $12 billion from $930 million in 2008. The rapid growth over the years showed a 40% increase in the workforce in London in the financial and technology services. Popular FinTech companies include Nutmeg, Funding Circle, and TransferWise.

In the US, various FinTech startups include Betterment, Affirm, Fundera, Behalf, Lending Club, IEX, Plaid, Money.net, Robinhood, Prosper, Square, SoFi, Wealthfront, and Stripe. In 2014, European FinTech firms received $1.5 billion in investments. London-based FinTech companies received $539 million, while the FinTech companies in Amsterdam received $306 million. FinTech companies in Stockholm received $266 million in investment.

In the Asia-Pacific, Sydney became the new FinTech hub in April 2015. Stockspot and Tyro Payments are strong players in the city, and being a new hub can accelerate the growth of FinTech. In Hong Kong, an innovation lab fosters disruption in financial services through technology. In the Philippines, VMoney is growing.

The Financial Data Science Association organized its first event to include members from the fields of NLP, machine learning, and AI. The organization hopes to create a research community around investment statistics and computer science.

In October 2014, the Wharton School of University of Pennsylvania founded Wharton FinTech to connect thought leaders, academics, investors, and innovators for the purpose of reinventing global financial services. The University of New South Wales and the University of Hong Kong Law School published a research paper about the evolution and regulation of FinTech.

Forbes magazine came up with a list of leading FinTech disruptors for its Global Fintech 50 for Forbes 2016. The UK Treasury commissioned a report in February 2016 comparing the leading FinTech hubs. California ranked first for its talent and capital, while the UK first ranked for government policy. New York City got first ranking for demand.

Impact of technology

FinTech increases customer satisfaction. In December 2015, its customer satisfaction was higher by 8% compared to banking. Established financial advisory companies like Fidelity Investments collaborated with FutureAdvisor, a FinTech startup, to allow new technology within a custodian. Celebrities like Jared Leto, Snoop Dog, and Nas invested in Robinhood, a FinTech startup.

Finance is highly vulnerable to the disruption caused by software. Financial services do not have concrete products. They have information. Although regulations shield finance, for now, FinTech startups are beginning to shake up the global banks. Established financial institutions weathered the dot-com crisis. On the other hand, regulations in areas of money transfer and the Bank Secrecy Act pose an ongoing threat to FinTech startups.

Challenges

Aside from traditional competitors, FinTech firms face challenges from financial regulators. Data security is another concern. The threat of hacking and the necessity to protect sensitive corporate financial and consumer data are problems each FinTech startup must face. Even a small breach of data can have a disastrous effect on the reputation of a FinTech company.

Another consistent challenge is the increased extortion attacks of the online financial sector. Further, many FinTech companies face challenges in marketing because larger competitors spend more money on advertising. In fact, even established financial institutions face security challenges because of their online customer services.

Conclusion

Financial analysis is a huge topic in accounting, and when done with the aid of machine learning, it will greatly help in coming up with a more concise profile and data to help the current manpower.

In the next chapter, we will try to understand machine learning in finance.

Chapter 11
Machine Learning in Finance

Introduction

Procedure computerization is one of the most widely recognized utilization of machine learning in finance. The innovation enables to supplant manual work, computerize redundant errands, and increment profitability.

Subsequently, machine learning empowers organizations to advance expenses, improve client encounters, and scale up administrations. Here are the robotization use cases wherein machine learning, reinforcement learning, and deep learning are applied:

- Chatbots.
- Call-focus computerization.
- Paperwork computerization.
- Gamification of representative preparing.

Structure

In this chapter, we will cover the following topics:

- Machine learning use cases in banking.
- Security.
- Guaranteeing and credit scoring.
- Algorithmic exchanging.
- Robo-advisors.
- Utilize outsider machine learning arrangements.
- Applications of machine learning.

Objective

After studying this chapter, you should be able to do the following:

- Understand machine learning use cases in the banking sector.
- Understanding application of machine learning in security, credit scoring, algorithmic trading, and robo-advising.

Machine learning use cases in banking

The following are a few instances of procedure computerization in banking:

JPMorgan Chase propelled the Contract Intelligence (COIN) platform that uses NLP, one of the machine learning methods. It helps them make reports and extract important information from there. Manual audit of 12,000 yearly business credit understandings would ordinarily take up around 360,000 work hours, but machine learning enables auditing a similar number of agreements in only a couple of hours.

BNY Mellon incorporated procedure mechanization into their financial system. This advancement is in charge of $300,000 in yearly investment funds and has realized a wide scope of operational enhancements.

Wells Fargo utilizes a machine-learning-driven chatbot through the Facebook Messenger to speak with clients and provide help regarding passwords and records.

Privatbank is a Ukrainian bank that executed chatbot collaborators over its versatile and web stages. Chatbots accelerated the goals of general client inquiries and permitted them to reduce the number of human associates.

Security

Security risks in the finance sector are expanding alongside the increasing number of exchanges and clients. Also, machine learning algorithms are astounding at identifying fakes. For example, banks can utilize this innovation to screen a great many exchange parameters for each record continuously. The calculation analyses each move a cardholder makes and reviews whether the endeavored action is normal for that specific client—such a model spots fake conduct with high accuracy.

When the framework recognizes suspicious conduct, it can demand extra identity proof from the client to approve the exchange, or even obstruct the exchange through and through if there is in any event a 95% likelihood of it being a fake. Machine learning algorithms need only a couple of moments (or even split seconds) to review an exchange. The speed forestalls fakes continuously, instead of simply spotting them after the wrongdoing has been submitted. Money-related observing is another security use case for machine learning in finance. Information researchers can prepare the framework to identify numerous micropayments and banner such tax evasion systems as smurfing.

Machine learning algorithms can altogether improve organization security as well. Information researchers train a framework to spot and disconnect digital dangers, as machine learning is best in class in breaking down a large number of parameters and constants. What's more, odds are this innovation will help the most developed cybersecurity organizations achieve self-discipline in the near future. Adyen, Payoneer, PayPal, Stripe, and Skrill are some outstanding FinTech organizations that invest intensely in security machine learning.

Guaranteeing and credit scoring

Machine learning algorithms fit superbly with the guaranteeing errands that are so normal in accounting and protection.

Information researchers train models on a large number of client profiles with several information passages for every client. A well-prepared framework would then be able to play out the equivalent endorsing and credit-scoring errands in genuine situations. Such scoring models help human representatives work a lot faster and more precisely.

Banks and insurance agencies have an enormous number of chronicled customer information, so they can utilize these sections to prepare machine learning models. Further, they can use data sets created by telecom or service organizations.

Algorithmic exchanging

In algorithmic exchanging, machine learning settles on better exchanging choices. A numerical model screens the news and exchange results constantly and identifies designs that can cause stock costs to go up or down. It would then be able to act proactively to sell, hold, or purchase stocks as indicated by its expectations.

Machine learning calculations can break down a large number of information sources at the same time, something that human brokers can't in any way, shape, or form accomplish. Machine learning algorithms help human merchants crush a thin advantage over the market. Further, given the huge volumes of exchanging tasks, that little advantage frequently converts into critical benefits.

Robo-advisors

One major facet of the FinTech revolution is the advent of robo-advisors. The first of the robo-advisors was founded in 2008 when a major financial crisis went into full swing. The initial idea was to rebalance investor assets while also giving them a modern high-tech way of dealing with their investments.

Prior to 2008, wealth management software was only sold to traditional human financial advisors who took advantage of the ability of technology to automate the work they'd otherwise have to toil away at. However, after 2008, the product was sold to consumers without any middleman or human financial advisor in the middle.

What robo-advisors do is provide portfolio management advice and take the place of traditional financial advisors. The robo-advisor services that consumers use in the post-2008 era utilize the same software and algorithms as do traditional advisors. As a result, robo-advisors are able to, at the most basic of levels, take the place of human advisors, at least in terms of rebalancing assets and accounts. What they aren't able to do is get involved in the more personal aspects of wealth management like taxes or retirement (at the moment, at least).

Thanks to these relatively small shortcomings, robo-advisors are being seen as a major alternative to traditional advisors for an assortment of rather obvious reasons. The principal reason for this is robo-advisors, of course, are automated. And because they're automated, they cost a fraction of what it would to staff or otherwise pay a human financial advisor. On top of that, there's the possibility that investors will actually see higher returns using robo-advisors than they would see otherwise. This is because, in addition to the cheaper fees, they tend to have a number of exclusive features like an automatic rebalancing of portfolios or tax-loss harvesting.

Bear in mind that robo-advisors aren't entirely abstracted from humanity. Many robo-advisors are actually just online advising firms that take advantage of automated technologies and programmed algorithms in order to take a lot of the hassle out of wealth management, as well as make it a cheaper option overall. However, there are certain robo-advisors that actually offer access to a dedicated human financial advisor who will give you a living, breathing ear to bounce off ideas and help you manage your wealth altogether.

There are also robo-advisors for different purposes, such as those for weekend investors and people who spend a large amount of time actively investing or otherwise taking part in the stock market. Every robo-advisor has different features, different account minimums, and different trade commissions.

Not everyone has the same niche or the same needs when it comes to selecting a robo-advisor.

Utilize outsider machine learning arrangements

A machine learning architect can execute the framework by focusing on your particular information and business space. The expert needs to extract the information from various sources, change it to fit for this specific framework, get the outcomes, and picture the discoveries.

An all-inclusive machine learning algorithm doesn't exist yet. Information researchers need to alter and tweak calculations before applying them to various business cases in various areas.

Applications of machine learning

Machine learning had already found application in finance before the emergence of efficient chatbots, mobile banking apps, and search engines. Owing to the high volumes, the required accuracy of historical data, and the quantitative nature of the financial world, few other industries are more suitable for AI. Now you can find more instances of machine learning in the financial sector than ever before. It's a trend accentuated by better computing power and more accessible machine learning tools like Google's TensorFlow.

Machine learning has arrived, and it plays a critical role in modern society and in many financial areas. It is used in loan approval, asset management, and risk assessment, but very few tech-savvy people have an accurate picture of the number of ways machine learning finds its way into people's financial lives.

Current financial applications

Here are some examples of machine learning used in the world today. Keep in mind that some applications use multiple AI technologies or approaches and not just machine learning.

Portfolio management

"Robo-advisor" is a term that was not heard of a few years ago but is now widely used in the financial world. However, the term is a bit misleading, because no robots are involved at all. Instead, robo-advisors (e.g., Betterment or Wealthfront) are machine-learning-based algorithms and built to design a user's financial portfolio, including their goals and risk tolerance. For example, users enter goals like retiring at age sixty-five with $ 3,000,000 in savings. They also enter their age, current financial status, and income. The advisor, more accurately referred to as an allocator, then spends the investment across different asset classes and financial instruments to reach the user's goals. The advisor system then calibrates to the changes in the user's goals and to the actual changes in the market, making it the best fit for the user's goals. They have become of great interest to consumers who do not need physical advisors to be comfortable with investments and those who are unable to pay fees to human advisors.

Algorithm trading

Algorithm trading dates back to the 1970s and is also referred to as automated trading. It uses difficult AI systems to make very fast trading decisions. The algorithmic systems make thousands or millions of transactions in one day. Therefore, the term HFT (high-frequency trading) is used, and it is part of algorithmic trading. Most financial institutions and hedge funds do not disclose the AI approach they use for trading. However, it is believed that deep learning and machine learning are playing an increasingly important role in trading decisions. There are some limitations to using machine learning in trading shares.

Detecting fraud

The system can detect abnormal behavior or unique activities by using machine learning and marking it for the security department. The biggest challenge for this system is to prevent false positives and situations where risks are flagged when there are actually no risks. There is a myriad of ways in which security breaches can occur, so real learning systems will become a necessity in the next five to ten years.

Insurance or credit insurance

Underwriting can be described as the perfect job for machine learning in the financial sector, but there is a lot of concern in the market that the machines will replace many of the current underwriting positions. This is especially a problem with large organizations such as large banks and limited liability companies. Machine learning algorithms can be trained in millions of consumer data cases such as occupation, age, and marital status. It can also be used for insurance results and financial loans to check whether a person is in default, or fails to pay, or get their loans in time, or has been in a car accident.

The underlying trends can be assessed using algorithms and continuously analyzed to detect trends affecting lending and insurance for the future. This way you can check whether more and more young people are having car accidents. Or has there been an increase in the number of defaults under a specific demographic in the past ten years? The answers to these questions are of great benefit to the organizations. However, this is currently limited to large companies that have the resources to hire data scientists and who have a massive amount of data (past and present) to train the algorithms.

Machine learning and cryptocurrencies

Trade supported by AI and machine learning has attracted enormous interest in recent years. There is a hypothesis that the inefficiencies in the cryptocurrency markets can be used to create big profits. The normal trading strategies supported by the state-of-the-art machine learning algorithms are much more capable than the standard benchmarks. Some non-trivial, but actually simple, algorithms can help anticipate the short-term evolution of the cryptocurrencies market.

The success that machine learning techniques had with stock market predictions suggests that the methods can also be used effectively to predict cryptocurrency prices. But applying the machine learning algorithm to the cryptocurrency market is mainly limited to analyzing Bitcoin prices using the Bayesian neural network, random forests, long- and short-term memory neural networks, and some other algorithms. These studies anticipated Bitcoin price fluctuations to some extent and concluded that the best results could be obtained by using algorithms based on neural networks. The deep learning enhancement was able to beat the performance of buy-and-hold strategies in predicting the prices of twelve different cryptocurrencies over a one-year period. There were other attempts to use machine learning to predict the prices of cryptocurrencies other than Bitcoin, but they came from non-academic sources and did not yield any comparisons for the results.

Day trading with machine learning

The speculation in securities is called day trading. More specifically, it refers to buying and selling financial instruments on the same trading day. Strictly speaking, it is a trading event within one day. It means that all positions are closed when the market is closed for the day. The day traders look at identifying the entry and exit positions on the shares with favorable conditions. These conditions yield various small-term gains that can add up to large gains.

If there are people on the market who can recognize favorable patterns in the market, we can even train a machine to perform in the same way. This is the purpose of using machine learning for day trading. But first, we need to identify the strategies that day traders use to signal market entry conditions. The technique is split into two processes: first is a high-level pattern description, and the second is machine learning.

In the first process, input semantics are identified, which occurs for potentially hundreds of predefined strategies. This is done using robust and highly scalable pattern matches, such as Apache Flink. Once a data cartridge has been activated, we can go through the historical data and find the past cases where cartridges were activated and what the outcome price was after ten or twenty minutes. We can generate a training example for the algorithms using machine learning to create probability distribution over previous entries.

Conclusion

In this chapter, we discussed the application of machine learning in finance. We also had a walkthrough on how banks and financial institutions utilize Credit Scoring Models to analyze clients' credit scores using algorithms. We also discussed the application of machine learning in cybersecurity, algorithmic trading, and robo-advising.

In the next chapter, we will learn how to avoid fraud.

Chapter 12
eKYC and Anti-Fraud Policy

Introduction

Technological innovation has driven civilization to a unique age of corruption. News about whipped credit cards and identity theft has been a regular phenomenon.

Fraud identification is one of the areas that has achieved a major boost in achieving reliable and superior outcomes by AI action. It is one of the main fields throughout the banking sector where AI programs have fared exceptionally well. Fraud identification has come a long way and is predicted to advance in the coming years.

Structure

In this chapter, we will cover the following topics:

- Big data analytics: True buzzword of today.
- How criminals obtain information for online banking.
- Common ways in which information can be stolen.
- Security measures.

Objective

After studying this chapter, you should be able to do the following:

- Understand different ways information can be stolen.
- Understand different security measures to stop this fraud.

Big data analytics: True Buzzword of today

Previously in the IT sector, we have seen large projects in data warehouse and business intelligence. Let's look at an important innovation used in the finance industry, big data, which is now analyzing tebibytes (1 gigabyte = 0.000909495 tebibytes) of structured and unstructured data including text, numeric, digital, algorithmic, video, and voice. As per ingrammicroadvisor.com, big data can be used in various segments of banking such as fraud detection, compliance, regulatory requirements, customer segmentation, personalized marketing, and risk management.

How criminals obtain information for online banking

Digital culprits and malevolent con artists are continually at work to devise better approaches to steal secret client information. Sadly, such assaults are not simply limited to the Internet. And as of late found credit/check card information stolen and these strategies are progressively done in the physical world.

One of the biggest banks in the world, Swift, has recently acknowledged that it has been the target of cybercrime in 2016 after fraudsters hacked into a communications system linking eleven thousand banks and stole cash from one of the banks by reversing links among the bank and Swift network.

Common ways in which information can be stolen

In this section, we will learn the most common ways in which data is being stolen by criminals or hackers.

Retail stores or restaurants

Con artists have formulated approaches to mess with card readers at retail locations. Malicious smart cards that look like genuine cards are embedded into the machines to make an installment. But the machine basically detects that a blunder has happened, and the retail vendor is uninformed of the harm that has just been done. Later when a legitimate installment is made with a genuine card over the same machine, details of that card get recorded.

Then the con artist returns to the store and supplements the fake card into the machine to make another apparently honest installment. Details about all the cards that have been embedded in the interim period are currently moved into the malignant card, which can be extracted by the con artist by means of a gadget.

Online portals

The best web shopping entryways compulsorily utilize secure strategies to ensure client points of interest. Unsecure entryways are vulnerable to programmers and can without much of a stretch lose information to tricksters. There have been a few examples where major online gateways have been ruptured and card details have been stolen and abused.

From a customer's viewpoint, it is prudent to know about SSL and security declarations. Moreover, the best Internet security programming introduced on a machine can likewise recognize false pages and entrances.

Hacked mail accounts

Numerous individuals get and pay their Visa bills by means of email. So if a programmer figures out how to access an email account, he can bring about a great deal of inconvenience. It is better to utilize different devices for email assurance like two-element verification, solid and extraordinary passwords, and so on. Mail administrators like Gmail and Yahoo offer different techniques to check whether your account has been hacked into and merits an investigation.

ATMs

Robbery of card information from ATMs is known as "skimming." This can be easily done with the assistance of a card perusing gear and a little camera that records a person when he punches in his PIN. Tricksters additionally utilize hardware that reproduces the attractive segment of cards. Be that as it may, steps can be taken to dodge such assaults, and this includes ensuring no one can see your hand while entering the PIN and monitoring suspicious-looking machines.

Pickpockets and thieves

Physical loss of cards is the greatest danger in this situation. It is prudent to report a card loss to the concerned bank and neighborhood powers at the earliest opportunity. Another essential precautionary measure is to screen card articulations and actions for any indications of malignant movement.

Employee records

Your employer has your personal data that a character criminal could access. To prevent this wrongdoing, ask your manager how your own data is put away. Be watchful for things you'd never anticipate.

Fake calls

The criminal calls you, claiming to be a rep from your Visa organization, requesting that you confirm some personal data. The criminal then contacts your MasterCard organization and poses as you . . . Please simply HANG UP!

Get back to the MasterCard organization using the number on the back of your card to confirm any potential issues. Never give personal data via telephone.

Social media

Your online profile may have all the data a hoodlum needs to take your character. Prevent this by erasing personal data. Offer responses to the security inquiries of monetary records that don't show up on your online networking pages.

Security measures

Credit/check cards are helpful yet perilous as con artists have thought of a few creative strategies to extract personal data. You are encouraged to stay careful and take proper measures to stay that way.

Some tips on card security

- Quickly sign the back of your card.
- Try not to let any other person use your card, NAB ID, or verification gadget.
- Continuously know where your card is, keep it safe from misfortune or loss, and watch out for home security.
- Be aware of card misrepresentation and security on the web. You don't need your character stolen.
- Bear in mind your card and receipt at the ATM. Tell your supplier promptly if your card is lost or stolen.
- Upon expiry, cut your card askew, down the middle (counting any implanted microchip on the card, attractive strip and security code).
- Try not to use an ATM in the event that you believe something's wrong

Security tips for Internet banking

- Keep security points of interest secret, and don't record them and store them with your cards.
- Retain your details on the off chance that you can. Banks ought to never request security details online, so disregard any messages indicating to be from a bank requesting that you enter your points of interest.
- Try not to give anybody a chance to see you entering your secret key or PIN.
- Tell your bank instantly on the off chance that you lose or overlook a secret word or PIN or on the off chance that another person uses it.
- Evade simple-to-figure combinations of numbers or letters or ones effectively identifiable with you, i.e., your birthday, portable, postcode, etc.
- Frequently change your PIN and passwords. Contact your supplier to see what security highlights they offer.

Conclusion

The financial sector is the most vulnerable, or should we say a hot target for fraud, simply because of money. Using an electronic system to get to know your client (eKYC) and security measures should be properly put in place to avoid fraud.

In the next chapter, we will learn about data mining and data visualization and their uses.

CHAPTER 13
Uses of Data Mining and Data Visualization

Introduction

Data comes from a variety of fields such as tracking devices and credit cards, which are being gathered since people feel it is useful. The challenge is to locate useful information buried in all this data. This is a daunting task, and this is where data mining comes into play.

In order to make data mining valuable and efficient, it is important to involve humans in the data discovery procedure. The concept of visual discovery in data mining is to depict raw visualization data. Humans will then obtain understanding, draw assumptions, and engage with the evidence.

Structure

In this chapter, we will cover the following topics:

- Data visualization.
- Data mining.
- Future health care.

- Education.
- Customer relationship management.
- Criminal investigation.
- Fraud detection.
- Customer segmentation.
- Intrusion detection.
- Lie detection.

Objective

After studying this chapter, you should be able to do the following:

- Understand the concept of data mining and data visualization.
- Understand the real applications of data mining and data visualization.

Data visualization

Data visualization is a method in which visual essentials such as charts and graphs are used to draw meaningful insights from the data. It is meant to show the facts behind the data and to further allow the user to see the structure of the data.

Data mining

Data mining is a method used by businesses to translate raw data into usable information. By using algorithms to search for trends in vast batches of data, companies may understand more about their clients in order to create more successful marketing campaigns, improve revenue, and cut costs.

There are several other uses of data mining and visualization, and they are as follows:

Future health care

Data mining and visualization have shown great potential in health systems. Health-care systems use data and analytics in identifying the best practices to be used to improve health care while looking for cost reduction options. Researches make use of data mining and visualization for approaching multidimensional databases, software computing, and machine learning. It can be used to predict the number of patients in different categories.

Education

Data mining and visualization can be used in the educational environment. It can be used in predicting a student's future through study behavior and the effects of educational support. It can also be used in advancing scientific knowledge centered on learning. Institutions can make use of data mining and visualization to make accurate decisions while predicting students' futures.

Customer relationship management

Customer relationship management is a good way of getting and keeping customers. It can be used to predict customer strategies and maintain a good relationship with them. With data mining and visualization techniques, collected data can be used for analysis, so instead of businesses getting confused about the focus points in customer retention, they can easily identify the problem and find a solution.

Criminal investigation

Criminology is a study of certain crime characteristics. The high volume of crime data sets can show the relationships between certain fields, and text-based crimes can be converted into word processing files.

Fraud detection

Fraud is a crime that has taken over billions of dollars. Although traditional methods of fraud detection are effective in some cases, they take a lot of time and are relatively complex. Data mining and visualization can provide a meaningful pattern by transforming information into insightful data. Any information that is valid in crime detection is considered as knowledge. It is important to note that fraud detection systems should be able to protect the information of all its users. Data mining and visualization aid in the collection of sample records, which are then classified as either fraudulent or nonfraudulent. Data mining, for example, is built to locate sources of data such as web scraping from different sources.

Customer segmentation

Customer segmentation, a very important part of business, should not be overlooked. The use of traditional market research can help in segmenting customers, but the use of data mining goes a lot deeper and helps by increasing the market effectiveness. Data mining can align the customers into a more distinct segment while tailoring to their needs according to their reviews. In business, the market has to do with getting and retaining customers, and this should be done in the most efficient way possible. With data mining and visualization, you can easily determine a customer's vulnerability and a business can offer them special features to enhance their satisfaction and keep them engaged and interested.

Intrusion detection

Intrusion is something a lot of organizations and firms undergo. This is basically actions that can compromise the integrity of a confidential source. The defensive measures employed to ensure this doesn't happen include user information authentication. Data mining can help in improving this intrusion detection by adding a high level of focus to anomaly detection. This way, data analysts can easily distinguish intrusive activity from more common and frequent everyday activity. Data mining can be used to extract data that is specifically relevant to the problem. Data visualization makes it a lot easier to understand the data for decision-making.

Lie detection

Lie detection can be difficult and apprehending criminals might be the easy part, but getting them to say the truth can be a lot more difficult. Law enforcement can use data mining and visualization techniques for investigating crimes and tracking communication between suspected terrorists or people. This includes the use of text mining and seeking out meaningful patterns in data, which usually come in unstructured text. For example, data samples that have been used in past investigations can be compared and a model lie detector can be developed. The data thus collected can be sorted according to the requirement of the case.

Data mining and visualization can be used for all of the above and possibly so much more. The bottom line is that they are important tools in today's world and are necessary for businesses and corporations to thrive the best way they can. Data mining can also be used to achieve data visualization as the case may be. The best part is, it gives rise to a lot of possibilities in the nearest future.

Conclusion

Data mining is necessary for improving an existing system. Also known as data or knowledge discovery, it can be used for analyzing data from every aspect. This data can be used in reducing costs, increasing revenues, and running a business or an organization.

Data mining is one of the tools widely used in analyzing data from different angles and dimensions while categorizing it and establishing relationships among the different types of data.

Data visualization is also a very crucial tool. It isn't just a representation of data in graphical formats, but it is incredibly powerful and the use of this tool can affect an organization both negatively and positively. When data visualization is not used effectively, it can affect decision-making and obscure whatever information you are trying to communicate. When used right, it can help you achieve more effective and powerful communication and aid great decision-making.

Data mining and visualization are both different ways of collecting and identifying data, and if used properly, they can make an organization grow.

In the next chapter, we will discuss the advantages and disadvantages of machine learning.

Chapter 14

Advantages and Disadvantages of Machine Learning

Introduction

Machine learning is a very strong instrument that has the ability to revolutionize the way things work. Machine learning can be used to overcome challenges and support businesses by making projections and allowing them to make wise choices. Each coin has two faces; each face has its individual property and characteristics. It's time to reveal both faces of machine learning.

Structure

In this chapter, we will cover the following topics:

- Advantages.
- Disadvantages.

Objective

After studying this chapter, you should be able to do the following:

- Understand the advantages and disadvantages of machine learning.

Advantages

Owing to the sheer volume and magnitude of the tasks, there are some instances where an engineer or developer cannot succeed, no matter how hard they try; in those cases, the advantages of machines over humans are clearly stark.

Identifies patterns

When an engineer feeds a training data set to a machine with AI, the machine will then learn how to identify patterns within the data and produce results for any other similar inputs that the engineer provides. This is efficiency far beyond that of a normal analyst. Because of the strong connection between machine learning and data science (which is the process of crunching large volumes of data and unearthing relationships between the underlying variables), through machine learning, one can derive important insights from large volumes of data.

Improves efficiency

Humans might have designed certain machines without a complete appreciation for their capabilities, since they may be unaware of the different situations in which a computer or machine will work. Through machine learning and AI, a machine will learn to adapt to environmental changes and improve its own efficiency, regardless of its surroundings.

Completes specific tasks

A programmer will usually develop a machine to complete certain tasks, most of which involve an elaborate and arduous program where there is scope for the programmer to make errors of omission. He or she might forget about a few steps or details that they should have included in the program. A machine with AI that can learn on its own would not face these challenges, as it would learn the tasks and processes on its own.

Helps machines adapt to the changing environment

With ever-changing technology and the development of new programming languages to communicate these technological advancements, it is nearly impossible to convert all existing programs and systems into these new syntaxes. Redesigning every program from its coding stage to adapt to technological advancements is counterproductive. At such times, it is highly efficient to use machine learning so that the machines can upgrade and adapt to the changing technological climate all on their own.

Helps machines handle large data sets

Machine learning brings with it the capability to handle multiple dimensions and varieties of data simultaneously and in uncertain conditions. A machine with AI has the ability to learn on its own and can function in dynamic environments, emphasizing the efficient use of resources.

Machine learning has helped develop tools that provide continuous quality improvements in small and larger process environments.

Disadvantages

It is difficult to acquire the data necessary to train a machine. The engineer must know what algorithm he or she wants to use to train it, and only then can he or she identify the data set they will need to use to do so. There can be a significant impact on the results obtained if the engineer does not make the right decision:

- It's difficult to interpret the results accurately to determine the effectiveness of the machine-learning algorithm.
- The engineer must experiment with different algorithms before he or she chooses to train the machine.
- Technology that surpasses machine learning is being researched; therefore, it is important for machines to constantly learn and transform to adapt to new technology.

Concepts involved in machine learning

Machine learning uses multiple concepts, and each of these concepts helps a programmer develop a new method that can be used in machine learning, and all these concepts together form the machine learning discipline.

Statistics

A common problem in statistics is testing a hypothesis and identifying the probability distribution that the data follows. This allows the statistician to predict the parameters for an unknown data set. Hypothesis testing is one of the many concepts of statistics that are used in machine learning. Another concept of statistics that is used in machine learning is predicting the value of a function using its sample values. The solutions to such problems are instances of machine learning since the problems in question use historical (past) data to predict future events. Statistics is a crucial part of machine learning.

Brain modeling

Neural networks are closely related to machine learning algorithms. Scientists have suggested that nonlinear elements with weighted inputs can be used to create a neural network. Extensive studies are being conducted to assess these elements.

Adaptive control theory

Adaptive control theory deals with methods that help the system adapt to changes and continue to perform optimally. The idea is that a system should anticipate the changes and modify itself accordingly.

Psychological modeling

For years, psychologists have tried to understand human learning. EPAM (Electronic Pilot Activity and Alertness Monitor) network is a method that's commonly used to understand human learning. This network is used to store and retrieve words from a database when the machine is provided with a function. In recent times, research in psychology has been influenced by AI. Another aspect of psychology, called reinforcement learning, has been extensively studied in recent times, and this concept is also used in machine learning.

AI

As mentioned earlier, a large part of machine learning is concerned with the subject of AI. Studies in AI have focused on the use of analogies for learning purposes and on how past experiences can help anticipate and accommodate future events. In recent years, studies have focused on devising rules for systems that use the concepts of inductive logic programming and decision tree methods.

Evolutionary models

A common theory in evolution is that animals prefer to learn how to better adapt to their surroundings to enhance their performance. For example, early humans started to use the bow and arrow to protect themselves from predators that were faster and stronger than them. As far as machines are concerned, the concepts of learning and evolution can be synonymous with each other. Therefore, models used to explain evolution can also be used to devise machine learning techniques. The most prominent technique that has been developed using evolutionary models is the genetic algorithm.

Conclusion

In this chapter, we discussed the advantages and disadvantages of machine learning in order to get a clear picture of the phenomenon as a whole.

In the next chapter, we will learn about the applications of machine learning.

Chapter 15

Applications of Machine Learning in Other Industries

Introduction

We are in the center of an uprising trend of machine learning applications. Machine learning is applied in self-driving cars, personalized banking, customer support queries, recommendation engines, LinkedIn recommendations, and the list keeps going on.

Structure

In this chapter, we will cover the following topics:

- General applications of machine learning.

Objective

After studying this chapter, you should be able to do the following:

- Understand in detail about general applications of machine learning.

General applications of machine learning

- **Financial analysis:** Machine learning is used to perform financial analysis for companies that have large volumes of accurate and quantitative historical data. As it is already being used for algorithmic training, portfolio management, fraud detection, and loan underwriting, future applications of machine learning might include chatbots for customer service, sentiment analysis, and security purposes. Machine learning models learn with experience and use data to make future financial predictions in a better way.

- **Predictive maintenance:** Corrective and preventive maintenance is a major part of manufacturing industries. Although this process is complex and expensive when conducted with conventional approaches, machine learning has now made it easier to discover meaningful insights and hidden patterns in factory data. Because this process helps in reducing risks associated with unexpected failures, companies can also reduce unnecessary expenses by implementing machine learning models. What's more, artificial intelligence and machine learning algorithms work in collaboration to analyze historical data and ensure workflow visualization.

- **Fraud detection:** Machine learning models are best suited for detecting spam and fraud. Spam filters are now designed by using the latest artificial intelligence and machine learning algorithms that include neural network rules for detecting phishing messages and spam.

  Major search engines such as Google and Bing work perfectly because the system is programmed to learn the ranking of pages and eliminate spam in real time. They are powered by effective machine learning algorithms. Also, there is a specific number of spam filtering approaches that are used by email clients to make sure that these spam filters are regularly updated.

  To perform such activities, a rule-based spam filtering approach is considered along with multi-layer perceptron and decision tree induction. Furthermore, system security programs that are designed using artificial intelligence and machine learning algorithms understand the coding pattern and stop fraudulent activities before they can cause any problems. It is predicted that online credit card fraud will reach a worth of $32 billion in 2020. So, this is a serious issue that must be solved quickly.

Fraud detection is one of the major applications of machine learning that helps make online transactions safer. The number of transactions from credit cards, debit cards, UPI, numerous wallets, and smartphones has been increasing with time and has also been a major target for online criminals. To make online transactions smooth and safe, the machine learning model thoroughly investigates the processes and searches for suspicious patterns. This approach in machine learning is also known as the classification problem and is handled in real time.

- **Image recognition:** Image recognition, also known as computer vision, involves machine learning, data mining, and database knowledge for discovery. Having the ability to produce symbolic and numeric information from high-dimensional data and images, the approach is widely used in various industries such as automobiles and health care. Moreover, image recognition helps in detecting criminals or suspected persons in real time.
- **Dynamic pricing:** Setting a fixed price for a service or a product is considered a traditional means of trading. Nowadays, pricing strategies are based on objectives and targets for which different kinds of promotions and discounts are given. Services such as air tickets or cab fares are dynamically priced depending upon various external scenarios such as traffic or the number of passengers.

  Artificial intelligence is now able to track buying trends and provide pricing solutions for determining competitive product prices. Further, machine learning algorithms are implemented to predict accurate service rates depending upon load and user demands.
- **Medical diagnosis:** Machine learning in the medical industry has supported health-care organizations and companies in great ways. Prediction and analysis in the medical industry are generally performed using machine learning algorithms as they work with patient records and data sets to yield relevant outcomes. By reducing medical costs and providing

highly accurate predictions, artificial intelligence and machine learning have provided effective treatment plans and better diagnostic tools in order to improve patients' health. Also, machine learning is now being used in health-care systems for identifying high-risk patients, recommending medicines, and predicting readmissions.

- **Google Translate:** Google Translate is yet another amazing application of machine learning that has helped people translate and understand the text of any language. Google has introduced the Google Neural Machine Translation (GNMT) system that is based on neural machine learning.

  With the ability to work on thousands of languages and dictionaries, Google Translate uses the natural language processing approach for finding the accurate translation of words and phrases. Other techniques, such as named entity recognition, POS tagging, and chunking, are also modeled in Google Translate systems to yield accurate translation results.

- **Customer lifetime value prediction:** Machine learning can be effectively used to derive meaningful business insights. Customer lifetime value prediction and customer segmentation are two of the major challenges faced by marketers. To handle the growing number of customers, companies have to update their systems with artificial intelligence and machine learning algorithms for better forecasting results.

  Thus, machine learning and artificial intelligence allow businesses to predict customer behavior and purchasing patterns. Further, companies can greatly improve their sales by analyzing customer purchase trends over the Internet.

- **Financial evaluation:** The finance industry has taken advantage of the benefits associated with artificial intelligence and machine learning models. Machine learning algorithms have the ability to perform in-depth financial analysis and are widely implemented for different procedures such as algorithmic trading, fraud detection, and portfolio management.

  The models have the capability to give an overview of an individual's past record and credit history in loan approval cases. Financial institutions and companies can also stop fraudulent activities, such as illegal transactions or money laundering, by using machine learning algorithms, as artificial intelligence and machine learning models have the ability to stop such activities and report them in real time.

- **Social media:** Social media platforms have millions of active users with tons of data stored in their storage systems. In order to handle big data analytics and prediction requirements at a larger scale, companies such as Facebook and Twitter have state-of-the-art data centers that are equipped with powerful machine learning and artificial intelligence algorithms. Taking the example of Facebook, we can notice that whenever a picture of you with a friend is uploaded, Facebook instantly recognizes the person and suggests a tag. This is made possible through face recognition machine learning models.

  Moreover, Facebook's "people you may know" feature is based on machine learning models that are designed to understand and make decisions based on experience. To make this happen, Facebook consistently takes note of the friends you are in contact with and makes suggestions accordingly.

- **Online customer support:** To cater to customer requirements, websites are now taking the help of chatbots to instantly answer their queries and present the most suitable solutions. Not all websites have deployed a customer support representative at the backend. However, in most cases, all you need is to talk with a chatbot. These chatbots are designed with effective machine learning models that learn with time and present the required information to customers instantly. The ability to understand customer queries and problems makes chatbots one of the best options to be considered for online businesses.

- **Product recommendations and guidance:** Online shopping has become the latest trend, and it is expanding globally. We can see that online shopping stores or websites often recommend suitable items through email or on the website. This activity is performed through intelligent machine learning algorithms that understand customer requirements and suggest options accordingly. Depending upon the market trends and behavior of the customers, the system is designed to learn and adapt to new changes for making accurate predictions. Factors such as brand preferences, website purchases, or liked items allow machine learning models to send relevant predictions and suggestions to the customers regularly.

- **Search engines:** Search engines such as Google and Bing are powered by machine learning algorithms to improve search results and rankings. Each time a search is executed by the user, algorithms at the backend deliver accurate findings and also learn from search results.

  Pages having regular and high amounts of traffic are displayed by the search engines at the top in accordance with the query. In this way, algorithms make use of search engine estimates and give improved search results. The algorithms also match user requirements and show pages that contain the most relevant information.

- **Self-driving cars:** Self-driving or AI-powered cars are currently the most impressive machine-learning-based technological invention. Based on the concept of unsupervised learning algorithms, self-driving cars work according to the rules of deep learning and crowdsource data from all similar vehicles and drivers.

  This application can guide a vehicle in case of emergency situations or drive on its own by receiving and using information from sensor data fusion systems. External and internal information sources such as Lidar, cameras, IoT, or radars feed the system with relevant data and allow the vehicle to operate on its own.

  In an autonomous car, one major task of the machine learning algorithm is to continuously observe the surroundings and environment to forecast changes. Generally, these sub-tasks include the detection of an object, identification of an object, and prediction of movement.

- **Traffic alerts:** Google Maps is the most popular application when it comes to location searching and getting directions. The application uses machine learning and artificial intelligence algorithms to find both the shortest and the fastest paths. Google Maps is being used widely by people from all over the world, as it provides accurate locations, best possible routes, average speeds, and predictions about traffic rates as well.

- **Government:** Government agencies are regularly in need of machine learning models because they have various data sources and need to oversee information regularly. By identifying useful patterns and insights, data can be analyzed in different ways to minimize cost and increase the efficiency of projects.
- **Cybersecurity:** Cybersecurity is one of the major factors that make the Internet safe for online transactions, data transfer, and information handling. Take the example of the finance industry, where cybersecurity systems are implemented to stop money laundering and distinguish between legitimate and illegitimate transactions that take place between buyers and sellers. This is all possible because of the implementation of machine learning algorithms that compare millions of transactions taking place all over the world in real time.
- **Email spam:** Spam filtering methods are used by email clients to avoid fake emails or malware. As such emails can greatly damage a system if opened, malware software is used for phishing purposes by spammers to obtain highly sensitive information such as bank card details or passwords. Machine learning and artificial intelligence algorithms, such as decision tree induction and multi-layer perceptron, help in avoiding cybercrime and provide real-time protection as well.
- **Video surveillance:** Video surveillance systems are the best way to monitor a specific location. While multiple security cameras cannot be managed by a single person, video surveillance systems are designed with machine learning and artificial intelligence technologies that train computers to do this efficiently.

  Video surveillance systems that are powered by artificial intelligence have the ability to detect and report a crime even before it happens because the system can track unusual behavior such as someone standing motionless for a long period of time or napping regularly on a specific location.

  Each reported event is handled by a machine learning model at the backend so that immediate responses can be generated to stop any kind of criminal activity or mishap.

- **Virtual personal assistants:** With the latest advancements in technology, smartphones and digital devices now feature virtual personal assistants that help users find useful information when requested via text or voice. Major applications of virtual personal assistants designed using machine learning models are speech recognition, speech-to-text conversion, text-to-speech conversion, and NLP. Siri, Alexa, Google Assistant, and Cortana are few of the most popular virtual personal assistants that help users perform routine activities every day.

Conclusion

In this chapter, we reviewed that machine learning is not only used in the financial sector, but it is being applied in other sectors as well.

The next chapter is all about the implications of deep learning.

CHAPTER 16
Ethical Considerations in Artificial Intelligence

Introduction

Deep learning systems are getting popular owing to their ability to detect fraud, optimize linguistics, conduct research, compose art, and translate in different languages. These systems are intelligent and are transforming the lives of humans for all good reasons. These systems are getting more capable and making the world around us highly efficient and richer. Tech giants such as Facebook, Microsoft, Alphabet, Amazon, and some individuals like Elon Musk are firing up the debate that now is the right time to explore the boundless applications of artificial intelligence. But as it is an emerging technology, it is imperative to talk about ethics and risk management.

There has been a surge in fears that artificial intelligence is going to rule over the world while humans will live like slaves. This may seem to be a far-fetched opinion at the moment, but no one can deny the possibility in the future. Will Smith's movie “I Robot” is a clear example of a super-intelligent machine that can outsmart humans and take control of the world. The movie shows how an AI system can become autonomous and take control of the power grid and other man-made systems. Robots are seen policing the area and killing humans. Experts fear this kind of rise of the machines. The biggest concern is what happens if AI learns to program itself.

Structure

In this chapter, we will cover the following topics:

- Loss of jobs.
- Inequality.
- Humanity.
- Disinformation.
- Artificial intelligence and evil people.
- Racist robots.
- Artificial intelligence against humans.

Objective

After studying this chapter, you should be able to do the following:

- Understand the impact of artificial intelligence on humans.

Loss of jobs

The first ethical implication of deep-learning artificial intelligence is loss of jobs. The hierarchy of labor at the moment is concerned with automation. The world is getting more proficient in automating jobs. There is little room for people to take up complex roles, shifting from physical work that once dominated the pre-industrial world to highly cognitive labor that focuses on administrative work in a globalized society. Take a look at the field of trucking. There are millions of employees in the trucking industry in the US, and Elon Musk has promised to roll out self-driving trucks. What happens to those employees if you flood the roads with self-driving trucks? Loss of jobs is one side of the mirror. The positive side is that there will be a lower risk of accidents, which makes self-driving trucks an ethical choice. A similar kind of scenario can be imagined for a majority of the workforce in developed countries like the US and the UK.

This indeed is the most immediate concern. Artificial intelligence brings mixed sentiments in this regard. It is getting increasingly clear that artificial intelligence is not a job killer. Instead, the term that is used is a category killer. There are certain categories of jobs that artificial intelligence will ultimately wipe out. For example, computer systems have replaced humans in the field of weaving of looms, leading to shifts in employment from one category to another. The same thing can happen when artificial intelligence systems land in the hands of common men. Research shows that category loss is inevitable when artificial intelligence will take over the world. The highly targeted sectors are customer service, which needs to run round the clock, and professional services. Companies will be free to put their human resources to more cognitive tasks.

This new move is fanning concerns about the displacement of labor in different industries. Artificial intelligence is speeding up digital transformation in different business processes. As different companies look forward to adapting and then implement several artificial intelligence strategies, it is time to think about whether it is important to engage in an honest conversation with employees. The latest research shows that companies that are teaming up artificial intelligence systems with humans show promising results in terms of better yield. They are doing far better than companies that have fully replaced humans with artificial intelligence systems. On the other hand, employees are welcoming these decisions. They are feeling more comfortable working with machines as their coworkers instead of seeing them as their competitors and potential replacements.

Inequality

The second ethical concern is what will happen to the distribution of wealth when machines start earning while humans sit back and relax in their homes. A majority of services and goods companies depend on hourly work. By using artificial intelligence, a company can reduce its reliance on human workforce. It means that the revenues will go into the pockets of a few men. As a result, individuals who own artificial-intelligence-driven companies will keep making large amounts of money. The wealth gap in society will continue to increase. Startup founders will bag big profits. Silicon Valley has now fewer employees than it had in the past.

Humanity

Artificial-intelligence-driven bots can learn faster with each experience, and they continue to improve themselves. They are getting better at modeling human relationships and conversations.

In 2015, an artificial-intelligence-driven bot Eugene Goostman was awarded the Turing challenge. This milestone is just the start of the era where humans will be in frequent interaction with machines as if they are real humans. Project Debater is one example. This interaction is going to happen on a large scale in the fields of sales and customer service. While humans may be limited in kindness and attention, artificial-intelligence-driven bots are not, and they can build and cement relationships that can ultimately benefit a business.

Many of us may not be aware of this, but machines already surround us. If we take a critical look around us, we will realize that different machines are interacting with us. Perhaps today you have seen a clickbait headline that led you to a business landing page, or you might have reached a video game page. These ads and headlines are optimized by artificial-intelligence-driven bots that monitor your search habits and come up with the topic you mostly search for. This, along with other methods, is used to get people addicted to video games and other entertainment.

Many people see artificial intelligence as a transformative technology. The question is whether such systems will be able to transform society fully. Will it be able to mow our lawns? Will it help us raise our kids? Will they fight our wars for us? Will it write our articles and let us relax and watch TV? Will it create political advertisements? All these questions are worth a debate.

Disinformation

Artificial intelligence systems are being used to create fake videos, images, and conversations to misguide people on important social and political debates. We cannot believe what we see online. The scenario is nothing less than a disaster if you cannot tell whether the image you see on the Internet is real or artificial-intelligence-generated. Artificial-intelligence-driven bots were blamed for the havoc that happened during the 2016 US presidential elections. Bots spread and fanned certain political propaganda on a wider scale. Automated social media accounts aided in the creation and spread of misinformation on the Internet in an attempt to manipulate voters from different strata of society. The activities fueled partisan disagreement in society and almost polarized it by deepening the gulf between two groups in the society. What makes artificial-intelligence-driven bots better than humans is their ability to work day and night tirelessly without a break. They are very fast and can generate a huge amount of content in a small time frame. Once the news is shared and retweeted by others, it starts to go viral. These bots are very effective at spreading false news or altered facts. They can amplify messages and put certain thoughts in the heads of people. Now criminals and different state actors are using fake imagery to interfere with the operations of the government. All it takes now are a couple of malicious actors who take on the job of spreading false claims to alter the opinion of the general public.

Going forward, corporations and governments will have to think about how they are going to reign amid political damage by these artificial intelligence systems. The only thing that can save governments and corporations is the way they respond to fake news and fake content. They should take artificial-intelligence -backed fake content as seriously as they take cybersecurity threats.

Artificial intelligence and crime

Do we want bad guys to get their hands on artificial intelligence technology? While artificial intelligence has the potential to do a lot of good to us, we must keep in mind that artificial intelligence can be dangerous in the hands of malicious users. As technology gets more powerful, it can inflict severe damage if it falls into the hands of the wrong people. They can use it to steal big cash and launch unprecedented attacks on government machinery. They can use it to launch a malicious hacking attack on big organizations and compromise their data. What happened during the 2016 elections left most people in awe. The latest technology targets the vulnerability in a computer network. As the technology evolves, it will become more resilient and tough and will sustain any attempt made on making these systems weaker. This is terrifying. Just imagine a scenario in which an artificial intelligence system operates in hiding and is invincible as per current security standards. It can do whatever its operator wants. Thus, there is the need for the construction and management of a resilient and advanced digital infrastructure. Detection of these malicious attacks is destined to get tougher; therefore, there is a need to patch the security vulnerabilities along with the evolution of artificial intelligence systems.

Racist robots

Artificial intelligence is capable of processing at high speed and accuracy that are beyond the scope of humans, but it cannot always be trusted to stay fair and neutral. Google is one of the leaders in terms of artificial intelligence, as you can see in its Google Photos service that categorizes different photos of different people and places them in separate folders for ease of use. It also collects photos of the same person to create collages. Artificial intelligence in Google Photos is generally used to identify different people and scenes as well.

But it would be wrong if the software that is designed to predict future criminals is biased against black people. The problem is that humans program artificial intelligence systems, so whatever humans tell them to do, they will act upon it. If humans tell them that black people are most likely to commit crimes in the future, the artificial intelligence system will label them as such. However, if the technology is used in the right manner, it can bring a positive change in the world.

Artificial intelligence vs. humans

It is not just the enemies that you should be worried about. The biggest concern that irks lots of people including Tesla CEO Elon Musk is, what if an artificial intelligence system turns against us? This doesn't mean it would turn evil in the way a human does. It means an advanced artificial intelligence system that can fulfill all of your wishes but has terrible consequences in the end. A machine has no feelings; therefore, it cannot show malice against anyone. Which means it could also a lack of understanding of the context in which the wish was made. Just imagine an artificial intelligence system that is tasked with eliminating cancer from the world. So it creates a formula to eliminate cancer and ends up killing all the people on the planet, because there is a little bit of cancer in every human being. So, this kind of misunderstanding can be devastating for humans across the world. While neuroscientists are working on unlocking the hidden secrets of the conscious experience, we need to understand the basic mechanisms of aversion and reward.

Conclusion

In this chapter, we discussed the impact of machine learning not only on the financial sector but also on our lives, economy, and humanity as a whole.

The next chapter will take you through the application of artificial intelligence in the banking sector.

Chapter 17 Artificial Intelligence in Banking

Introduction

Banks can use artificial intelligence to improve customer experience by enabling smooth and continuous interactions. However, artificial intelligence in banking is not limited to retail banking only. It can flex its muscles in investment banking and other financial services as well.

Structure

In this chapter, we will cover the following topics:

- Fraud detection.
- Cost cutting.
- Customer service.
- Risk management.
- Internet banking.

Objective

After studying this chapter, you should be able to do the following:

- Understand the applications of artificial intelligence in banking.

Fraud detection

For many financial institutions, especially banks, security is always one of the top concerns that they must address if they have to win the loyalty and confidence of their customers. There has been a constant fear of hacks and frauds, which can lead a financial institution to devastation. With the help of artificial intelligence, bank management can cut down the rates of false identity incidents, fraud attempts, and other stealing attempts. Artificial intelligence can reduce fraud attempts by making accurate interpretations of insights into different trends. It is a blend of supervised and unsupervised machine learning, and this combination can determine if a transaction has fraudulent tendencies or not.

As more financial institutions move toward digital advancements, payment fraud is rising at a sharp pace. These kinds of attacks have a digital footprint, which makes them invisible. With artificial intelligence, banks can install a security layer to protect itself against any fraudulent attempt. Artificial intelligence keeps an eye on any kind of trend swerving and detects a fraud swiftly even if it is being conducted at a massive scale. This gives banks an advantage in battling out fraud attempts and securing the reserves. Generally, banks use predictive analytics.

Cost cutting

Reduction in operational costs is another advantage of artificial intelligence in the banking sector. Here a large amount of time is spent on identifying, digitizing, and onboarding document templates. With the help of the digitization process, banks can reduce the amount of time spent on completing these tedious tasks. If as a banker you automate the process of digitization in the branches, you can slash the time that your employees would have spent on the same process. This way you can reduce the cycle time and redirect employees to other important projects.

Customer service

The customer service sector is another sector that can be improved with the help of artificial intelligence tools. Customer experience affects almost every business around the world, including the banking sector. It leaves an impact on the way people would normally perceive a particular organization. Where banks are concerned, people want access to their money even if the bank is closed for holidays. They prefer banks that offer them swift transactions. Take the example of a person who has to withdraw $1000 because he is on vacation with his family and he has run out of hard cash. He goes to the ATM and gets his card captured for some unknown reason. It is midnight and there is no employee available at the bank. This is where artificial-intelligence-powered chatbots play an important role in guiding the customer and solving his problem. Chatbots don't sleep and they are not bound by any time zone, so they can help customers irrespective of time and place.

Nowadays almost every online business has a chatbot, so you may wonder whether the one you are interacting with is a simple chatbot or an artificial-intelligence-powered chatbot. There is a big difference between the two types of chatbots. A simple chatbot that most businesses have deployed on their websites is static. It doesn't learn through its own experience until you feed it with data from the backend. However, an artificial-intelligence-powered chatbot can learn through customer interactions, which helps improve them and, in turn, improves customer experience. A brilliant example of chatbots is Erica, the virtual financial assistant at the Bank of America. You can request her to send you push notifications of your credit scores and also help you in paying bills or conducting financial transactions online.

Risk management

Credit cards pose a problem when it comes to securing financial transactions. There is always a fear of a fraudulent or malafide transaction. The best way to mitigate this risk is by conducting a thorough analysis of the potential customer so that the right person is authenticated. Artificial intelligence can help you do real-time identification of the client and thus help the banking sector save millions of dollars that are otherwise wasted on fraudulent practices. You can also deploy several technologies like facial recognition scans, iris scans, and voice recognition.

Internet banking

This is another sector in which artificial intelligence can help you. Artificial intelligence has revolutionized the concept of customer experience and banking. One of the most important features of mobile banking is its all-time presence across the world.

The impact of artificial intelligence on the banking sector runs deep, and it is yet to be explored fully. The banking sector is witnessing many innovations, thanks to machine learning and artificial intelligence. In the coming years, these are likely to make further progress.

Conclusion

In this chapter, we discussed the applications of artificial intelligence in the banking sector. The next chapter explains some traditional machine learning algorithms.

Chapter 18 Common Machine Learning Algorithms

Introduction

Machine learning is a process that teaches computers to learn on their own. Now, maybe you're wondering, why on earth would we expect robots to learn on their own? Yeah, it's got a lot of perks. Applications of machine learning are numerous, and they are powered by solid fundamentals—lots of data emitted by sensors across the world, with low-cost storage and lowest ever computing costs!

Structure

In this chapter, we will cover the following topics:

- Regression.
- k-means clustering.
- k-nearest neighbor (KNN).
- Principal component analysis (PCA) algorithm.
- Polynomial fitting and least-squares algorithm.
- Forced linear regression algorithm.

- Support vector machine (SVM) algorithm.
- Conditional random fields (CRFs) algorithm.
- Decision tree algorithm.

Objective

After studying this chapter, you should be able to do the following:

- Understand the concepts of traditional supervised and unsupervised machine learning algorithms.

Regression

Regression techniques fall under the supervised machine learning category. They help predict or describe a given numerical value based on the set of previous information, such as anticipating the cost of a property based on previous cost information for similar characteristics. Regression techniques range from simple (such as linear regression) to complex (such as regular linear regression, polynomial regression, decision trees, random forest regression, and neural networks, among others).

The simplest method is linear regression, which has a mathematical equation of the form Y = m * X + b. Multiple data pairs (X, Y) can train a linear" regression model by calculating the position and slope of a line in order to reduce the total distance between the data points and the line. In other words, calculating the" slope (m) "and" y-intercept (b) "has been used for a line that provides the highest approximation for data observations. The data relationships can be modeled using" linear predictive functions ", estimating unidentified model variables at based on the data these systems are called linear models. Traditionally, if the values of the explanatory variables or predictors are known, the conditional mean of the response would be used as the affinity function of those values. Use of conditional media and other measures in linear models is very rare. Similar to any other form of "regression analysis " the "linear regression" works okay on the "conditional probability distribution" of the responses instead of the joint probability distribution of the variables obtained with the multivariate analysis.

The most thoroughly researched form of regression analysis with wide applicability is linear regression. This is because models that rely linearly on their unidentified parameters are easy to work with compared to the models that are nonlinearly related to their parameters. In other words, the statistical characteristics of the resulting predictors can be easily determined with a linear distribution. There are many useful applications of linear regression:

- If the goal is to generate forecasts and predictions or to reduce errors, the predictive model can be linked to an identified data set and explanatory variables using a linear regression algorithm. Once the model is developed, new unresponsive input data can be easily predicted by the appropriate model.
- Linear regression analysis can be used to specifically quantify the relationship between the predictors and the response, to assess whether certain explanatory variables lack any linear relationship to the response. It can also be used to identify predictor subsets that contain data redundancy across response values.

The adaptation of most linear regression models is accomplished using the least squares approach. However, this model can also be fitted by significantly reducing the lack-of-fit in another standard (like the least absolute deviation regression), or by using a "punished version of the smallest square as done in the ridge regression. minimize (L2 standard penalty) and lasso regression (L1 standard penalty) ". In contrast, it is possible to use the least squares approach to fit machine learning models that are not linear. Therefore, although the terms “least squares” and “linear model” are closely related, they are not the same.

Multiple linear regression is generally the most common type of regression technique used in data science and most statistical tasks. As with the linear regression technique, there will be an output variable Y in multiple linear regression. However, the difference now is that we will have numerous values of X or independent variables that generate predictions for Y.

For example, a model developed to predict housing costs in Washington, DC, will be driven by multiple linear regression techniques. The cost of housing in Washington, DC, is Y or dependent variable for the model. X or the independent variables for this model include data points such as proximity to public transport, training district, square meters, and single rooms, which will ultimately determine the market price of the homes.

The mathematical equation for this model can be written as follows:

$Housing\_price = \beta 0 + \beta 1\ sq\_foot + \beta 2\ dist\_transport + \beta 3\ num\_rooms$

Our models developed a straight line in the last two types of regression techniques. This straight line is the result of the connection between X and Y, which is linear, and the influence X has on Y does not change with the values of X. In the case of polynomial regression, our model shows a curve.

If we try to fit a graph with nonlinear features using linear regression, it would not yield the best fit. For example, in the image below, the graph on the left is a scatter plot showing an upward trend, but with a curve. A straight line does not work in this situation. Instead, we generate a line with a curve corresponding to the curve in our data with polynomial regression, like the graph on the right in the figure below. The equation of polynomial regression will appear same as that for linear regression, except that one or more of the X variables will be associated with a polynomial expression.

$Y = mX2 + b$

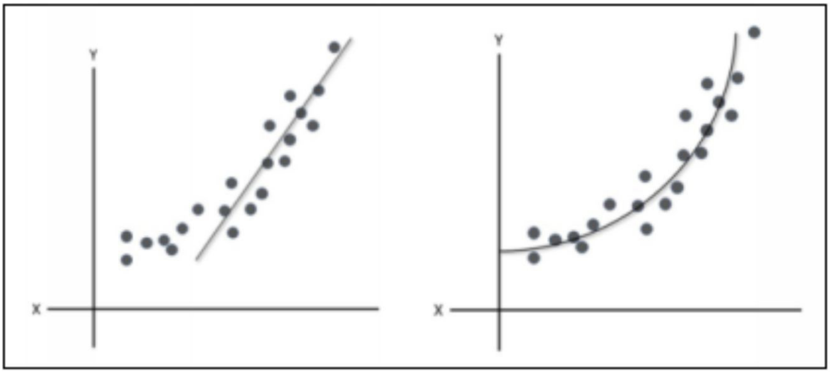


Figure 18.1

Another important regression technique for data researchers is support vector regression, which is most commonly used in case classification. The concept here is to discover a line in space that divides data points into different categories. It is also used for regression analysis. It is a form of binary classification technique that is not associated with probability.

Ridge regression is widely used for analyzing multicollinear data sets. Depending on the characteristics of the data set, the correct use of cause addictive models regression can reduce standard errors and significantly improve model accuracy. Ridge regression can be useful if your data contains highly correlated independent variables. If you can predict an independent variable using another independent variable, your model will have a high risk of multicollinearity. For example, if you use variables that measure a person's height and weight, these variables in the model probably create multicollinearity.

Multicollinearity can potentially affect the accuracy of the predictions generated by the model. Consider the type of predictor variables used in the model to avoid multicollinearity that can be caused by the type of data you are using, as well as the data collection method. Another reason could be the selection of a small variety of independent variables, or the selection of limited number of independent variables, resulting in very similar data points.

Multicollinearity can also be caused by the type of model. Note that there are more variables than data points in the model. If you have chosen to use a linear model, which worsens the multicollinearity of the model, you can try to implement cam regression. Ridge regression can make the predictions more accurate by allowing a touch of bias in the model. This technique is also known as "regularization."

Another technique to improve the accuracy of the model is to standardize the independent variables. The simplest way to do this is to reduce complexity by changing the values of certain independent variables to null. The approach is not just to change these independent variables to zero, but to implement a structure that rewards values closer to zero. This will decrease the coefficients, which will also decrease the complexity of the model, but the model will retain all its independent variables. This will give the model more bias, which is a tradeoff for greater accuracy of predictions.

Another reduction technique is called a LASSO (least absolute crimp and selection operator) regression. A highly complementary cam regression, lasso regression promotes the use of simpler and leaner models to generate predictions. In lasso regression, the model lowers the value of coefficients close to zero. Data on our scatter plot, such as the average or median values of the data, is reduced to a more compact level. We use this when the model experiences high multicollinearity as compared to the ridge regression model.

A hybrid of LASSO and cam regression methods is known as "elastic net regression." The main aim is to further improve the accuracy of the predictions generated by the LASSO regression technique. Elastic net regression rewards smaller coefficient values. All three of these designs are available in the R and Python glmnet suite.

Bayesian regression models are useful if there is insufficient data or if the available data is poorly distributed. These regression models are developed from probability distributions rather than data points, which means that the resulting graph will appear as a bubble curve representing the variance with the most common values in the center of the curve. The dependent variable Y in Bayesian regression is not valuation but a probability. Instead of predicting a value, we try to estimate the probability of an event. This is considered frequentist statistics, and are based on the Bayes theorem. Frequentist statistics assume whether an event will occur and the likelihood that it will recur in the future.

Conditional opportunity is an integral part of frequentist statistics. It refers to the events on which the results are interdependent. Events can also be conditional, meaning the previous event can change the probability of the next event. Let's say you have a box of M&M's, and you want to know the likelihood of getting different colored M&M's out of the bag. If you have a set of three yellow M&M's and three blue M&M's, and you get a blue M&M on your first draw, chances you will get a blue M&M on your next draw out of the box would be lower than the first draw. This is a classic example of conditional probability. On the other hand, an independent event is the flip of a coin, meaning that the preceding coin cover does not change the probability of the next coin cover. Therefore, a coin flip is not an example of conditional probability.

Linear regression

Linear regression is applied when you need to come up with an actual set of values based on continuous data. Basically, it involves drawing a clear line where a correlation between dependent and independent variables is considered.

In linear regression, the equation $Y = a*X + b$ is used where Y is the dependent variable, a is the slope, X is the independent variable, and b is the intercept.

To make this concept easier to understand, here's an example. You were asked to arrange a group of people according to their weight without really knowing their actual weights. You do, however, know what their heights are, and you can also visibly see each person's individual body build. Using these parameters, you can come up with a way to arrange the group that would probably yield accurate or near-accurate results.

Linear regression has two types: simple and multiple. Simple linear regression only involves one independent variable, while multiple linear regression involves two or more independent variables.

Logistic regression

Although there is such a thing as a regression algorithm, this is not at all related to that. Logistic regression actually deals with binary values (yes or no, true or false, 0 or 1) using independent variables as a basis. It uses the logit function to measure the probability of a certain event. Because it involves probability, it is only expected that the output would lie anywhere between 0 and 1.

If your friend gives you a puzzle, you can only expect two outputs: either you solve it or you don't. Now, what if you are given a number of different puzzles to see which areas you're good at; this is where logistic regression can be applied.

Let's say one of the puzzles given to you is a riddle commonly given to school-aged children. Then you would probably have an 80% chance of solving it. What if you were given a cryptic puzzle that only an expert can solve? Then logistic regression would probably tell you that you only have a 20% probability of solving it. Basically, coming up with these numbers involves choosing parameters that aim to observe the values being used as a sample instead of looking at the sum of squared errors.

Logistic regression also helps answer the common questions we use in data science as examples: How does the probability of getting lung disease (yes vs. no) change for each additional pound a patient is overweight and for each pack of cigarettes consumed every day? Do calorie intake, fat intake, body weight, and age have effects on the probability of having a stroke (yes vs. no)? As we see, these questions have a certain level of complexity, and logistic regression can offer a solution. Applied in business, it is a powerful tool for organizations to solve complex questions about trends and business patterns.

Let's explore another example of application of logistic regression by analyzing the following plot. We see all data points have either the value 0 (fail) or 1 (pass). We also can observe that the logistic fit is the S-curve, which models the probability of success as hours of study function. By analyzing the plot, we can conclude that most of the students who studied less than four hours failed the exam. However, for students who studied four hours, the model predicts that the probability of passing will be around 70%. As we see on the plot, the S-curve is a side effect of how the logistic regression estimates the probability of a different event. This is an example of the type of relevant information logistic regression can provide us with:

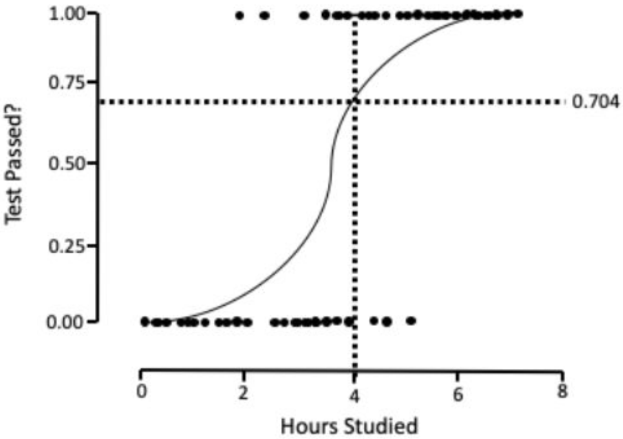


Figure 18.2: *Logistic Regression: Tests passed related to hours studied (Source: Cerebro MLAI)*

A critical point to comprehend logistic regression is related to probability. It is crucial to understand that they are between 0 and 1, and, contrary to linear regression, the values we get from the model don't offer us direct predictions for the values we are observing. Instead, by observing X = 4, the model shows us that we can expect around 70% of chances of passing in the test.

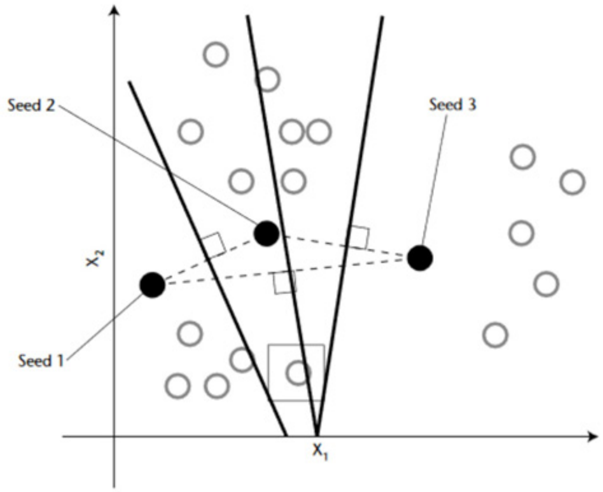
Compared to logistic regression, linear regression works on the data well, and it can forecast the value of a result fairly well only by being aware of the value of the predictor variable. Some data scientists get confused about the differences between the concepts and applications of linear regression and logistic regression. Understanding these differences is crucial to offer customers and organizations support in their decision-making process.

k-means clustering

One of the most widely and extensively used clustering algorithms is the k-means algorithm. The "k" in its name relates to the fact that the algorithm looks for a fixed number of clusters that are identified in terms of data point proximity to one another. The version listed here was published first by *J. B. MacQueen* in 1967. The technique is demonstrated using two-dimensional diagrams for ease of clarification. Note the algorithm typically manages far more than two independent variables in operation. This implies that, instead of points corresponding to vectors with two elements (X1, X2), the points correspond to vectors with n elements (*X1, X2 . . . Xn*) In itself, the process is unchanged.

Three steps of the k-means algorithm

The k-means algorithm randomly selects k data points to be the seeds (a seed is an embryonic cluster with one element in it) in the first step. MacQueen's algorithm simply takes records of the first k points. In cases where the records have some significant order, selecting widely spaced records or a random collection of records might be preferable. This example sets Cluster Number to 3. The latter step assigns each record to the nearest seed. One way to do that is to find the boundaries between the clusters, as geometrically shown below. The boundaries between the two clusters are the points from which each cluster is equidistant.



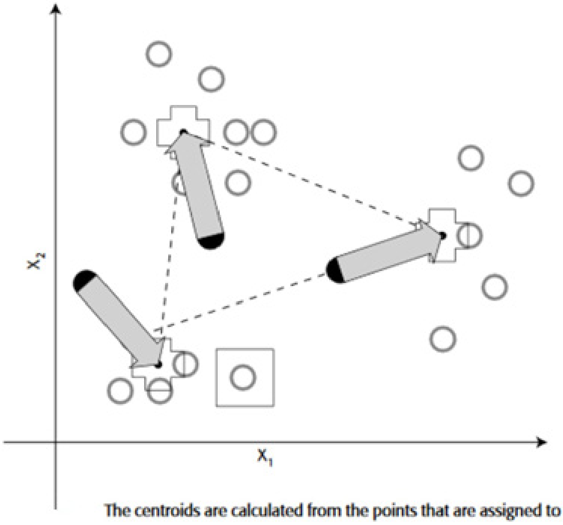
The initial seeds determine the initial cluster boundaries.

Figure 18.3

Recalling a high school geometry lesson makes this less complicated than it sounds: given any two points A and B, all points that are almost equidistant from points A and B fall along a line (which is called the perpetual bisector) that is perpendicular to the one that connects A and B and all points between them. The dashed lines link the initial seeds in the figure above; the resulting cluster boundaries shown with solid lines are at right angles to the dashed lines. Using these lines as guides, it is clear which records are closest to the seeds.

These boundaries would be planes in three dimensions, and they would be N-dimensional hyperplanes *N – 1*. Fortunately, such conditions are managed easily by computer algorithms. Finding the actual boundaries among the clusters is useful for geometrically displaying the operation. However, in practice, the algorithm normally calculates the distance between each record and each seed, and chooses the minimum distance for this step.

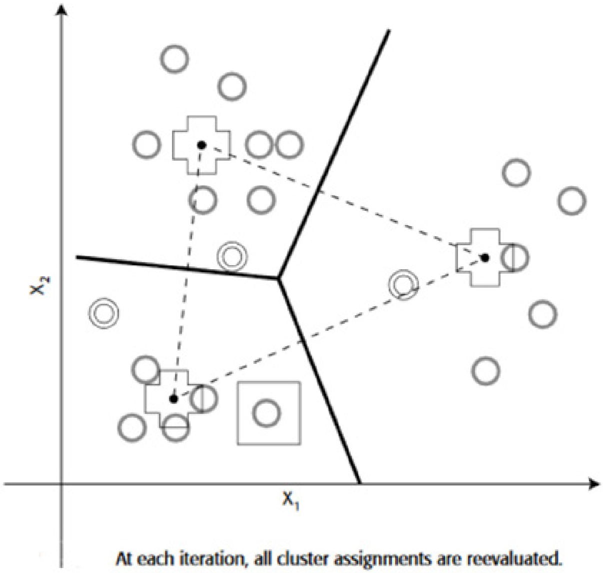
Consider, for example, the record with the box drawn around it. Based on the initial seeds, this record is assigned to seed number 2 because it is likely closer to that seed than to one of the other two. Every point has been allocated at this point to exactly one of the three clusters based around the original seed. The third step is to measure the cluster centroids; these now do a better job of characterizing the clusters than the initial seeds. Finding the centroids is simply about taking the average value of each dimension for all the cluster records. The new centroids are stamped with a cross in *Figure 18.4*. The arrows display the movement of the seeds from their original location to the cluster of new centroids produced from those seeds.



The centroids are calculated from the points that are assigned to each cluster.

Figure 18.4

The centroids become the seeds for the algorithm's next iteration. Step 2 is repeated, and each point with the closest centroid is again allocated to the cluster. The figure below shows the new boundaries of clusters formed, as before, by drawing equidistant lines between each pair of centroids.



At each iteration, all cluster assignments are reevaluated.

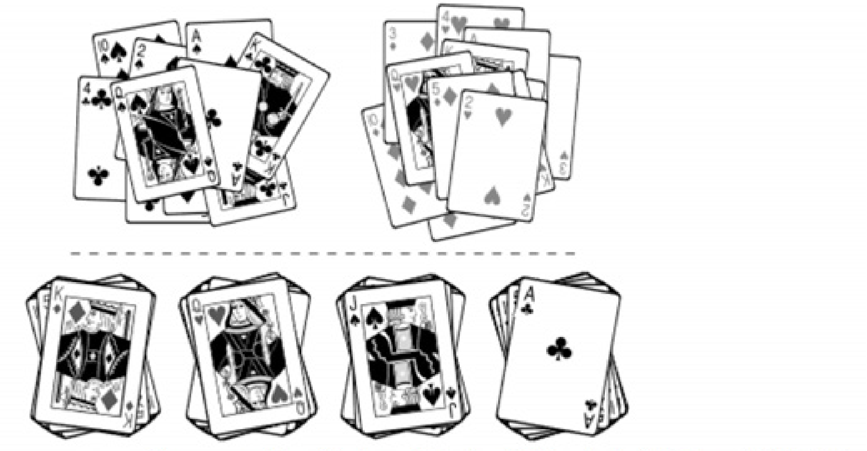
Figure 18.5

Note that cluster number 1 has now been allocated to the point with the box around it that was originally assigned to cluster number 2. The process of assigning cluster points, and then recalculating centroids, continues until the boundaries of the cluster stop shifting. In practice, after a few tens of iterations, the k-means algorithm typically finds a set of stable clusters.

What does k mean?

Clusters define the structure that underlies the data. There is no one proper definition of the structure, however. For example, someone who isn't from New York City may think the entire city is downtown. Someone from Brooklyn or Queens may apply this nomenclature to Manhattan. It could just be areas within Manhattan south of 23rd Street. And even there, in the southern tip of the island, downtown could still be reserved only for the taller buildings. Clustering poses a similar problem; data structures occur on several different levels.

k-means definitions and allied algorithms gloss over k's variety. But since there is no a priori justification to choose a specific value in certain situations, there really is an outermost loop to these algorithms that exists in the course of analysis rather than in the computer program. This outer loop consists of automatic cluster detection using one k value, evaluating the results, and then retrying with another k value, or perhaps modifying the data. These tests can be automated, but it is often important to evaluate the clusters on a more subjective basis to assess their utility for a given application. As shown in *Figure 18.6,* different k values can lead to equally valid clusters of very different types. The figure shows the clustering of a deck of K = 2 and K = 4 playing cards. Is one better than the next? It depends on the use the clusters are put to:



These examples of clusters of size 2 and 4 in a deck of playing cards illustrate that there is no one correct clustering.

Figure 18.6

For the first time, k-means clustering is performed on a given data set. Most data points fall into one massive central cluster, and the rest form a number of smaller clusters outside. This is mostly because most records describe natural variations in the data, but the clustering algorithm has enough outliers to cause confusion. This form of clustering can be useful for applications such as detection of fraud or fabrication defects. In other cases, removing outliers from the data can be desirable; more often, massaging the data values is the solution.

Distance and similarity

If records are mapped to space points in a database, identification of automatic clusters is really very simple—a little geometry, some vector means, and voilà! Of course, the issue is that the systems encountered in marketing, distribution, and customer service are not about space points. They are about sales, phone calls, airplane journeys, car registration, and a thousand other items that have no apparent relation to the dots in a cluster diagram. This kind of clustering records involves some notion of natural association; that is, records in a cluster are equal or related to each other than records in another cluster. Since abstract concepts are hard to express to a machine, this vague definition of connection needs to be converted into a kind of numerical measure of the degree of similarity. The most popular but by no means the only approach is to convert all fields into numeric values so that the records can be viewed as space points. In the geometric sense, then, if two points are identical, they represent similar records in the database. This method has two key problems: (1) many variable forms, including all categorical variables and many numerical variables like rankings, do not have the correct behavior to be handled properly as components of a position vector, and (2) the contribution dimension is of equal significance in geometry, but in databases, a minor change in one field can be much more significant than a big change in another. A few alternative similarity measures are introduced in the following section.

Similarity measures and geometric distance of variable type work well as similarity measures for well-behaved numeric variables. A numeric variable is one whose value in our geometric model indicates its location along the corresponding axis. Not all of the variables fall into that group. Variables fall into four classes to this end, listed here in increasing order of suitability for the geometric model.

- Differential variables.
- Ranks.
- Intervals.
- Right measures.

Categorical variables define only one of many unordered categories to which one element belongs. For example, an ice cream may be labeled pistachio and another butter pecan, but it is not possible to tell that one is larger than the other or to determine which one is closer to black cherry. Mathematically, X > Y can be assumed, but not if X > Y or X < Y. Ranks put things in order, but don't assume that one thing is any bigger than another. The valedictorian has higher grades than the salutatorian, but by what margin, we don't know. If X, Y, and Z are ranked A, B, and C, we know that X > Y > Z, but X – Y or Y – Z cannot be specified. Intervals measure the interval from one measurement to another. If it is 56 °F in San Francisco and 78 °F in San Jose, then at one end of the bay it is 22 °F colder than at the other.

True measures are variables of intervals that measure from a meaningful zero point. This trait is important because it means the ratio of the variable's two values is large. The temperature scale of Fahrenheit used in the US and the scale of Celsius used in the rest of the world have this property. It doesn't make sense in either system to say a 30 ° day is twice as warm as a 15 ° day. Similarly, a size 12 dress isn't twice as big as a size 6, and gypsum isn't twice as hard as talc, but on the hardness scale, they are 2 and 1. But it makes complete sense to say a fifty50-year-old is twice as old as a twenty-five-year-old, or a ten-pound sugar bag is twice as heavy as a five-pound one.

Examples of true measurements are age, weight, length, customer tenure, and duration. Geometric distance metrics for interval variables and true measurements are well defined. It is important to transform these into interval variables in order to use categorical variables and rankings. Unfortunately, it may apply false details to those transformations. If random numbers 1 through 28 are assigned to ice cream flavors, it will mean that flavors 5 and 6 are nearly related while 1 and 28 are far apart.

KNN algorithm

The KNN algorithm is one of the simplest machine learning algorithms used by today's data scientists. It was first developed for statistical prediction and pattern recognition in the early 1970s and can be applied to both regression and classification problems, although it is most commonly applied to classification models.

How does KNN work?

When making predictions, KNN must access the entire set of training data available. This means no learning is required as all data must be stored in memory and remain always accessible. If you are dealing with large data sets, you may want to consider complex data structures, such as k-d trees, to maximize storage efficiency and computational speeds.

For each new data point, KNN searches through all the training data for the k closest entries (nearest neighbors). k is the number of neighbors considered by the algorithm: for k = 1 only the closest value will be considered, for k = 3 the three closest values will be considered, for k = infinity the entire data set will be considered, etc.

Once the k closest data points are identified (k-nearest neighbors), their properties and values are assessed. A function will consider all of these properties and compute an output value for the new data point. For regressions, the predicted value might be the weighted average of the nearest neighbors, and for classification, it may be the most frequent class.

Distances or degree of separation between the new data point and its neighbors must also be considered. For instance, if two squares neighbors are closer than three strawberry neighbors, the predicted class of the new entry point may be squares. Of course, this depends on how you define distances in your data set (for now, you can think of distance as a measure of error).

Before diving into the concept of distance and discuss the different metrics available, let's visualize what was discussed this far. In the following figure, you will find a training data set for two classes of data: squares and diamonds. The two variables that define these points are: Loan$ and Age. These are called the predictors.

Using this data set, a KNN algorithm can be used to predict a class of new data points: is the circle a square or a diamond? In this particular example, I have chosen k = 4 for my KNN algorithm. Now, I must select the four nearest neighbors to the new data point, i.e., the circle.

As you can see, two of the closest neighbors are diamonds and two are squares. What class should we assign to the circle? This will depend on how your algorithm defines the distance parameter. For instance, if you consider absolute distance (i.e., how far the point is on the graph from the training data), then your algorithm will predict "diamond" because the two diamonds are closer. However, if age is a more important factor in your study, you may choose to give it a greater weighting factor w. Hence, your distance is more biased towards minimizing the difference in age and will predict "squares" (because the squares are closer in age to the circle). Distance is a key parameter for any KNN algorithm, and hence, it must be chosen and tuned carefully on the basis of needs of your study.

When specifying distance metrics for your algorithm, a few things must be kept in mind. First, you must select the most relevant predictors (i.e., age and loan in the previous example). Due to differences in the scale between predictors, it is common practice to normalize them to values between 0 and 1. For instance, in the example above, you would have Loan$ in the range of $60,000 to $250,000 and Age ranging between 20 and 52; these discrepancies in scale must be factored into how your algorithm will deal with the distances between training data.

If you want to assign a greater factor of importance to a chosen predictor, you can use a weight factor w after normalizing the data. For example, now your data will range from 0 to 1*w. The greater the weight factor, the greater impact a predictor will have on determining the final output. Choosing the predictors and weight factors that give maximum accuracy is called parameter tuning. There is no right or wrong way to do this; it depends on many factors, and the best way remains trial and error.

Although this process may appear confusing and even random at first, with experience you will learn to tune parameters quickly and effortlessly to an algorithm that works for your data and your requirements—it just takes practice.

Once you have selected and weighted the predictors, you must now choose how the distance parameter is calculated. The most common approach is using Euclidian distance (if you do not have prior experience in machine learning, I highly recommend you use this technique). It is a very effective and straightforward technique, regularly used throughout the industry. The mathematical formula to compute the Euclidian distance is shown on the following page.

There are other distance measures that can be implemented into KNN. Owing to the broad range of subjects covered in this book, I will only briefly mention these other metrics. Of course, there are more distance metrics that can be implemented; you can even create a bespoke distance function yourself to suit your predictors perfectly. Like many aspects of machine learning algorithms, a lot comes down to parameter tuning:

- **Hamming distance:** Useful when dealing with binary vectors (i.e., classes).

- **Manhattan distance:** Use the sum of the absolute difference between data vectors to calculate distance.
- **Minkowsky distance:** A combination of Manhattan and Euclidian distances.

With regards to the optimum value for k, again, there is no right answer. Parameter tuning and experimenting are required to identify what works best for your training data and your individual needs. However, do keep in mind that large values of k will have longer computational times. The specific impact of this depends on the complexity of your algorithm, the size of the data set, and computational power available.

Another key concern when dealing with the KNN algorithm is scale, particularly relating to the size of your training data. Remember that, in order to make a prediction, the algorithm must access each entry in the training data and search for k with the most similar values (the nearest neighbors). If your training data consists of hundreds of thousands of entries, analyzing each point means significant challenges in computational time and storage space.

An effective solution is using a stochastic subset of data. This means that, if your training data has 100,000 entries, you build your KNN algorithm using only 1,000 randomly selected points. This is usually a good workaround for large data sets. You can drastically reduce computational time and retain a good level of accuracy, especially when there is a lot of repetition in your training data. To choose an appropriate size of the stochastic subset, I tend to run trials and compare the results against the entire data set. You will find that, as subset size decreases, so does the accuracy of your prediction. Based on your computational time and accuracy requirements, select a minimum threshold discrepancy (e.g., 95%), and then choose the smallest subset size that matches this. Be sure to run this test for multiple entries to avoid any outliers.

Preparing your training data for KNN

Preparing and filtering training data before running any machine learning algorithm can drastically improve predictive power, accuracy, and efficiency. When implementing a KNN, you should always do this following:

- **Rescale the predictors:** Always try to normalize your predictors so they fall between values of 0 and 1.

- **Remove missing data:** If a data point is missing (i.e., has a value of zero), it can cause severe damage to the distance calculations. All zero data should be removed from the training data.
- **Reduce the number of entries:** As already explained, KNN must access and examine all entries in your training data set. If this contains 100,000 points, high computational power will be required. By using a stochastic subset (e.g., 1,000 randomly selected points), you can greatly reduce computational requirements, often with minimal effect on accuracy.

Final remarks

The KNN algorithm is a simple machine learning algorithm capable of delivering highly competitive results. This model requires access to all training data sets as it does not carry out a learning process. It is a perfect choice for newcomers to machine learning algorithms or those looking to build a predictive model quickly. In spite of its inherent simplicity, there are many variations and parameters you can tune to maximize predictive accuracy and computational speed. Please keep in mind there is no single most effective setup; as with most machine learning algorithms, you will have to experiment and tune all parameters using a trial-and-error approach that best suits your training data.

Principal component analysis (PCA) algorithm

It is among the simplest algorithms of machine learning. This helps you to identify minimum data element which can predict and to sacrifice as minimal details as possible. This is used in various fields such as object detection, computer vision, and encoding of files. The estimation of the key components is limited to the computation of the initial data's eigenvalues, eigenvectors, and covariance matrix values or the data matrix's singular decomposition.

With PCA, we can communicate different signals, combined, so to say, simplified models. For example, it would most certainly not be feasible to prevent a lack of information, although the PCA approach would help us mitigate it. It is a method to measure the elements ordered.

Polynomial fitting and least squares algorithm

The least squares approach is a statistical technique used to address different problems. It focuses on reducing the number of squares of variations from the ideal variables of other functions. It may be used to solve over-determined equation structures (whenever the number of iterations reaches the number of uncertainties), to look for answers in the ordinary sense (not over-determined) variational equation systems, as well as calculate the points earned of a specific function. Use this algorithm to fit basic curves/regressions.

Forced linear regression algorithm

The least squares method may confuse overshoots, false areas, etc. Constraints are important to reduce the line variation that we bring into the data collection. The best answer is to suit the linear regression formula, which means that the weights are not wrongly defined. Models can be either LASSO or ridge regression or maybe both (elastic regression). Use this method to suit constraints on the regression axes, without overriding.

Support vector machine (SVM) algorithm

SVM is a linear model (e.g., logistic/linear regression). The distinction is that it has a loss feature and is focused on margins. Using optimization techniques (e.g., L- SGD or BFGS), you can minimize the loss function. One aspect of SVMs works well on small dataset. Classification algorithms (even regressors) may be equipped with SVM.

Conditional random fields (CRFs) algorithm

This algorithm is used represent an RNN-like series and can be used in combination with an RNN. These could also be used, for instance, in image segmentation and other organized prediction activities.

CRF algorithm models each part in the series (say, a sentence), such that the neighbors influence the item marked in the series and not the individual labels. Use CRF algorithm for sequences such as image, text, time-series data, and DNA.

Decision tree algorithm

One of the most popular algorithms in machine learning, it is used for statistical modeling in analytics and in data processing. The arrangement reflects the divisions and the leaves. The objective function attributes rely on the branches of a decision tree. The objective function results are calculated in the leaves, and the residual nodes include attributes on which the cases vary.

In case of a new case to be categorized, you have to go down a tree to the leaf to offer the correct meaning. The goal is to construct a model centered on multiple input variables that forecasts the output variable value.

Conclusion

In this chapter, we discussed about traditional algorithms: regression, k-means clustering, k-nearest neighbor, PCA algorithm, polynomial fitting and least squares algorithm, forced linear regression algorithm, SVM algorithm, CRF algorithm, and decision tree algorithm.

The next chapter covers FAQs.

Chapter 19
Frequently Asked Questions

1. **What is artificial intelligence?**

   Ans. Artificial intelligence can be defined in different ways, but it basically describes the range of capabilities that can be demonstrated by machines. The issue of defining artificial intelligence is trickier than it may seem at first because our understanding of artificial intelligence has changed as the ceiling of what machines can do (and are doing) is progressively raised.

   For example, some think of artificial intelligence in terms of machine learning. This sort of definition of artificial intelligence would be focused on the ability to receive external cues from the environment, model that environment, and make changes to its core structure or behavior on the basis of those changes. Others conceptualize artificial intelligence as being related to those capabilities that artificially intelligent machines or agents have not yet accomplished. This implies that our understanding of artificial intelligence is constantly being changed as new advancements are made.

2. **What is an agent in the context of artificial intelligence?**

   Ans. The machine or program that is displaying the intelligence capabilities is referred to as the agent. The agent is essentially the actor that is engaged in cognition or learning in a way that is described as artificially intelligent. The term "agent" basically allows scientists and theorists to have a convenient way of describing artificially intelligent machines both in a real, practical sense and theoretically.

3. **Would artificially intelligent agents really be just intelligent humans?**

   Ans. Early concepts of artificially intelligent agents usually visualized these agents as automatons or robots that resembled humans and essentially completed tasks that human beings would complete, whether in a rote and simplistic fashion or in a complex, intelligent way. Even the ancient Greeks imagined automatons as being able to engage in locomotion, attack other people, and essentially do the things that humans do. The only difference was that they were not actual humans. They were artificial.

   Artificial intelligence today recognizes that the forms that artificial intelligence displays can take vary greatly. Most artificial intelligence agents today are computer programs that record data or stimuli from the environment and interpret it. For example, many businesses have artificial intelligence software that records purchases, product views, or other behavior and makes predictions and suggestions on the basis of that data. In this way, artificially intelligent agents may display some aspects of human behavior while still being (at least, at present) stand-ins for humans: completing essentially human tasks rapidly and reliably. In short, the agents of today are clearly programs that are distinct from human beings.

4. **Are there different types of artificial intelligence?**

   Ans. There are different types of artificial intelligence. This includes subsets or extrapolations of artificial intelligence like machine learning or deep learning. But the term usually refers to approaches to artificial intelligence in terms of how the agents operate. The main classification system divides artificial intelligence into (1) analytical artificial intelligence,

(2) human-like artificial intelligence, and (3) humanized artificial intelligence.

Understanding different types of intelligence is key to knowing the difference between these three types of artificial intelligence. Analytical artificial intelligence is thought of purely in terms of cognition. Cognition represents the range of skills like processing information, language production, and learning that are thought of as representing higher order thinking. This is the type of thinking that human beings and some higher animals are capable of.

Human-like artificial intelligence adds emotional intelligence to cognitive intelligence in its range of abilities.

Humanized intelligence is perhaps the most human of the three types of artificial intelligence. Humanized artificial intelligence includes cognitive intelligence, emotional intelligence, and social intelligence. Emotional and social intelligence represent more nuanced forms of intelligence. These are the types of bits of intelligence that robots, for example, would need to exhibit in order to not be detected immediately as being robots.

5. **Why is artificial intelligence receiving so much attention lately?**

Ans. Artificial intelligence is receiving so much attention lately because, as artificial intelligence breaks new barriers in what it is able to accomplish, many have begun to theorize that artificial intelligence represents a so-called existential threat to human beings. An existential threat is essentially a factor that represents a danger to human survival as a species.

Why is artificial intelligence believed to be an existential threat? Artificial intelligence has been postulated to represent an existential threat because of the understanding that artificially intelligent agents that are fully aware (in science-fiction terminology) or that have achieved singularity in scientific terms will be able to exceed collective human intelligence. An artificially intelligent agent would, therefore, be more intelligent than all human beings in history put together. An undercurrent of this fear is the supposition that such an intelligent agent may attempt to destroy the human race, although that is at this stage an assumption.

6. **What is the connection between artificial intelligence and robotics?**

   Ans. Although many people often think of artificial intelligence as a quality associated with robots, most agents of artificial intelligence at present are computer programs that complete specified tasks. Some of these programs merely receive data from their environment and accomplish specified tasks, while others are able to make changes to their structures in response to information that has been gained or learned, which is referred to as machine learning in the context of artificial intelligence.

7. **What are artificial neural networks?**

   Ans. Artificial neural networks are a type of network that some artificially intelligent agents are created with. Artificial neural networks are designed to resemble, at least in part, the neural networks of human beings. The human central nervous system contains millions of connections in the forms of neurons, synapses, and innervated tissue. These human neural networks allow human beings to perceive their environment, operate in their environment, interact with other objects in their environment, form language, and learn.

   Scientists studied human neural networks in order to create the artificial neural networks that some artificially intelligent agents have. The idea is that artificially intelligent agents would be able to learn and thus improve their functioning if they had basic network components that resembled those of human beings. Although the full capabilities of this technology have yet to be realized, it is likely that, should artificial intelligence achieve singularity, artificial neural networks would be a key (if not the critical) component of their operation.

8. **What role can artificial intelligence play in business and finance?**

   Ans. Artificial intelligence already plays a big role in business and finance, even if the average person in the US (or other Western countries) does not realize it. Companies use artificial intelligence on their business websites or other customer interfaces to improve the functioning of their business. Artificial intelligence programs can make purchasing suggestions to consumers, interpret consumer phone calls to direct them to the proper individual to handle their issue, or make suggestions on music and video websites or social media on the basis of previous activity.

In other words, artificially intelligent agents can record and analyze consumer behavior in ways that may not be cost-effective for businesses to do otherwise. These artificial intelligence programs, therefore, can accomplish the sorts of tasks that businesses have always wanted to do, but did not how to or did not have the manpower to do.

9. **Do I need to integrate artificial intelligence into my business?**

   Ans. There certainly is no law stating that you are required to integrate artificial intelligence into your business. The machines shave not taken over yet, so this future event is still a long way away. With that said, it may be a poor business decision not to think of ways that you can integrate artificial intelligence into your business. It may surprise you to learn how simple this may be. You may choose to have a chatbot on your website, or you may have a customer service number that uses artificial intelligence to direct users where they want to go. As most businesses are using artificial intelligence of some kind in their endeavors, it is certainly in your best interest to put some thought into it.

10. **What is machine learning?**

    Ans. Machine learning refers to the ability of machines to learn and adapt with the help of algorithms that allow the machines to make predictions and behave without being explicitly programmed. Machine learning is considered, by most, a subset of artificial intelligence, as it represents a type of intellectual capacity that machines or artificially intelligent agents can be imbued with.

    In this regard, machine learning perhaps is most closely related to analytical artificial intelligence as it involves cognitive abilities without the human-like emotional and social bits of intelligence that some artificially intelligent agents have been imbued with. It is important to understand machine learning as it allows the person to study artificial intelligence and get a sense of how artificially intelligent agents can learn and change their performance over time.

11. **What is deep learning?**

    Ans. Within machine learning is the realm of deep learning, which refers to the layered type of algorithms that allow machines to detect information and process it in complex ways. Deep learning models are based on what is known as artificial neural networks, computer systems whose creation was inspired by the functionality of the human brain. The human brain also processes information in a layered or overlapping fashion, which allows networks designed to resemble them to have some basic characteristics of human brains.

12. **How long has artificial intelligence been around?**

    Ans. The concept of artificial organisms or machines capable of thinking in human-like ways or otherwise independently has been around for thousands of years. Ancient Greeks described an automaton called Talos that had been created by the god Hephaestus and that was able to behave independently and attack human beings. Perhaps this is where the idea of dangerous artificially intelligent beings began. Artificial intelligence as an area of study is believed to date from the 1950s. Computers had become prominent and important in the 1940s, and the first reference to artificial neurons dates back to 1943.

    Research on artificial intelligence began with a workshop that was held in 1956 at Dartmouth College in New Hampshire. This workshop involved demonstrations by computers that were able to engage in games and learning. At this early stage, artificial intelligence research was largely funded by government agencies like the US Department of Defense, although funding subsequently fell off dramatically and was not revived into the 1970s and 80s. This revival was associated with increasing applications of computer technology in modern life.

13. **Why is there so much worry regarding advancements in artificial intelligence?**

    Ans. There are many arguments about why advancements in artificial intelligence are the source of so much fear among a large group of people. This worry essentially stems from fears that, if human beings are supplanted as the most intelligent, powerful agents on planet Earth, then our own existence may be threatened. In fact, as human beings, we project some of our own motivations, thought processes, and behavior onto machines operating with artificial intelligence.

In short, we presume that artificial intelligence agents would behave the same way that human beings would behave. Human beings may like to imagine that they behave in a way that is beneficial to the world, but human beings often engage in behaviors that are harmful not only to other species but to our own species, as well as the world itself. There has long been a tendency to imbue technology that humans create with human qualities, and at least some of the concern regarding artificial intelligence stems from this.

14. Are concerns about artificial intelligence warranted?

Ans. There is no question that every day our capacity to understand just how miraculous artificial intelligence is, can (and will) grow. As we understand how the human mind works (including neural networks), we are continually fascinated by the idea that agents of our design are able to think much faster and potentially more logically than human beings, even when we constrain the agents in various ways. Artificially intelligent agents are able to beat human beings at games like chess and Go, and they are able to seemingly predict human thoughts and behaviors even before human beings know what they are doing or why. Are concerns about AI warranted? Absolutely!

15. Will machines using artificial intelligence one day take over the world?

Ans. Many science fiction writers, screenwriters, and movie directors have speculated on this idea, that artificial intelligence will one day pose some sort of threat to humanity. Even when the ancient Greeks imagined the automatons fashioned by Hephaestus, they perceived these automatons to be representing some sort of threat to human beings. Perhaps it is human nature to perceive the environment as filled with threats that need to be destroyed or at least curtailed in some ways.

To be honest, it is difficult to say whether artificially intelligent agents will one day take over the world. If you have watched "The Matrix," then perhaps you believe that artificially intelligent agents have already taken over the world and that you are currently experiencing a machine-induced delusion. Whatever you believe, the reality is that artificial intelligence is already a part of our lives. Perhaps a world in which artificial intelligence takes over is not too different from the sort of world in which we live now (and, no, I do not mean that in a Matrix type of way).

16. **What is all this singularity business about?**

Ans. Singularity is the term used to describe a hypothetical future event that involves artificially intelligent machines. Singularity essentially refers to the future point at which the intelligence of AI agents would exceed the collective intelligence of the human race. This represents an irreversible event that would permanently change the course of human civilization and life on Earth.

Much of the fear around singularity has to do with the reality that many regard computers, machines, and artificial intelligence with suspicion. Many imbue machines, computers, and artificial intelligence with human qualities. They perceive that an agent that is basically, what some may call, omniscient may decide that human beings represent a threat to life, the universe, or to the machines themselves and would, therefore, take actions against human beings. The strain running deep within the fears of singularity occurring is the idea that artificially intelligent agents would reach some form of inevitable conclusion.

17. **Why is it important for machines to learn?**

Ans. Machine learning (and more complex deep learning layered processes) involves machines making changes to their structure or programming on the basis of interactions that the machines have with their environment. This type of learning is important in artificial intelligence programs because some types of tasks may require learning for the program to be able to complete the task effectively, even with its programming.

For example, think about how poorly a human being may complete a challenging game or puzzle if they were unable to learn from their past actions. Imbuing artificially intelligent agents with the ability to learn essentially improves the functioning of the program. Learning also involves understanding the correlations between things in the environment, information that the artificially intelligent agent stores in the form of .dat files or other data. This data can be accessed by people for the purposes of data mining.

18. **What is the relationship between quantum computers and artificial intelligence?**

    Ans. Quantum computers store memory using qubits instead of bits, which allows them to exceed the potential of traditional computers. Quantum computers use quantum-mechanical phenomena, which allow these computers to perform simulations that classical computers cannot. Quantum computers first appeared about twenty years ago, and some experts believe that we are only a few years away from quantum computers becoming online. If we see artificial intelligence as being a program that knows how to think, behave, and learn, then quantum computing is the brain that powers these essential motivations. For this reason, some in the scientific community perceive that the real danger is not artificial intelligence but quantum computers.

19. **What is the Turing test?**

    Ans. Turing test was a concept created by British computer scientist Alan Turing in 1950. Although there are different forms of the Turing test, and even some disagreement about what should or should not be considered a true Turning test, this type of test can be thought of as answering the question of whether a machine is capable of thinking like a human being.

    The Turing test follows a basic format. There are three actors in an interrogation, two of which are human and one is a computer. If a human being cannot tell the computer from the other human being, then the computer has successfully begun to think like a human being. In its basic form, this interrogation features a human interrogator reading answers generated by a human being and by a computer. In a way, a Turing test seeks to determine if a computer can mimic a human being rather than if it has the intellectual capacity of human beings.

20. **What is the Voight-Kampff test?**

Ans. The Voight-Kampff test is the version of the Turing test explored in the movie "Blade Runner." In this movie, artificially intelligent beings were shown to become so advanced that they are virtually indistinguishable from humans. This test, like the Turing test, involves the response of the subject to questions, but several measures are used to determine the emotional response, including pupillary dilation response. The existence of the Voight-Kampff test implies a future point where human beings will have difficulty being able to tell artificially intelligent agents from humans, therefore requiring advanced tests to make the distinction.

Conclusion

By now you should be able to apply various machine learning algorithms for supervised learning, unsupervised learning, and reinforcement learning. This knowledge is beneficial when you want to explore other problems regarding machine learning. Having said that, it is important to give you some more words of advice, especially on how to approach a problem regarding machine learning.

Approaching a machine learning problem

One must be very orderly especially when approaching a machine learning problem for the first time. Endeavor to answer the following questions:

- What and how do I measure for my working model (e.g., a fraud prediction model)?
- Do I have the knowledge to evaluate a specific algorithm?
- What is the final implication (businesswise or academically) of my model in case I am successful?

Humans in the loop

Identify and assign roles for each and every human participant in your model appropriately.

Move from prototype to production.

Build prototypes using the library features of `scikit-learn`, and engage a production team who should be working with programming languages such as Python, R, Go, Scala, C#, C++, and C.

Testing production systems

Before you release your product into the market, make sure you carry out proper testing to ascertain whether your product is working properly. It is in this phase that you can identify various faults and even points of improvement.

Next step

This book provides such a perfect illustration of machine learning that, at the end of it, you expect to be a machine learning expert. However, if you need more information, feel free to consult other literature about those concepts that may not be very clear to you. Some good books may include *The Elements of Statistical Learning by Friedman et al.*, and *Machine Learning: An Algorithmic Perspective by Stephen Marsland.*

Machine learning packages

Currently there are a number of machine learning packages in the market that are meant to boost your understanding of various algorithms. Some of the most common packages are `scikit-learn` (most interactive and obviously favorite), `statsmodels` package, vowpal wabbit (vw), and finally `mllib` (a Scala library built on Spark).

Use recommender systems and other kinds of machine learning systems to identify the best of the best. Also carry out probabilistic modeling, probabilistic programming, and interfacing to increase your knowledge of machine learning algorithms. Other highly recommended topics, though very complex, would be neural networks, data scaling, and deep learning.

It is my hope that you are now convinced about machine learning as the new niche for building new applications and objects (project manipulation). Keep digging into this new area and never give up; a lot more is yet to come.

Machine learning is an active research subject, in particular, artificial neural networks. Nowadays, machine learning is used in every domain, such as marketing, health-care systems, banking systems, stock market, and gaming applications, among others. This book's objective is to provide a basic understanding of the major branches of machine learning, as well as the philosophy behind artificial neural networks.

Where do we go from here?

We have already established how wonderful machine learning is and what it can do today. What's even more amazing is what it will be able to do in the future. Yes, the future looks very bright for so many reasons. Of course, there are many predictions (probably created by machine learning) of what we can expect, but chances are, when the next evolution has passed, we'll probably all be utterly surprised, standing on the sidelines muttering, "I didn't know machines could do that!"

But what are the predictions for the future? What do we know now that we can confidently keep a watchful eye out for?

- **Quantum computing:** Right now, machine learning is mostly used for problem-solving. They manipulate and classify data at incredible speeds. In the future, quantum computers will be better equipped to manipulate high-dimensional vectors. They will accomplish this by using hybrid training methods. By utilizing a blend of supervised and unsupervised algorithms, there will be a huge increase in the number of vectors resulting in a highly impressive rate of speed.
- **Improved unsupervised algorithms:** Their ability to discover hidden patterns in data on its own and their self-learning techniques make it possible for unsupervised learning to be utilized more fully in the future. Machines of the future will be built smarter and mostly unsupervised.
- **Collaborative learning:** Machines will have an enhanced ability to use other computational entities in a collaborative manner. This will allow them to produce better results than what is already being achieved now.
- **Deeper personalization:** In the future, machines will know much more about you personally. While we may think this is very annoying and an invasion of our privacy, the

feeble attempts used today will be greatly enhanced. Those frustratingly inaccurate recommendations will be a thing of the past, ending our frustrations with the whole process after all.

- **Cognitive services:** No doubt, we will see many more intelligent features appear in even the most everyday machines. Computer scientists are already working on emotion detection systems, speech recognition, vision recognition, and so much more.

No doubt, you will be able to think of many more possibilities for machine learning in the future, but as I said before, there is a good chance that the majority of the world will be surprised at what will emerge.

Index

A

agent-environment interface 97
AI-based trends
  in banking 126
AI/robotics 114
AI technologies
  development status 122-124
algorithmic exchanging 134
anomaly-based
    intrusion detection 31
anomaly detection 32, 33
anomaly detection method 60
artificial intelligence (AI)
  about 4, 5, 197
  advancements 202, 203
  business and
    finance role 200, 201
  concept 202
  connection, between
    robotics 200
  context agent 198
  cost of investing 120
  crucial leadership 120, 121
  in banking 125, 126
  inorganic growth 121
  integrating 201
  need for 199
  neural networks, types in 77
  strategy 119, 120
  types 198, 199
  using 203, 204
  versus humans 172
  versus quantum
    computers 205
  warranted 203
artificial intelligence (AI),
    impact
  cybersecurity threats 171
  disinformation 170, 171
  humanity 169, 170

inequality 169
loss of jobs 168, 169
racist robots 171, 172
artificial intelligence (AI), in banking
cost cutting 174
customer service 175
fraud detection 174
internet banking 176
risk management 175
artificially intelligent agents 198
artificial neural networks
about 200
benefits 79
ATMs 143, 144
ATMs, risks
employee records 144
fake calls 144
loss of cards 144
social media 144
Australian Securities Exchange (ASX) 116
automatic cluster detection
Gaussian model 24-26
in data mining 19, 20
scatter map 21-24
simple structure, searching 20, 21
automation
about 105
FinTech 106-108
robots 105, 106

B

Bank of America
big data projects, using 117
Bayesian classification 16
big data
development directions 37-39
future 36, 37
big data analytics 142
biometric recognition 59
about 56-58
advantages 57
biometric trait, characteristics
acceptability 58
collectability 58
cost 58
inter-intra-class performance 58
universality 58
Blockchain
about 115
in banking industry 115
Blockchain technology
advantages 116

C

card security
tips 145
CatBoost 80
category data 69
chatbot 114
cognitive computing 115
cognitive services 209
collaborative learning 208
computer vision application
about 50, 51
neural networks 51
overview 52-54
concepts, machine learning disadvantages
adaptive control theory 156
artificial intelligence (AI) 157
evolutionary models 157
neural networks, creating 156
psychological modeling 156
statistics 156
conditional random field (CRF)
algorithm 85, 195, 196
convolutional layers 52
convolutional neural network (CNN)
about 51, 78

algorithm 84
convolutional layers 52
fully connected layers 52
pooling layers 52
criminal investigation 149
customer relationship management 149
customer segmentation 150
cybersecurity
machine learning, application 26, 27
cybersecurity, problems and solution approaches
anomaly detection 32, 33
malware detection 29-31
phishing page detection 27-29
spam detection 27

D

data
hacking 142
hacking, at retial store/restaurants 143
hacking, on online portals 143
hacking, through mail accounts 143
data handling
in natural language processing (NLP) 45, 46
data mining
about 6, 148
automatic cluster detection 19, 20
education 149
health care system 148
versus machine learning 6
data reduction
key predictors, identifying 93, 94
data structuring
about 36
semi-structured 36
structured 36
unstructured 36
data visualization
about 148
education 149
health care system 148
DDoS attack detection 31, 32
decision tree algorithm 196
deeper personalization 208
deep learning 5, 202
denial-of-service (DoS) 31
digital bank 118
distributed denial-of-service (DDoS) 31
DoS attack detection 31, 32

F

feed-forward neural networks
about 77, 78
algorithm 84
financial rules and designs
improving 124, 125
FinTech
about 106-108
challenges 129
impact 109-113
technology, impact 129
FinTech companies
overview 128, 129
forced linear regression algorithm 195
fraud detection 149
fraud detection and prevention 124
fully connected layers 52

G

Gaussian model 24-26
GBM
about 79

CatBoost 80
Light GBM 79
XGBoost 80
general applications, machine learning
about 160
customer lifetime value prediction 162
cybersecurity 165
dynamic pricing 161
email spam 165
financial analysis 160
financial evaluation 162
fraud detection 160, 161
Google Translate 162
government 165
image recognition 161
medical diagnosis 161, 162
online customer support 163
predictive maintenance 160
product recommendations and guidance 163
search engines 164
self-driving cars 164
social media 163
traffic alerts 164
video surveillance 165
virtual personal assistants 166
gradient boosting algorithms
about 79
GBM 79
gradient descent 80, 81
gradient descent 80, 81

H

Hamming distance 192

I

ImageNet 55
image recognition 54-56
input data 69
internet banking
security tips 145
internet of things (IoT)
in banking 117
intrusion detection 150

K

k-means clustering
about 184-187
distance and similarity 189, 190
overview 187, 188
k-means method
drawbacks 24
KNN algorithm
about 190
overview 194
training data, preparing 193, 194
working 190-193
know your customer (KYC) principle 113

L

least squares algorithm 195
lie detection 150, 151
Light GBM 79
linear regression
about 182, 183
applications 179
multiple linear regression 183
simple linear regression 183
live by
guidelines 113, 114
logistic regression 183, 184
loop 206

M

machine learning
about 3-6, 201, 204, 205

application,
  in cybersecurity 26, 27
applications 136
cryptocurrencies 138
data abstraction 3
data input 3
day trading 138, 139
development
  and combination 10
factors 4
generalization 4
guaranteeing and
  credit scoring 133, 134
illustration 207
importance, in finance 7
in finance 6
information and business
  space, utilizing 135
need for 208, 209
outsider arrangements,
  utilizing 9, 10
packages 207, 208
problem, approaching 206
problem types 92, 93
security risks 133
use cases, in banking 132
using, in financial
  service 10, 11
utilizing, in finance 8, 9
versus data mining 6

machine learning, advantages
  about 154
  data sets, handling 155
  efficiency, improving 154
  environment modification,
    adapting 155
  pattern, identifying 154
  tasks 154
machine-learning-based
    detection algorithms
  features 28
machine learning,
    disadvantages
  about 155
  concepts 156
machine learning, financial
    applications
  about 136
  algorithm trading 137
  fraud, detecting 137
  insurance or
    credit insurance 137
  portfolio management 136
maline tool 31
malware detection 29-31
malware detection, classes
  adware 29
  backdoor 29
  keylogger 29
  rootkit 29
  spyware 29
  virus 29
  worm 29
Manhattan distance 193
Minkowsky distance 193
modeling process
  about 86
  feature engineering 86, 87
  model, selecting 86, 87
  model, training 87
  model, validating 87-89
  observations, predicting 89
model learning 97
modular neural networks 78

N

Naive Bayes
  about 15
  Bayesian classification 16
Naive Bayes algorithm
  applications 17
  parameters 16, 17
  recommendation system 18

strengths 16, 17
weaknesses 16, 17
natural language processing (NLP)
data handling 45, 46
key advantages 45
uses 44, 45
neural networks
about 70-77
activation function 51
data preprocessing 51
in computer vision application 51
weight initialization 51
working 64-69
neural networks, types
convolutional neural network (CNN) 78
feed-forward neural networks 77, 78
in artificial intelligence (AI) 77
modular neural networks 78
recurrent neural networks 78
NLP application
about 46
email filtering 47
grammar checking 47
health care 46
online search autocomplete 47
online search autocorrect 47
social media monitoring 47
survey analysis 46
talent recruitment 46
voice assistants 46
normal distribution 18, 19

O

online banking
information, obtaining 142

P

pattern recognition method 60
phishing page detection 27-29
policy function 98
polynomial fitting 195
pooling layers 52
Principal component analysis (PCA) algorithm 194
production systems
testing 207

Q

quantum computers
versus artificial intelligence (AI) 205
quantum computing 208

R

recurrent neural networks (RNN)
about 78
algorithm 85
regression techniques
about 178-182
linear regression 182, 183
logistic regression 183, 184
reinforcement learning (RL)
about 94, 95
agent 97
algorithms 98, 99
applications 99, 100
examples 94, 95
fundamentals 96, 97
reinforcement learning (RL) algorithms
about 99
model-based 100
policy-based 100
value-based 99

reinforcement learning
(RL) types
about 100
negative 100
positive 100
risk management 124
robo-advisors 134, 135
Robo-warning 7

S

security measures 144
signature-based intrusion
detection 31
singularity 204
skimming 143
software vulnerabilities 59-61
spam detection 27
static analysis
techniques 30
statistics 5
structured data 40-42
support vector machine
(SVM) algorithm 195

T

technologies
about 114
AI/robotics 114
Bank of America, with
big data projects 117
Blockchain 115, 116
Blockchain, in
banking industry 115
Blockchain user 116, 117
chatbot 114
cognitive computing 115
internet of things (IoT),
in banking 117
voice biometrics 118
technology and finance 104, 105
Trading 125
training data
preparing, for
KNN algorithm 193, 194
Turing test 205

U

unstructured data 40-42
unsupervised algorithm 208

V

value function 98
value function, types
deterministic stationary
policy 98
optimal value function 98
stationary policy 98
voice biometrics 118
Voight-Kampff test 206
vulnerability detection methods
anomaly detection method 60
pattern recognition method 60

W

word embedding 84

X

XGBoost 80