FDP INSTITUTE®

FDP Institute Study Guide

October 12 – November 8, 2020

Brought to you by CAIA ASSOCIATION®

FDP Institute Study Guide

October 12 – November 8, 2020

Brought to you by CAIA ASSOCIATION

Outline of FDP Exam Study Guide Contents

Introduction to the FDP Program

The FDP Institute® was founded by the Chartered Alternative Investment Analyst Association® to create the FDP® charter. It is the only globally recognized professional designation in the area of financial data science, an increasingly important part of the financial services industry.

In recent years, the financial industry has been disrupted by the digital revolution. It is critical for industry practitioners to have a working knowledge of the increasingly important roles played by big data, machine learning, and artificial intelligence in the financial industry. The FDP Institute has designed this self-study program to provide the finance professional with an efficient path to learn about the essential aspects of financial data science. The FDP curriculum introduces candidates to the central concepts of machine learning and big data, including ethical and privacy issues, and their roles in various segments of the financial industry. Candidates will earn their FDP Charter once they pass the FDP exam and complete two short online classes, which can be done before or after the FDP exam.

The university faculty and industry practitioners who have helped create the FDP Charter program bring years of experience in the financial services industry. Consequently, the curriculum is consistent with recent advances in the applications of data science to the financial industry.

Passing the FDP examination is an important accomplishment and will require a significant amount of preparation. All candidates will need to study and become familiar with the FDP curriculum material to develop the knowledge and skills necessary to be successful on examination day.

This study guide is organized to facilitate quick learning and easy retention. Each topic is structured around learning objectives that define the content to be tested on the exam. The learning objectives are an important way for candidates to organize their study, as they form the basis for examination questions. All learning objectives reflect the content in the FDP curriculum, and all exam questions are written to address the learning objectives directly. A candidate who can meet all learning objectives in the study guide should be well prepared for the exam. For these reasons, we believe that the FDP Institute has built a rigorous program with high standards, while also maintaining an awareness of the value of candidates' time.

Candidates for the FDP Charter are required to complete both the FDP exam and the online classes. Since the FDP program is designed for finance professionals, it is assumed that candidates have an understanding of finance. This includes awareness of the roles and characteristics of various financial institutions and instruments as well as the financial models employed by these institutions to value the instruments and measure risk. These concepts are covered in CAIA®, CFA®, and FRM® exams, and dedicated undergraduate or graduate courses covering financial markets, investments, and risk management.

FDP Program: Online Requirements

While the FDP exam will not contain any coding questions, all candidates must complete the following two components before they can obtain their FDP charters. The classes can be completed before or after a candidate sits for the FDP exam:

- **FDP Exam**
- **Two online classes covering the basics of Python or R programming, or the single class offered by Metis.**

Online classes are offered by three approved providers. Depending on the candidate's background, the online classes are estimated to take 8-10 hours. No programming background is required to complete the online classes.

As of now, pre-selected online classes offered by the following organizations have been approved by the FDP Institute.

- **Datacamp:** https://www.datacamp.com/
- **Dataquest:** https://www.dataquest.io/
- **Metis:** https://www.thisismetis.com/

The list of online classes for each approved online provider appears on FDP Institute's website as well as in this document.

The approved online classes offered by Dataquest and Datacamp are available as soon as a candidate registers on their respective sites. The approved class offered through Metis is offered once a month throughout the year. All three providers offer limited free access to their classes. Candidates should take advantage of the limited free access to determine which platform's approach is more suited to their needs. Candidates cannot mix and match classes from different providers. Finally, the approved classes offered by the three platforms assume no prior knowledge of Python or R language or any specific computer programming language.

The Candidates' Handbook, which can be found on FDP's website, describes the procedure for sending proof of successful completion of the online classes to the FPD Institute. The following classes should be completed to satisfy the FDP Charter requirements.

DataCamp

Candidates can access Datacamp classes through their website at https://www.datacamp.com/. Candidates are responsible for the cost of classes offered at Datacamp. Candidates are encouraged to take advantage of limited free access offered by Datacamp to evaluate its method of teaching. The classes listed below are short and, depending on the candidate's background, each one can be completed within 4 - 6 hours. Candidates can choose between either two (2) R or two (2) Python classes.

DataCamp: Python

1. Introduction to Python

Python is a general-purpose programming language that is becoming ever more popular for data science. Companies worldwide are using Python to harvest insights from their data and gain a competitive edge. Unlike other Python tutorials, this course focuses on Python specifically for data science. In this Introduction to Python course, you'll learn about powerful ways to store and manipulate data, and helpful data science tools to begin conducting your own analyses.

https://www.datacamp.com/courses/intro-to-python-for-data-science

2. Intermediate Python for Data Science

Intermediate Python for Data Science is crucial for any aspiring data science practitioner learning Python. Learn to visualize real data with Matplotlib's functions and get acquainted with data structures such as the dictionary and the pandas DataFrame. After covering key concepts such as Boolean logic, control flow, and loops in Python, you'll be ready to blend everything you've learned to solve a case study using hacker statistics.

https://www.datacamp.com/courses/intermediate-python-for-data-science

DataCamp: R

1. Introduction to R

In Introduction to R, you will master the basics of this widely used open-source language, including factors, lists, and data frames. With the knowledge gained in this course, you will be ready to undertake your first very own data analysis. Oracle estimated over 2 million R users worldwide in 2012, cementing R as a leading programming language in statistics and data science. Every year, the number of R users grows by about 40%, and an increasing number of organizations are using it in their day-to-day activities.

https://www.datacamp.com/courses/free-introduction-to-r

2. Intermediate R

Intermediate R is the next stop on your journey in mastering the R programming language. In this R training, you will learn about conditional statements, loops, and functions to power your R scripts. Next, make your R code more efficient and readable using the application functions. Finally, the chapter utilities gets you up to speed with regular expressions in R, data structure manipulations, and times and dates. This course will allow you to take the next step in advancing your overall knowledge and capabilities while programming in R.

https://www.datacamp.com/courses/intermediate-r

DataQuest

Candidates can access DataQuest classes through their website at https://www.dataquest.io/. Candidates are responsible for the cost of classes offered at DataQuest. Candidates are encouraged to take advantage of limited free access offered by DataQuest to evaluate its method of teaching. The classes listed below are short and, depending on the candidate's background, each one can be completed within 4 - 6 hours. Candidates can choose between either two (2) R or two (2) Python classes.

DataQuest: Python

1. **Python for Data Science: Fundamentals**
   In our introductory course on Python for data science, you'll get an overview of the Python programming language and how you can use it for data science. You will learn to code using real-world mobile app data while learning key Python concepts such as lists and for loops. You'll also learn how to update variables, how to work with different kinds of data, how to manipulate Python dictionaries, and how to use custom functions to speed up your workflow. Additionally, we'll cover some coding best practices that'll help you build good habits right from the start, and show you how to use Jupyter Notebook, a popular tool used in the Data Science workflows for easy sharing of data science projects. At the end of the course, you will combine all the skills you have learned to create your data science portfolio project. In this guided project, you'll analyze different app profiles on the iOS App Store to make recommendations for the most profitable types of apps to develop.
   https://www.dataquest.io/course/python-for-data-science-fundamentals/

2. **Python for Data Science: Intermediate**
   In our Python for Data Science Intermediate course, we'll cover some essential techniques for working with the Python programming language for data science. To start, you'll learn how to clean and prepare data in Python, a critical skill for any data analyst or data scientist job. To do this, you'll dig into some real-world data about the artwork at the Museum of Modern Art and learn to manipulate text, clean messy data, and more. You'll also get to practice summarizing numeric data and formatting strings in Python. Next, you will unlock the true power of Python as we dive into object-oriented programming (OOP) and how it relates to data science. Then, you'll apply this new understanding by building your class. Finally, you'll learn how clean, standardize, and analyze time-series data using Python's datetime module. At the end of the course, you will combine all the skills you learned to create a portfolio project centered around Hacker News post titles to find out what types of posts are most likely to be successful at what times.
   https://app.dataquest.io/course/python-for-data-science-intermediate/

DataQuest: R

1. **Introduction to Programming in R**
   In the world of data science, R is a popular programming language for a reason. It was built with statistical manipulation in mind, and there's an incredible ecosystem of packages for R that let you do amazing things – particularly in data visualization – that would be much more difficult in Python. In this free introductory course on R, you'll go hands-on with R for data science, learning critical R concepts such as matrices, vectors, lists, and more, and writing your code to practice them right in your browser window. And you'll learn all of this while working with real-world data, much as you would for a real data science project. You will also learn how to update variables, work with different kinds of data, and how to import data into R and save it as a dataframe. We'll also cover how to how to install packages to extend R's functionality for working with dataframes, a crucial skill for extending your data science toolkit. And you'll learn the basics of using R Studio, which is a popular free and open-source development environment that's widely used in the R data science community so that you can easily share projects.
   https://www.dataquest.io/course/intro-to-r/

2. **Intermediate R Programming**
   IIn our Intermediate Programming in R course, you will continue building your R data science skillset. We'll take you beyond the basics to enhance your understanding of R, supercharge your workflow, do some pretty neat stuff along the way. To start, you will learn how to use control structures in your R programming to control the flow of your code. Then, you will learn to work with vectorized functions to make the most of R's functionality. You will also learn how to use functions in your code to speed up your workflow and write better code to avoid common pitfalls. Next, you will learn about how to work with functionals and understand why they're suitable alternatives to loops, and you'll get hands-on practice with single and multivariable functions. Towards the end of the course, you will learn the basics of working with strings and string manipulation as you analyze with real-world data from the World Cup. By the time you get to the end of this course, you'll be quite comfortable with programming in R, and you'll have built the fundamental skills you need to dive into a variety of unique data science projects of your own!
   https://www.dataquest.io/course/intermediate-r-programming/

Metis

Candidates can access the Metis course through their website at https://www.thisismetis.com/. Candidates are responsible for the cost of the course offered at Metis. Candidates are encouraged to take advantage of free sample videos that are offered by Metis to evaluate its method of teaching.

Unlike the classes offered through Dataquest and Datacamp, which consist of pre-recorded videos and texts, Metis offers live online classes with dedicated instructors who are ready to answer your questions during the live sessions as well as later during office hours. Further, while Dataquest and Datacamp classes can be taken at any time by a candidate, Metis's live

Metis *continued*

classes are offered monthly throughout the year. Enrolled candidates will be able to watch a video of the class should they miss a session. The approved class offered by Metis lasts 6 weeks.

The single course offered by Metis and approved by the FDP exam is titled Beginner Python and Math for Data Science, and it consists of the following 6 topics, which are usually covered over 12 live sessions.
https://www.thisismetis.com/courses/beginner-python-and-math-for-data-science/

1. **Python Basics**
   Candidates are introduced to programming in Python. Candidates will learn about Jupyter Notebooks – a popular platform for running Python programs. This part of the course will cover the basics of programming, including data structures, data operations, if-else statements, for and while loops, and logical operations.

2. **Python Advanced**
   This segment of the course covers advanced functionality in Python, including functions, debugging, error handling, string manipulations, and writing efficient code.

3. **Python Mathematical Libraries**
   Candidates will learn about using libraries that are useful for data manipulation and visualization. Candidates will learn to use NumPy, Pandas, and Matplotlib. These libraries will allow candidates to load and save data, manipulate data such as aggregating, filtering, detecting outliers, and visualizing.

4. **Linear Algebra**
   This segment of the course is a refresher in linear algebra. It will cover the fundamentals of linear algebra, including vectors, and vector manipulations, matrices and matrix manipulations, linear equations and solutions, eigenvalues, and eigenvectors.

5. **Calculus and Probabilty**
   This module is a refresher in the fundamentals of calculus. It reintroduces students to such central concepts of calculus such as derivatives, integrals, determining local maximum and minimum, and limits. In addition, the module provides a refresher on central concepts of probability such as random variables, mean, variance, probability mass and density functions, and cumulative distribution functions.

6. **Statistics**
   This final refresher module covers a few important statistical concepts such as ANOVA, hypothesis testing and p-value, and confidence intervals.

FDP Examination

The FDP examination, administered twice annually, is a four-hour computer-administered examination that is offered at test centers throughout the world. The FDP examination is comprised of 75 multiple choice questions weighted as 60% of the total points and two to three constructed response questions (multi-part essay type) weighted as 40% of the total points. The FDP exam will not contain any Python or R programming questions.

The FDP examination is based on this study guide, which is organized to facilitate quick learning and easy retention. Each topic is structured around learning objectives and keywords that define the content to be tested on the exam. The learning objectives and keywords are an important way for candidates to organize their study, as they form the basis for examination questions. All learning objectives reflect the content in the FDP curriculum, and all examination questions are written to address the learning objectives directly.

For additional information about the FDP examination, please see the Candidate's Handbook, which can be found on the FDP Institute website.

The FDP Curriculum: Outline

Candidates for the FDP Charter will have to enroll in the self-study program created by the FDP Institute and follow its carefully designed study guide. To become an FDP Charterholder, candidates must pass the FDP exam and submit their certificates of learning of the required online classes. The rest of this document discusses the FDP curriculum. Below is the outline of the curriculum:

Topics		Approximate Weight %
1.	Introduction to Data Science & Big Data	5-10
2.	Machine Learning: Introduction to Algorithms	5-10
3.	Machine Learning: Regression,Support Vector Machine & Time Series Models	5-10
4.	Machine Learning: Regularization, Regression Trees, Random Forest & Overfitting	5-10
5.	Machine Learning: Classification & Clustering	5-10
6.	Machine Learning: Performance Evaluation, Backtesting & False Discoveries	5-10
7.	Data Mining & Machine Learning: Naïve Bayes & Text Mining	5-10
8.	Big Data & Machine Learning: Ethical & Privacy Issues	5-10
9.	Big Data & Machine Learning in the Financial Industry	30-50

Other Study Tools and Resources

In addition to this study guide and candidate's handbook, the FDP Institute website directs you to the readings that are covered in the curriculum. The readings are detailed below by topic area and include textbooks, often used across topics, as well as several individual articles that are usually topic-specific. Both types of readings can be purchased from Amazon or the publisher, and whenever possible, they are posted on the FDP Institute website. They will be freely available to registered candidates.

Page Number References for Keywords

For candidate's convenience, a set of six articles published by PMR Journals is provided in one collection titled Big Data & Machine Learning in the Financial Industry: Readings for the Financial Data Professional Exam and is available at a discounted price of $99 for registered candidates. In this collection, there are two sets of pages numbers: one corresponding to the collection's table of contents, and one corresponding to each article's page number in the original journal. The page numbers appearing next to the keywords refer to the page numbers as they appeared in the original article.

The FDP Curriculum: Reading List

Topic 1: Introduction to Data Science & Big Data

- Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc. Chapters 1 & 2. Retrieved from https://www.amazon.com/Data-Science-Business-Data-Analytic-Thinking-ebook/dp/B00E6EQ3X4
- Guida, T. (2019). Big Data and Machine Learning in Quantitative Investments. West Sussex, UK: John Wiley & Sons Ltd. Chapters 2, 4 & 5. https://www.amazon.com/Machine-Learning-Quantitative-Investment-Finance/dp/1119522196

Topic 2: Machine Learning: Introduction to Algorithms

- James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapters 1, 2.1 & 2.2. http://www-bcf.usc.edu/~gareth/ISL/ https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1461471370
- Nielsen, M. A. (2015). Using Neural Networks to Recognize Handwritten Digits. In Neural Networks and Deep Learning, Determination Press. Retrieved from http://neuralnetwork-sanddeeplearning.com/chap1.html

Topic 3: Machine Learning: Regression, Support Vector Machine & Time Series Models

- Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapters 3 & 4
- James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapter 3, Sections 1-3.
- Aas, K. and X. K. Dimakos. (2004). Statistical modeling of financial time series: An introduction. Oslo Norway: Norwegian Computing Center. Retrieved from https://www.nr.no/files/samba/bff/SAMBA0804.pdf Sections 1-4.

Topic 4: Machine Learning: Regularization, Regression Trees, Random Forest & Overfitting

- Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 5
- James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapters 6.1, 6.2, 8.1, and 8.2.

The FDP Curriculum: Reading List *continued*

Topic 5: Machine Learning: Classification & Clustering

- Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 6 & 7

Topic 6: Machine Learning: Performance Evaluation, Backtesting & False Discoveries

- Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 8.
- Arnott, R., C. B. Harvey, and H. Markowitz. (2019). A Backtesting Protocol in the Era of Machine Learning. Journal of Financial Data Science, 1(1), 64-74. DOI: **https://doi.org/10.3905/jfds.2019.1.064**
- Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, London, U.K.: Royal Society Open Science, Retrieved from **https://royalsocietypublishing.org/doi/full/10.1098/rsos.140216**
- López de Prado, M. (2019). A Data Science Solution to the Multiple-Testing Crisis in Financial Research. Journal of Financial Data Science, 1(1), 99-110. DOI: **https://doi.org/10.3905/jfds.2019.1.099**

Topic 7: Data Mining & Machine Learning: Naïve Bayes & Text Mining

- Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapters 9 & 10.
- Jurafsky, D. and J. Martin. (2018). Chapter 4. Naïve Bayes and Sentiment Classification, In Speech and Language Processing. Retrieved from **https://web.stanford.edu/~jurafsky/slp3/4.pdf**

Topic 8: Big Data & Machine Learning: Ethical & Privacy Issues

- Institute of Business Ethics. (2016, June). Business Ethics and Big Data (IBE Issue 52). London, U.K.
- Institute of Business Ethics. (2018, January). Business Ethics and Artificial Intelligence (IBE Issue 58). London, U.K.
- Institute of Business Ethics. (2018, May). Beyond Law: Ethical Culture and GDPR (IBE Issue 62). London, U.K.
- Loukides, M., M., H. Mason and DJ. Patil. Ethics and Data Science **Free e-book** **https://www.amazon.com/Ethics-Data-Science-Mike-Loukides-ebook/dp/B07GTC8ZN7**

The FDP Curriculum: Reading List *continued*

Topic 9: Big Data & Machine Learning in the Financial Industry

- Financial Stability Board. (2017) Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications. Retrieved from http://www.fsb.org/wp-content/uploads/P011117.pdf
- Monk, A., M. Prins, and D. Rook. (2019). Rethinking Alternative Data in Institutional Investment. Journal of Financial Data Science, 1(1), 14-31. DOI: https://doi.org/10.3905/jfds.2019.1.1.014
- Simonian, J., C. Wu, D. Itano and V. Narayanan. (2019). A Machine Learning Approach to Risk Factors: A Case Study Using the Fama–French–Carhart Model. Journal of Financial Data Science, 1(1), 32-44. DOI: https://doi.org/10.3905/jfds.2019.1.032
- Rasekhschaffe, K. and R. Jones. (2019). Machine Learning for Stock Selection. Financial Analyst Journal, 13 May 2019 Volume 75 Issue 3. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3330946
- Gu, S., B. Kelly, and D. Xiu. (2018). Empirical Asset Pricing via Machine Learning. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281018 https://www.bryankellyacademic.org/ http://dachxiu.chicagobooth.edu/download/ML.pdf **The paper posted on the FDP Institute's website will be the version used for exam questions.**
- López de Prado, M. (2018). The 10 Reasons Most Machine Learning Funds Fail. The Journal of Portfolio Management, 44 (6) 120-133; DOI: https://doi.org/10.3905/jpm.2018.44.6.120
- Harvey, C. R. and Y. Liu. (2014). Evaluating Trading Strategies. [Special 40th Anniversary Issue]. The Journal of Portfolio Management, 40(5), 108-118. DOI: https://doi.org/10.3905/jpm.2014.40.5.108
- Raman, J., and R. Lam (2019). Artificial Intelligence Applications in Financial Services https://www.oliverwyman.com/our-expertise/insights/2019/dec/artificial-intelligence-applications-in-financial-services.html PDF: https://www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2019/dec/ai-app-in-fs.pdf
- Zappa, D., M. Borrelli, G.P. Clemente, N. Savelli. Text Mining In Insurance: From Unstructured Data To Meaning https://www.variancejournal.org/articlespress/articles/Text_Mining-Zappa-Borrelli-Clemente-Savelli.pdf

Learning Objectives

Topic 1. Introduction to Data Science & Big Data

Readings

1.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc. Chapters 1 & 2. Retrieved from http://data-science-for-biz.com/ https://www.amazon.com/Data-Science-Business-Data-Analytic-Thinking-ebook/dp/B00E6EQ3X4

1.2 Guida, T. (2019). Big Data and Machine Learning in Quantitative Investments. West Sussex, UK: John Wiley & Sons Ltd. Chapters 2, 4 & 5. https://www.amazon.com/Machine-Learning-Quantitative-Investment-Finance/dp/1119522196 Guida, T. (2019).

Reading 1.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc. Chapters 1 & 2.

Keywords

Data mining (p. 2)
Data science (p. 4)
Churn (p. 4)
Data-driven decision making (p. 5)
Data engineering (p. 5, 7)
Data-analytic thinking (p. 12)
Target (p. 24)
Label (p. 24)
Unsupervised data mining (p. 24)
Supervised data mining (p. 25)

Learning Objectives

Demonstrate proficiency in the following areas:

1.1.1 Data analytic thinking (Ch. 1)

For example:

A. Discuss the ubiquity of data opportunities.
B. Compare and contrast data science, engineering, and data-driven decision making.
C. Explain data and data science capability as a strategic asset.
D. Describe data-analytic thinking.
E. Compare data science and the work of the data scientist.

Learning Objectives *continued*

1.1.2 Business problems and data science solutions (Ch. 2)

For example:

A. Describe how one transitions from business problems to data mining tasks.

B. Compare supervised methods to unsupervised methods.

C. Describe the difference between data mining and using the results of data mining.

D. Describe key aspects of the data mining process, including business understanding, data understanding, data preparation, modeling, and evaluation.

Reading 1.2 Guida, T. (2019). Big Data and Machine Learning in Quantitative Investments. West Sussex, UK: John Wiley & Sons Ltd. Chapters 2.1-2.4, 4.1-4.6 & 5.1-5.4.

Keywords

Quant quake (p. 110)
Quantamental investing (p. 127)
Fundamental law of active management (p. 127)
Exhaust data (p. 151)
Nowcasts (p. 151)

Note: These page numbers correspond to the e-book.

Learning Objectives

Demonstrate proficiency in the areas of:

1.2.1 Defining big data

For example:

A. Discriminate between alternative data and big data.

B. Contrast drivers of adoption of alternative data with its challenges in the investment community

C. Identify the largest categories of alternative data types in use today.

D. Evaluate the usefulness of an alternative data set.

E. Describe the likely attributes that differentiate alternative data sets in terms of cost.

F. Discuss some of the most prominent alternative data trends.

1.2.2 Implementing alternative data in an investment process (Ch. 4.1-4.6)

For example:

A. Describe the "quant quake" and how it motivated the search for alternative data.
- Note that Table 4.1 is cut off in the ebook version. The full version of the table appears at the end of this section

B. Discuss reasons for "the chasm" in the alternative data adoption life cycle and reasons that the chasm has been difficult to cross for many fund managers.

Learning Objectives *continued*

C. Discuss methods for improving the efficiency of evaluating data sets for finding alpha.

D. Describe issues involved with selecting a data source for evaluation within the context of a quant equity process.

E. Explain why and under what circumstances a fundamental prediction may be more appropriate than an asset price prediction when working with alternative data.

F. Apply the fundamental law of active management and describe how it applies to discretionary managers and how it applies to quant managers.

G. Describe the transition from fundamental analysis to "quantamental analysis."

H. Describe how alternative data can be used to generate a trading signal using examples including blogger sentiment, online consumer demand, transactional data, and environmental, social, and governance (ESG) data.

1.2.3 Using alternative and big data to trade macro assets (Ch. 5.1-5.4)

For example:

A. Define general concepts and terms for the use of big data and alternative data, including "exhaust data."

B. Compare traditional model building approaches and machine learning.

C. Discuss how big data and alternative data can be used to improve economic forecasts and "nowcasts."

D. Describe how case studies show that alternative data is related to the following types of macro data: US Treasury yields, implied volatility in the foreign exchange market, and investor anxiety.

Learning Objectives *continued*

TABLE 4.1
Average annualized return of dollar-neutral, equally-weighted portfolios of liquid US equities

	More crowded factors				Less crowded factors			
	Earnings yield (%)	Momentum (%)	Simple reversal (%)	Average (%)	TM1 seasonality (%)	CAM 1 volume (%)	CAM1 Skew (%)	Average (%)
2001-2007								
Avg. Ann return	11.00	14.76	35.09	20.28	8.64	3.60	17.10	9.78
Daily factor return in August 2007								
7 Aug. 2007	-1.06	-0.11	-0.34	-0.50	-0.06	0.33	-0.85	-0.19
8 Aug. 2007	-2.76	-4.19	0.23	-2.24	-0.21	-0.04	0.21	-0.01
9 Aug 2007	-1.66	-3.36	-3.41	-2.81	-0.29	-1.27	-0.23	-0.60
10 Aug. 2007	3.91	4.09	12.45	6.82	0.71	-0.01	1.70	0.80

Learning Objectives *continued*

Topic 2. Machine Learning: Introduction to Algorithms

Readings

2.1 James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapters 1, 2.1 & 2.2
http://www-bcf.usc.edu/~gareth/ISL/
https://www.amazon.com/Introduction-Statistical-Learning-Applications-statistics/dp/1461471370

2.2 Nielsen, M. A. (2015). Using Neural Networks to Recognize Handwritten Digits. In Neural Networks and Deep Learning, Determination Press. Retrieved from http://neuralnetworksanddeeplearning.com/chap1.html

Reading 2.1 James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapters 1, 2.1 & 2.2

Keywords

Statistical learning (p. 1)
Classification problems (p. 28)
Semi-supervised learning (p. 28)
Quantitative variables (p. 28)
Qualitative response (p. 28)
Binary response (p. 28)
Regression (p. 28)
Predictors (p. 29)
Mean squared error (MSE) (p. 29)
Test MSE (p. 30)
Test data (p. 30)
Training MSE (p. 30)
Flexibility (p. 31)
Degrees of freedom (p. 32)
Cross-validation (p. 33)
Expected test MSE (p. 34)
Bias (p. 35)
Bias-variance trade-off (p. 36)
Error rate (p.37)
Indicator variable (p. 37)
Training error (p. 37)
Test error (p. 37)
Bayes classifier (p. 37)
Conditional probability (p. 37)
Bayes decision boundary (p. 38)
Bayes error rate (p. 38)
K-nearest neighbors (p. 39)

Note: The page numbers for the keywords correspond to those indicated by a PDF reader when the web page is printed to a PDF file.

Learning Objectives

Demonstrate proficiency in the following areas:

2.1.1 Organization and resources of the book An Introduction to Statistical Learning: with applications in R (Ch. 1)

This chapter is assigned to facilitate your studies, but no exam questions will be drawn from this chapter.

Learning Objectives *continued*

2.1.2 Statistical learning (Ch. 2.1)

For example:

A. Explain why we estimate a function with data, including the role of input and output variables and their synonyms, as well as error terms (reducible and irreducible), expected values, and variance.

B. Compare and contrast parametric and non-parametric learning methods.

C. Describe the trade-offs between prediction accuracy, flexibility, and model interpretability, including the role of overfitting.

D. Determine when a supervised learning model is preferable to unsupervised or semi-supervised learning models.

E. Explain how the appropriateness of regression problems relative to classification problems may be related to whether responses are quantitative or qualitative.

2.1.3 Assessing Model Accuracy (Ch. 2.2)

For example:

A. Recognize and explain the equation for mean squared error.

B. Explain the goal of measuring the quality of fit by minimizing training and test mean square errors (MSEs) and the implications of different levels of flexibility (degrees of freedom) for both training and test MSEs.

C. Explain the purpose of cross-validation.

D. Explain the bias-variance trade-off with an MSE decomposition into three fundamental quantities.

E. Explain the salient features of a simple Bayes classifier (for two classes), including the Bayes decision boundary and Bayes error rate.

F. Explain how the K-nearest neighbors classifier is related to the Bayes classifier and how the choice of K impacts results.

Reading 2.2 Nielsen, M. A. (2015). Using Neural Networks to Recognize Handwritten Digits. In Neural Networks and Deep Learning, Determination Press.

Learning Objectives *continued*

Keywords

Perceptron neurons (p. 3)
Weights (p. 4)
Threshold value (p. 4)
Layer (p. 6)
Bias (p. 6)
NAND gate (p. 7)
Input layer (p. 9)
Learning algorithms (p. 10)
Sigmoid neuron (p. 11)
Sigmoid function (p. 12)
Activation function (p. 14)
Input neurons (p. 16)
Output neurons (p. 16)
Hidden layer (p. 16)
Multilayer perceptrons (p. 16)
Feedforward neural networks (p. 17)
Recurrent networks (p. 17)
Cost function (p. 24)
Loss function (p. 24)
Objective function (p. 24)
Quadratic cost function (p. 25)
Mean squared error (MSE) (p. 25)
Gradient descent algorithm (p. 25)
Gradient vector (p. 29)
Learning rate (p. 30)
Stochastic gradient descent (p. 34)
Mini-batch (p. 34)
Epoch (p. 35)
Validation set (p. 37)
Hyper-parameters (p. 37)
Deep neural networks (p.55)

Note: These page numbers refer to the pages indicated in a PDF reader when the webpage is printed to a PDF file.

Learning Objectives

Demonstrate proficiency in the following areas:

2.2.1 Motivation for using neural nets to recognize handwritten digits

For example:

A. Describe the use of a training set as an alternative to a rules-based program to recognize digits.

2.2.2 Perceptron neurons

For example:

A. Calculate the output of a perceptron neuron.

B. Describe the intuition of a perceptron as a decision-making device.

C. Describe a perceptron as a NAND gate and what it implies for perceptron networks concerning computing logical functions.

D. Explain how perceptron neurons are more than new types of NAND gates.

Learning Objectives *continued*

2.2.3 Sigmoid neurons

For example:

A. Recognize a limitation of perceptron neurons that can be overcome by sigmoid neurons.

B. Recognize and differentiate perceptron neurons from sigmoid neurons.

C. Calculate the output of a sigmoid function, which is also referred to as a logistic function.

D. Explain the importance of the smoothness of the sigmoid function.

2.2.4 The architecture of neural networks

For example:

A. Identify components of a simple network with appropriate terminology.

B. Describe the central feature of a feed-forward network.

C. Compare and contrast feedforward networks with recurrent networks.

2.2.5 A simple network to classify handwritten digits

For example:

A. Argue for a natural order for solving the two problems of segmenting digits and classifying digits.

B. Calculate the required input neurons for classifying an individual digit in an image of a specific size in pixels.

C. Explain the choice to use ten output neurons instead of four for classifying an individual digit.

2.2.6 Learning with gradient descent

For example:

A. Recognize a quadratic cost function of weights and biases and alternative terminology for the cost function.

B. Explain why minimizing a quadratic cost function is preferable to working with other types of cost functions.

C. Recognize an equation for an update rule that defines the gradient descent algorithm and explain the purpose of each component in the equation.

D. Explain how quickly stochastic gradient descent can speed up learning given a training set size n and a mini-batch size, m.

Learning Objectives *continued*

2.2.7 Implementing a network to classify digits

For example:

A. Understand the role of hyper-parameters and their impact on output for each epoch.

2.2.8 Why deep learning matters

For example:

A. Describe deep learning in terms of neural networks and their performance relative to networks that are not based on deep learning methods.

Learning Objectives *continued*

Topic 3. Machine Learning: Regression, Support Vector Machine, Time Series Models

Readings

3.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapters 3 & 4

3.2 James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapter 3, Sections 1-3

3.3 Aas, K. and X. K. Dimakos. (2004). Statistical modeling of financial time series: An introduction. Oslo Norway: Norwegian Computing Center. Sections 1-4. Retrieved from https://www.nr.no/files/samba/bff/SAMBA0804.pdf

Reading 3.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapters 3 & 4.

Keywords

Information (p. 43)
Tree induction (p. 44)
Predictive model (p. 45)
Descriptive modeling (p. 46)
Target variable (p. 46)
Attributes or features (p. 46)
Model induction (p. 47)
Deduction (p. 47)
Training data (p. 47)
Labeled data (p. 47)
Supervised segmentation (p. 48)
Information gain (p. 51)
Entropy (p. 51)
Parent set (p. 52)
Child set (p. 52)
Variance (p. 56)
Entropy graph/chart (p. 58)
Decision nodes (p. 63)
Classification tree (p. 63)
Regression tree (p. 64)
Probability estimation tree (p. 64)
Tree induction (p. 64)
Decision surface or boundary (p. 69)
Frequency-based estimation of class membership probability (p. 72)
Laplace correction (p. 73)
Parameter learning or parametric modeling (p. 81)
Linear classifier (p. 85)
Linear discriminant (p. 86)
Hyperplane (p. 86)
Parameterized model (p. 86)
Objective function (p. 88)
Margin (p. 92)
Support vector machine (SVM) (p. 92)
Hinge-loss (p. 94)
Zero-one loss (p. 95)
Squared error (p. 95)
Odds ratio (p. 98)
Log-odds (p. 99)
Logistic function (p. 101)
Nonlinear SVM (p. 107)
Neural networks (p. 108)

Learning Objectives *continued*

Learning Objectives
Demonstrate proficiency in the following areas:

3.1.1 Models, induction and prediction (Ch. 3)

For example:

A. Define the information and tree induction.
B. Define prediction in the context of data science.
C. Compare and contrast predictive modeling with descriptive modeling.
D. Define attributes or features.
E. Describe model induction.
F. Compare and contrast induction with deduction.
G. Define the training data and labeled data.

3.1.2 Supervised segmentation (Ch. 3)

For example:

A. Describe supervised segmentation.
B. List the complications arising from selecting informative attributes.
C. Define entropy and information gain.
D. Calculate the value of entropy.
E. Recognize and apply entropy with the maximum and minimum disorder.
F. Contrast parent set with child set.
G. Calculate information gain for a child relative to a parent.
H. Discuss the issues with the numerical variables for supervised segmentation.
I. Define variance and discuss its application to numeric variables for supervised segmentation.
J. Define an entropy graph/chart.
K. Describe how an entropy chart can be used to select an informative variable.
L. Define a classification tree, decision nodes, a probability estimation tree, and tree induction.

3.1.3 Visualizing segmentations (Ch. 3)

For example:

A. Define a decision surface or decision boundaries.
B. Describe the relationship between the decision surface and the number of variables.
C. Define frequency-based estimation of class membership probability.
D. Calculate probability at each node of a decision tree.

Learning Objectives *continued*

E. Describe how Laplace correction is used to modify the probability of a leaf node with few members.

F. Calculate the value of the Laplace correction.

3.1.4 Classification via mathematical functions (Ch. 4)

For example:

A. Define a linear classifier.

B. Recognize and apply the equation of a straight-line using the slope and intercept.

C. Define a linear discriminant.

D. Describe decision boundaries in 2-dimensions, 3-dimensions, and higher dimensions.

E. Interpret the magnitude of a feature's weight in a general linear model.

F. Describe how linear discriminant functions can be used for scoring and ranking instances.

G. Describe the objective function of the Support Vector Machine (SVM).

H. Describe the important ideas behind the SVM.

I. Define margin for the SVM.

J. Define the hinge-loss function and zero-one loss function and squared error.

K. Describe the reason for not using a squared loss function in classification problems.

3.1.5 Regression via mathematical functions (Ch. 4)

For example:

A. Describe the major drawback of the least-squares regression.

B. Define odds and log odds.

C. List the important features of the logistic regression.

D. Recognize and apply the logistic function.

E. Describe the shape of the logistic function.

F. Describe how an objective function is formed in logistic regression.

G. Compare and contrast classification trees with linear classifiers.

H. Explain the basic idea behind nonlinear SVMs and neural networks.

Reading 3.2 James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapters 3.1, 3.2, 3.3

Learning Objectives *continued*

Keywords

Residual (p. 62)
Residual sum of squares (p. 62)
Population regression line (p. 63)
Least squares line (p. 63)
Bias (p. 65)
Unbiased (p. 65)
Standard error (p. 65)
Residual standard error (p. 66)
Confidence interval (p. 66)
Null hypothesis (p. 67)
Alternative hypothesis (p. 67)
t-statistic (p. 67)
R2 statistic (p. 70)
Total sum of squares (p. 70)
F-statistic (p. 75)
Forward selection (p. 78)
Backward selection (p. 79)
Mixed selection (p. 79)
Prediction interval (p. 82)
Dummy variable (p. 84)
Additive linear (p. 86)
Hierarchical principle (p. 89)
Polynomial regression (p. 90)
Residual plot (p. 93)
Heteroscedasticity (p. 95)
Outlier (p. 96)
Collinearity (p. 99)
Multicollinearity (p. 101)
Power (p. 101)
Variance inflation factor (p. 101)

Learning Objectives

Demonstrate proficiency in the following areas:

3.2.1 Simple linear regression (Ch 3.1)

For example:

A. Define a residual and a residual sum of squares (RSS).
B. Calculate the value of RSS.
C. Recognize and apply the least-squares coefficient estimates.
D. Interpret the least-squares coefficients.
E. Define the population regression line and least-squares line.
F. Define the concept of bias and unbiased estimators.
G. Define standard error and residual standard error.
H. Calculate the standard error of a statistic.
I. Calculate the 95% confidence interval.
J. Describe null and alternative hypotheses.
K. Calculate the t-statistic.
L. Assess the accuracy of linear regression.
M. Calculate the R2 statistic given TSS and RSS.
N. Interpret the given values of R2.
O. Describe the relationship between R2 and correlation.
P. Define the total sum of squares.

Learning Objectives *continued*

3.2.2 Multiple linear regression (Ch 3.2)

For example:

A. Interpret the coefficients of multiple linear regression.

B. Describe how the relationship between responses and predictors is tested in multiple linear regression.

C. Calculate the F-statistic given TSS, RSS, n, and p.

D. Describe how to determine the importance of variables in a given multiple regression.

E. Define forward selection, backward selection, and mixed selection.

F. Describe the tools used to examine model fit for multiple regression.

3.2.3 Considerations in the regression model (Ch 3.3)

For example:

A. Define dummy variables.

B. Describe how to use qualitative variables with more than two levels in multiple regression.

C. Interpret the coefficients of a dummy variable.

D. Describe additive and linear assumptions for the linear regression model.

E. Define the interaction effect.

F. Interpret the coefficients of an interaction term.

G. Describe the hierarchical principle for multiple regression.

H. Define polynomial regression.

I. Describe the potential problems, such as non-linearity, correlation of error terms, a non-constant variance of error terms, outliers, high-leverage points, and collinearity, for a linear regression model.

J. Describe the limits for high leverage for simple regression.

K. Define heteroscedasticity.

L. Define the power of a hypothesis test.

M. Define multicollinearity and the variance inflation factor.

N. Describe the range of values for the variance inflation factor.

O. Calculate the variance inflation factor.

Reading 3.3 Aas, K. and X. K. Dimakos. (2004). Statistical modeling of financial time series: An introduction. Oslo Norway: Norwegian Computing Center. (Sections 1-4)

Learning Objectives *continued*

Keywords

Arithmetic return (p. 3)
Geometric return (p. 3)
Time resolution (p. 5)
Time horizon (p. 5)
Random walk model (p. 7)
Autoregressive model (p. 8)
AR(1) model (p. 8)
Stationarity (p. 9)
Autocorrelation function (p. 10)
GARCH (1,1) (p. 13)
Marginal distribution (p. 18)
Conditional distribution (p. 18)
qq-plot (p. 19)
Shapiro-Wilk test (p. 19)
Scaled student's t-distribution (p. 20)
Extreme value theory (p. 22)

Learning Objectives

Demonstrate proficiency in the following areas:

3.3.1 Concepts of time series

For example:

A. Define arithmetic and geometric returns.

B. Recognize and apply the relationship between arithmetic and geometric returns.

C. Describe the shape of the plotted line when geometric returns are plotted against arithmetic returns.

D. Define time resolution and time horizon.

E. Describe how time resolution and time horizon affect the distribution of financial data.

3.3.2 Concepts of time series

For example:

A. Describe a random walk model and an autoregressive model.

B. Recognize and apply an AR(1) model.

C. Recognize, calculate and understand applications of the variances of random walk and autoregressive models.

D. Define stationarity and autocorrelation function.

E. Recognize and apply the formula for the autocorrelation function.

F. List the properties of an autocorrelation function for an AR(1) process.

3.3.3 Modeling volatility

For example:

A. Describe a GARCH(1,1) model.

B. List the conditions that must be satisfied by the parameters of a GARCH(1,1) model.

C. Recognize and apply the variance equation of a GARCH(1,1) model.

Learning Objectives *continued*

D. Describe the goodness-of-fit for a GARCH model.
E. Define persistence.
F. Describe marginal and conditional distributions.
G. Describe the marginal distribution for a time-series model with the variance that follows a GARCH process.
H. Describe a qq-plot.
I. Describe how a qq-plot can be used to test for data normality.
J. Compare the reliability of the Shapiro-Wilk test to that of the qq-plot.
K. Describe the scaled Student's t-distribution.
L. Describe the extreme value theory.

Learning Objectives *continued*

Topic 4. Machine Learning: Regularization, Regression Trees, Random Forest & Overfitting

Readings

4.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 5

4.2 James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapters 6.1, 6.2., 8.1 & 8.2

Reading 4.1. Provost, F. and T. Fawcett. (2019). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 5.

Keywords

Generalization (p. 112)
Overfitting (p. 113)
Fitting graph (p. 113)
Holdout data (p. 113)
Training set (p. 114)
Test set (p. 114)
Base rate (p. 115)
Sweet spot (p. 117)
Cross-validation (p. 126)
Folds (p. 127)
Learning curve (p. 131)
Sub-training set (p. 134)
Validation set (p. 134)
Nested holdout testing (p. 134)
Nested cross validation (p. 135)
Sequential forward selection (p. 135)
Sequential backward elimination (p. 135)
Regularization (p. 136)

Learning Objectives

Demonstrate proficiency in the following areas:

4.1 Overfitting and its avoidance

For example:

A. Define generalization, overfitting, fitting graph, holdout data, and base rate.
B. Apply the concept of fitting a graph to find the optimal tree induction model.
C. Define the sweet spot for a typical fitting graph.
D. Apply the concept of overfitting in mathematical functions.
E. Analyze overfitting for logistic regression and support vector machine.
F. Explain why overfitting should be of concern.
G. Define cross-validation and folds.
H. Define a learning curve.
I. Compare and contrast a learning curve with a fitting graph.
J. Describe the shape of learning curves for logistic regression and tree induction.

Learning Objectives *continued*

K. List strategies that can be used to avoid overfitting in tree induction.

L. Describe how the minimum number of instances in a tree leaf can be used to limit tree size.

M. Explain how hypothesis testing can be used to limit tree induction.

N. Define sub-training set, validation set, and nested holdout testing.

O. Explain nested cross-validation.

P. Describe sequential forward selection and sequential backward elimination.

Q. Describe the main idea behind regularization.

Reading 4.2. James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Sections 6.1, 6.2, 8.1, and 8.2.

Keywords

Feature selection, variable selection (p. 204)
Best subset selection (p. 205)
Deviance (p.206)
Forward stepwise selection (p. 207)
Backward stepwise selection (p. 208)
Cp (p. 211)
Akaike information criterion (AIC) (p. 211)
Bayesian information criterion (BIC) (p. 211)
Adjusted R2 (p. 211)
Ridge regression (p. 215)
Tuning parameter (p. 215)
Shrinkage penalty (p. 215)
l2 norm (p. 216)
Scale equivalent (p. 217)
Lasso (p. 219)
Sparse (p. 219)
Decision trees (p. 303)
Regression Tree (p. 304)
Terminal nodes or leaves (p. 305)
Internal node (p. 305)
Branch (p. 305)
Recursive binary splitting (p. 306)
Tree pruning (p. 307)
Cost complexity pruning, weakest link pruning (p. 308)
Classification trees (p. 311)
Classification error rate (p. 311)
Gini index (p. 311)
Cross-entropy (p. 311)
Bagging, bootstrap aggregation (p. 316)
Out-of-bag (OOB) observations (p. 317)
OOB MSE, OOB classification error (p. 318)
Variable importance (p. 319)
Random forests (p. 319)
Boosting (p. 321)
Stump (p. 322)
Interaction depth (p. 322)

Learning Objectives *continued*

Learning Objectives

Demonstrate proficiency in the following areas:

4.2.1 Subset selection (Ch 6.1)

For example:

A. Define the best subset selection.

B. List the steps used in the best subset selection.

C. Define deviance.

D. Describe forward stepwise selection and backward stepwise selection.

E. List the steps used in the forward stepwise selection and backward stepwise selection.

F. Recognize and apply the equations for C_p, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted R^2.

4.2.2 Shrinkage methods (Ch 6.2)

For example:

A. Define ridge regression, tuning parameter, and shrinkage penalty.

B. Define l2 norm and scale equivalent.

C. Define standardizing the predictors.

D. Describe the bias-variance tradeoff.

E. Describe the ridge regression.

F. Describe how ridge regression improves upon least squares.

G. Describe the advantage of Lasso over the ridge regression.

H. Define a sparse model.

I. Describe the variable selection property of the Lasso.

J. Compare the Lasso to the Ridge regression.

K. Describe how to select the tuning parameter.

4.2.3 Tree-Based methods (Ch 8.1-8.2)

For example:

A. Interpret as well as predict using a given decision tree.

B. Describe the advantages and disadvantages of decision trees compared to other classification and regression methods.

C. Describe recursive binary (greedy) splitting for constructing regression trees.

D. Describe tree pruning, specifically cost complexity (weakest link) pruning.

Learning Objectives *continued*

E. Describe the construction of classification trees using classification error rate, Gini index, and cross-entropy.

F. Contrast tree-based methods and linear models.

G. Describe bagging and out-of-bag error estimation.

H. Describe random forests.

I. Describe boosting as an approach for improving the prediction results from decision trees.

Learning Objectives *continued*

Topic 5. Machine Learning: Classification & Clustering

Readings

5.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 6 & 7

Reading 5.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapters 6 & 7.

Keywords

Euclidean distance (p. 144)
Nearest neighbors (p. 144)
Combining function (p. 147)
Weighted voting or scoring (p. 150)
Curse of dimensionality (p.156)
Manhattan distance (p. 158)
Jaccard distance (p.159)
Cosine distance (p. 159)
Edit distance or Levenshtein metric (p. 161)
Clustering (p. 163)
Hierarchical clustering (p. 164)
Dendrogram (p. 164)
Linkage function (p. 166)
Cluster center or centroid (p. 169)
k-means clustering (p. 169)
Distortion (p. 172)
Accuracy (p. 189)
Class prior (p. 201)
F-measure (p. 204)

Learning Objectives

Demonstrate proficiency in the following areas:

5.1.1 Calculating and interpreting similarity and distance (Ch 6)

For example:

A. Calculate the Euclidean distance.

B. Define nearest neighbors and combining function.

C. Explain how combining functions can be used for classification.

D. Define weighted voting (scoring) or similarity moderated voting (scoring).

E. Calculate contributions using weighted voting for classification.

F. Explain how k in k-NN can be used to address overfitting.

G. Discuss issues with nearest-neighbor methods with a focus on

- Intelligibility,
- Dimensionality and domain knowledge and
- Computational efficiency.

H. Define and discuss the curse of dimensionality.

Learning Objectives *continued*

5.1.2 Discussing technical details related to similarities and neighbors (Ch 6)

For example:

A. Calculate the Manhattan distance and the Cosine distance.
B. Define the Jaccard distance.
C. Define edit distance or the Levenshtein metric.
D. Define clustering, hierarchical clustering, and dendrogram.
E. Describe how a dendrogram can help decide the number of clusters.
F. Describe the advantage of hierarchical clustering.
G. Define linkage functions.
H. Describe how distance measures can be used to decide the number of clusters in a dendrogram.
I. Define "cluster center" or centroid and k-means clustering.
J. Compare and contrast k-means clustering with hierarchical clustering.
K. Describe the k-means algorithm.
L. Describe the reason for running the k-means algorithm many times.
M. Define a cluster's distortion.
N. Describe the method for selecting k in the k-means algorithm.

5.1.3 Describing and evaluating classifiers (Ch 7)

For example:

A. Define and calculate the accuracy and error rate.
B. Describe a confusion matrix.
C. Define false positives and false negatives.
D. Identify false positive and false negative within a confusion matrix.
E. Describe unbalanced data and the problems with unbalanced data.
F. Discuss the problems with unequal costs and benefits of errors.

5.1.4 Describing a key analytical framework and calculating expected values (Ch 7)

For example:

A. Calculate the expected value and expected benefit.
B. Describe how the expected value can be used to frame classifier use.
C. Describe how the expected value can be used to frame classifier evaluation.
D. Define class priors.
E. Calculate the expected profit using priors.
F. Describe and evaluate the two pitfalls common to formulating cost-benefit analysis.
G. Define and interpret precision and recall.
H. Calculate the value of the F-measure.

Learning Objectives *continued*

Topic 6. Machine Learning: Performance Evaluation, Support Vector Machines & False Discoveries

Readings

6.1 Provost, F., and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 8

6.2 Arnott, R., C. B. Harvey, and H. Markowitz. (2019). A Backtesting Protocol in the Era of Machine Learning. Journal of Financial Data Science, 1(1), 64-74.

6.3 Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, London, U.K.: Royal Society Open Science, Retrieved from https://royalsocietypublishing.org/doi/full/10.1098/rsos.140216

6.4 López de Prado, M. (2019). A Data Science Solution to the Multiple-Testing Crisis in Financial Research. Journal of Financial Data Science, 1(1), 99-110.

Reading 6.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 8

Keywords

Profit curve (p. 212)
Base rate (p. 214)
ROC graph (p. 215)
Hit rate (p. 216)
False alarm rate (p. 216)
AUC (p. 219)
Lift curve (p. 219)
Cumulative response curve (p. 219)

Learning Objectives

Demonstrate proficiency in the following areas:

6.1.1 Visualizing model performance

For example:

A. Describe a ranking classifier.
B. Define a profit curve.
C. Describe how thresholding can create different confusion matrices.
D. Calculate a confusion matrix using thresholding.
E. Describe the properties of a profit curve.
F. Calculate points of a profit curve.
G. Describe the ROC graph.
H. Calculate points on a ROC graph using data from a confusion matrix.
I. Define the base rate.
J. Describe the four corners and the diagonal of the ROC graph.

Learning Objectives *continued*

K. Define the hit rate and false alarm rate.
L. Describe how to use the ROC space to evaluate classifiers.
M. Define the AUC measure.
N. Describe the cumulative response curve, also known as the lift curve.
O. Calculate points on a cumulative response curve.

Reading 6.2 Arnott, R., C. B. Harvey, and H. Markowitz. (2019). A Backtesting Protocol in the Era of Machine Learning. Journal of Financial Data Science, 1(1), 64-74.

Keywords

Research protocol (p. 65)
Winner's curse (p. 67)
Exaggerated positive (p. 68)

Learning Objectives
Demonstrate proficiency in the following areas:

6.2.1 Backtesting protocol in the era of machine learning

For example:

A. Explain why cross-validation may not reduce the curse of dimensionality.
B. Describe research protocols and their impact on false-positive discoveries.
C. Explain the role of an economic foundation when applying machine learning tools.
D. Describe the winner's curse.
E. Define an exaggerated positive.
F. Describe the seven protocols suggested for avoiding false positives.

Reading 6.3 Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, London, U.K.: Royal Society Open Science.

Keywords

False-positive (p. 2)
Positive predictive power (p.2)
Specificity (p. 2)
Sensitivity (p. 2)
Power (p. 4)
Inflation effect (p. 9)

Learning Objectives *continued*

Learning Objectives
Demonstrate proficiency in the following areas:

6.3.1 An investigation of the false discovery rate and the misinterpretation of p-values

For example:

A. Define specificity and sensitivity.
B. Describe the false discovery rate with the help of a tree diagram.
C. Calculate the probability of real effect given a result is significant.
D. Define the power of a test.
E. Calculate the false discovery rate.
F. Describe an underpowered study.
G. Describe the inflation effect in the context of false discovery.
H. Describe what happens when we consider p=0.05 rather than p<=0.05.
I. Describe Berger's approach.
J. Calculate the false discovery rate using conditional probabilities.
K. Calculate the conditional probability of the real effect.
L. Calculate the odds ratio using the Bayes approach.

Reading 6.4 López de Prado, M. (2019). A Data Science Solution to the Multiple-Testing Crisis in Financial Research. Journal of Financial Data Science, 1(1), 99-110.

Keywords
Selection bias under multiple testing (SBuMT) (p. 99)
Clustering of trials (p. 102)

Learning Objectives
Demonstrate proficiency in the following areas:

6.4.1 A data science solution to the multiple-testing crisis

For example:

A. Define selection bias under multiple testing.
B. Describe the three properties that must be satisfied by trials to reduce SBuMT.
C. Describe the clustering of trials.
D. Describe the implications of using an optimal number of clusters.
E. Describe how clustering of strategies can reduce SBuMT.
F. Describe the implications for authors, journals, and financial firms.

Learning Objectives *continued*

Topic 7. Data Mining & Machine Learning: Naïve Bayes & Text Mining

7.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapters 9 & 10.

7.2 Jurafsky, D. and J. Martin. (2018). Chapter 4. Naïve Bayes and Sentiment Classification, In Speech and Language Processing. Retrieved from https://web.stanford.edu/~jurafsky/slp3/4.pdf

Reading 7.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 9 & 10.

Keywords

Independent events (p. 236)
Probability (p. 237)
Bayes' Rule (p. 237)
Prior (p. 238)
Posterior probability (p. 238)
Likelihood (p. 240)
Conditional independence (p. 241)
Joint probability using conditional Naïve Bayes Classifier (p. 242)
Lift (p. 244)
Naïve-Naïve Bayes (p. 244)
Linguistic structure (p. 250)
Dirty data (p. 250)
Document (p. 251)
Corpus (p. 251)
Token (p. 251)
Terms (p. 251)
Term frequency (p. 252)
Bag of words (p. 252)
Stemmed (p. 253)
Stopwords (p. 253)
Inverse document frequency (p. 254)
N-grams (p. 263)
Latent information model (p. 266)
Information triage (p. 274)

Learning Objectives

Demonstrate proficiency in the following areas:

7.1.1 Evidence and probabilities (Ch 9)

For example:

A. Define independent events.
B. Calculate the joint probability of two events.
C. Recognize and apply joint probability using conditional probability.
D. Calculate joint probability for independent and dependent events.
E. Explain Bayes' Rule with the help of an example.
F. Define posterior probability, prior, likelihood, and conditional independence.
G. Explain the naïve Bayes classifier.

Learning Objectives *continued*

H. Explain why we do not need to calculate the denominator of Bayes' rule for the naïve Bayes classifier.

I. List the advantages and disadvantages of the naïve Bayes classifier.

J. Define generative model, lift, and Naïve-Naïve Bayes.

7.1.2 Broad issues involved in mining text (Ch 10)

For example:

A. Explain why text is "dirty," which makes mining text difficult.

7.1.3 Text representation (Ch 10)

For example:

A. Understand the meaning of "terms" when used in the field of information retrieval.

B. Describe the "bag of words" approach, including the following steps:
- Measuring term frequency (TF)
- Measuring sparseness: inverse document frequency (IDF)
- Combining them: TFIDF

C. Apply appropriate methods to create a TFIDF representation of a query.

D. Express entropy, in terms of the IDF measure.

7.1.4 Additional text representation approaches beyond "bag of words." (Ch 10)

For example:

A. Describe N-gram sequences.

B. Describe named entity extraction.

C. Describe topic models

7.1.5 Mining news stories to predict stock price movement (Ch 10)

For example:

A. Describe how a given task, such as recommending a news story that is likely to result in a significant change in a stock's price, must be formulated into a problem with simplifying assumptions.

B. Describe the required considerations for data preprocessing.

C. Identify and discuss appropriate methods for analyzing the results.

Reading 7.2 Jurafsky, D., and J. Martin. (2018). Chapter 4. Naïve Bayes and Sentiment Classification, **In Speech and Language Processing.**

Learning Objectives *continued*

Keywords

Sentiment analysis (p. 1)
Probabilistic classifier (p. 2)
Generative classifier (p. 2)
Discriminative classifier (p. 2)
Linear classifier (p. 5)
Sentiment lexicon (p. 9)
Gold labels (p.11)
Precision (p.12)
Recall (p.12)
F-measure (p.13)
Macroaveraging (p.13)
Microaveraging (p.13)

Learning Objectives

Demonstrate proficiency in the following areas:

7.2.1 Classification

For example:

A. Describe typical applications of classifying text.

B. Describe tasks often involved in classifying text.

C. Compare alternative methods of classification.

7.2.2 Math behind Naïve Bayes classifiers

For example:

A. Explain, in the context of classifying a document, why the denominator can be dropped from Bayes Rule.

B. Explain the bag of words and the assumptions of the naïve Bayes classifier.

C. Explain why Naïve Bayes calculations are done in log space so that the predicted class is a linear function of input features.

7.2.3 Training the Naïve Bayes classifiers

For example:

A. Explain why Bayes text categorization often uses Laplace smoothing.

B. Explain how to treat stopwords and unknown words during training.

C. Calculate the prior probabilities of two classes, given a training set categorized into two categories.

D. Determine the class that a test sentence belongs to using the Naïve Bayes classifier.

7.2.4 Optimizing for sentiment analysis

For example:

A. Explain how binary, multinomial Naïve Bayes differs from Naïve Bayes.

Learning Objectives *continued*

B. Explain why binary, multinomial Naïve Bayes (also called binary NB) might improve results relative to the standard Naïve Bayes approach.

C. Describe two other methods (besides binary NB) that can improve the results of sentiment analysis.

7.2.5 Evaluation of sentiment analysis results

For example:

A. Calculate precision and recall statistics given system output and gold standard label results.

B. Describe the F-measure and various methods of weighting precision and recall.

C. Compare macroaveraging and microaveraging approaches to evaluating the categorization performance of multiple classes.

D. Compare 10-fold cross-validation with bootstrap tests.

Learning Objectives *continued*

Topic 8. Big Data & Machine Learning: Ethical & Privacy Issues

Readings

8.1 Institute of Business Ethics. (2016, June). Business Ethics and Big Data (IBE Issue 52). London, U.K.

8.2 Institute of Business Ethics. (2018, January). Business Ethics and Artificial Intelligence (IBE, Issue 58). London, U.K.

8.3 Institute of Business Ethics. (2018, May). Beyond Law: Ethical Culture and GDPR (IBE, Issue 62). London, U.K.

8.4 Loukides, M., M., H. Mason and DJ. Patil. Ethics and Data Science **Free e-book** https://www.amazon.com/Ethics-Data-Science-Mike-Loukides-ebook/dp/B07GTC8ZN7

Note: Registered candidates can find these readings on the FDP Institute website at https://fdpinstitute.org/Topics-in-Financial-Data-Science

Reading 8.1 Institute of Business Ethics. (2016, June). Business Ethics and Big Data (IBE Issue 52). London, U.K

Keywords

Data trust deficit (p. 2) Veracity (p. 6)

Learning Objectives

Demonstrate proficiency in the following areas:

8.1.1 Big data for business

For example:

A. Discuss the potential and concerns of big data for business.

B. Explain how the new term "data trust deficit" developed.

8.1.2 Ethical issues

For example:

A. List five methods of protecting human rights in the 'Era of Big Data.'

B. Provide an example of concern for each of the three main areas of privacy issues: customer profiling, group privacy, and data security.

C. Discuss what constitutes informed consent.

D. Provide an example of how to improve the veracity of data.

Learning Objectives *continued*

8.1.3 The ethics test

For example:

A. List six questions that ethics professionals within an organization using big data can ask themselves.

Reading 8.2 Institute of Business Ethics. (2018, January). Business Ethics and Artificial Intelligence (IBE Issue 58). London, U.K.

Keywords

Artificial intelligence (p. 1) *Code of ethics (p. 6)*

Learning Objectives

Demonstrate proficiency in the following areas:

8.2.1 The nature of and business risks of artificial intelligence (AI)

For example:

A. List three main features characterizing artificial intelligence.

B. List three immediate risks of artificial intelligence.

8.2.2 Values that form the cornerstone of an ethical framework for artificial intelligence in business

For example:

A. Discuss each of the following as they impact the ethical nature of applications of artificial intelligence in business:
- Accurate results
- Respect for privacy
- Transparency and openness
- Interpretability of algorithms
- Fairness to stakeholders
- Integrity and due diligence
- Control of humans relative to machines
- Impact of a new technology
- Accountability assignment
- Learning about how AI technologies work

8.2.3 The role of business decision-makers

For example:

A. List five measures organizations can take to minimize the risk of ethical lapses due to improper use of AI technologies.

B. List potential questions addressing the use of AI in a code of ethics.

Learning Objectives *continued*

Reading 8.3 Institute of Business Ethics. (2018, May). Beyond Law: Ethical Culture and GDPR (IBE, Issue 62). London, U.K.

Keywords
General Data Protection Regulation (p. 1) *People risk (p. 3)*

Learning Objectives
Demonstrate proficiency in the following areas:

8.3.1 General Data Protection Regulation (GDPR)

For example:

A. Describe the primary purpose of the GDPR.

B. Describe the key changes in data protection regulation including the meaning of
- Rights of the individual
- Informed consent
- Notification
- Data portability
- Supervision and enforcement, and
- Liability.

8.3.2 Separating ethics and compliance

For example:

A. Distinguish between two types of threats of personal data breaches.

B. Discuss 'people risk.'

C. List key questions around the role an ethical culture plays in preventing data breaches.

8.3.3 Maintaining privacy of personal data

For example:

A. Describe how an organization must build awareness regarding employees' roles in protecting data.

B. Discuss liability for missing the 72-hour notification deadline.

8.3.4 The GDPR Embedding Wheel

For example:

A. Describe how the tone from the top can help foster an ethical culture and compliance with the GDPR.

B. Describe how establishing the boundaries and standards can help foster an ethical culture and compliance with the GDPR.

Learning Objectives *continued*

C. Describe how communication and training can help foster an ethical culture and compliance with the GDPR.

D. Describe how the choice of the individual can help foster or hinder an ethical culture and compliance with the GDPR.

E. Describe how monitoring outcomes can help foster an ethical culture and compliance

Reading 8.4 Loukides, M., M., H. Mason and DJ. Patil. Ethics and Data Science Free e-book
https://www.amazon.com/Ethics-Data-Science-Mike-Loukides-ebook/dp/B07GTC8ZN7

Keywords

Kanban (p. 12)
Minimal viable product (MVP) (p. 31)
Social impact statement (p. 36)
Andon cord (p. 39)
Digital privacy law (p. 43)

Note: These page numbers refer to the pages indicated in a PDF reader when the webpage is printed to a PDF file.

Learning Objectives
Demonstrate proficiency in the following areas:

8.4.1 Doing good data science

For example:

A. Identify aspects of putting ethical principles into practice.

8.4.2 Oaths and checklists

For example:

A. Identify appropriate tools for implementing sound ethical practices.

8.4.3 The five C's

For example:

A. Recognize the main point behind each of the five "C's"
- Consent
- Clarity
- Consistency
- Control and transparency, and
- Consequences and harm.

B. Recognize when an organization does and does not follow one of the "C's" framing guidelines.

C. Explain how to implement the five C's.

Learning Objectives *continued*

8.4.4 Taking responsibility for our creations (Data's Day of Reckoning)

For example:

A. Identify major issues in ethics and security training.

B. Argue for a method of developing guiding principles.

C. Describe how to build ethics into a data-driven culture.
- Identify four methods for doing so.
- Describe the ideal role of teams and corporations.

D. Discuss the regulatory environment for data and new technologies in terms of consent.

Learning Objectives *continued*

Topic 9. Big Data & Machine Learning in the Financial Industry

Readings

9.1 Financial Stability Board. (2017). Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications. Retrieved from http://www.fsb.org/wp-content/uploads/P011117.pdf

9.2 Monk, A., M. Prins, and D. Rook. (2019). Rethinking Alternative Data in Institutional Investment. Journal of Financial Data Science, 1(1), 14-31. DOI: https://doi.org/10.3905/jfds.2019.1.1.014

9.3 Simonian, J., C. Wu, D. Itano and V. Narayanan. (2019). A Machine Learning Approach to Risk Factors: A Case Study Using the Fama–French–Carhart Model. Journal of Financial Data Science, 1(1), 32-44. DOI: https://doi.org/10.3905/jfds.2019.1.032

9.4 Rasekhschaffe, K. and R. Jones. (2019). Machine Learning for Stock Selection. Financial Analyst Journal, 13 May 2019 Volume 75 Issue 3. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3330946

9.5 Gu, S., B. Kelly, and D. Xiu. (2018). Empirical Asset Pricing via Machine Learning. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281018
https://www.bryankellyacademic.org/
http://dachxiu.chicagobooth.edu/download/ML.pdf
The paper posted on the FDP Institute's website will be the version used for exam questions.

9.6 López de Prado, M. (2018). The 10 Reasons Most Machine Learning Funds Fail. The Journal of Portfolio Management, 44 (6) 120-133; DOI: https://doi.org/10.3905/jpm.2018.44.6.120

9.7 Harvey, C. R. and Y. Liu. (2014). Evaluating Trading Strategies. *[Special 40th Anniversary Issue]*. The Journal of Portfolio Management, 40(5), 108-118. DOI: https://doi.org/10.3905/jpm.2014.40.5.108

9.8 Raman, J., and R. Lam (2019). Artificial Intelligence Applications in Financial Services https://www.oliverwyman.com/our-expertise/insights/2019/dec/artificial-intelligence-applications-in-financial-services.html PDF: https://www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2019/dec/ai-app-in-fs.pdf

9.9 Zappa, D., M. Borrelli, G.P. Clemente, N. Savelli. Text Mining In Insurance: From Unstructured Data To Meaning https://www.variancejournal.org/articlespress/articles/Text_Mining-Zappa-Borrelli-Clemente-Savelli.pdf

Learning Objectives *continued*

Reading 9.1 Financial Stability Board. (2017). Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications.

Keywords

Big data (p. 4)
Artificial intelligence (p. 4)
Machine learning (ML) (p. 4)
Natural language processing (p. 5)
Supervised learning (p. 5)
Unsupervised learning (p. 5)
Deep learning (p. 5)
Reinforcement learning (p. 5)
Sentiment indicators (p. 10)
Fraud detection (p. 11)
RegTech (p. 11)
Trading signals (p. 11)
InsurTech (p. 13)
Chatbots (p. 14)
Know your customer (KYC) (p. 20)
SupTech (p. 21)
Auditability (p. 33)
Fintech (p. 35)
Robo-advisors (p.35)
Tonality analysis (p.36)

Learning Objectives

Demonstrate proficiency in the following areas:

9.1.1 Regulatory and supervisory issues around FinTech

For example:

A. Identify factors that may contribute to increases in third party dependencies among financial institutions.

B. Explain why unexpected forms of interconnectedness among institutions could be created.

C. Explain why new forms of macro-level risks could emerge.

D. Explain why new risk management tools and techniques may be required.

9.1.2 Relationship between artificial intelligence, machine learning, and big data, and algorithms

For example:

A. Describe the two recent developments that have contributed to increased interest in AI.

B. List factors contributing to making the markets more efficient.

C. Describe the relationship between AI, machine learning, and the three algorithms appearing in Figure 1.

Learning Objectives *continued*

9.1.3 Categories of machine learning algorithms

For example:

A. Define four categories of machine learning algorithms based on the degree of human intervention.

B. Describe the role of machine learning algorithms in determining causality vs. correlation.

C. Define 'augmented intelligence.'

D. Explain the limitations of machine learning algorithms in determining causality and correlations.

9.1.4 Drivers of the growth in the use of fintech and adoption of artificial intelligence

For example:

A. Discuss the supply factors related to advances in computing technologies and changes in the financial sector.

B. Discuss the demand factors related to the search for higher profits, increased competition, and changes in the regulatory environment.

9.1.5Use cases of artificial intelligence and machine learning in the financial sector

For example:

A. Describe customer-focused uses, such as credit scoring, insurance, and client-facing chatbots.

B. Describe operations-focused uses, such as optimal allocation of capital, risk management modeling, and market impact analysis.

C. Describe portfolio management and trading uses.

D. Describe regulatory compliance and supervision uses by financial institutions, central banks, macroprudential authorities, and market regulators.

9.1.6 The micro-financial analysis of artificial intelligence and machine learning uses

For example:

A. Describe the uses of artificial intelligence and machine learning in information gathering and processing their potential impacts on financial markets.

B. Describe the uses of artificial intelligence and machine learning in improving the efficiency of financial institutions.

C. Describe the uses of artificial intelligence and machine learning by financial institutions and their potential impacts on customers and investors.

Learning Objectives *continued*

9.1.7 The macro-financial analysis of uses of artificial intelligence and machine learning uses

For example:

A. Describe economic growth and enhanced economic efficiency that could result from the applications of artificial intelligence and machine learning to financial services.

B. Describe the implications of uses of artificial intelligence and machine learning by financial institutions for market concentration and systemic importance of those institutions.

C. Describe how the uses of artificial intelligence and machine learning by financial institutions could be sources of greater instability and vulnerability in financial markets.

D. Describe how the uses of artificial intelligence and machine learning by the insurance industry could affect both moral hazard and adverse selection problems.

E. Describe challenges posed by the lack of interpretability or auditability in applications of artificial intelligence and machine learning in the financial industry.

9.1.8 Define the terms listed in the glossary

For example:

A. Describe the following terms: Algorithm, Artificial intelligence, Augmented intelligence, Big data, Chatbots, Cluster analysis, Deep learning, FinTech, InsurTech, Internet of things, Machine learning, Natural Language Processing, RegTech, Reinforcement learning, Robo-advisors, Social trading, SupTech, Supervised learning, Tonality analysis, Topic modeling, and Unsupervised learning.

Reading 9.2 Monk, A., M. Prins, and D. Rook. (2019). Rethinking Alternative Data in Institutional Investment. Journal of Financial Data Science, 1(1), 14-31.

Keywords

Alternative data (p. 14)
Social media (p. 14)
Microdata (p. 14)
Data exhaust (p. 14)
Rivalry (p. 16)
Excludability (p. 16)
Defensive strategies (p. 17)
Defensible strategies (p. 18)
Operational alpha (p. 19)
Aggregation (p. 19)
Disaggregation (p. 19)
Volume (p. 21)
Velocity (p. 21)
Variety (p. 21)
Veracity (p. 21)
Granularity (p. 21)
Relationality (p. 21)
Flexibility (p. 21)
Actionability (p. 22)
Excludable (p. 28)
Data hoarding (p. 29)

Learning Objectives *continued*

Learning Objectives
Demonstrate proficiency in the following areas:

9.2.1 Alternative data and institutional investors

For example:

A. Define alternative data and list examples of alternative data.

B. List the most commonly used types of alternative data.

C. Explain why the alternative data's core value proposition is different for institutional investors.

D. Discuss the advantages and disadvantages that institutional investors may have in using alternative data.

E. Discuss why the most in-depth value proposition alternative data has for institutional investors entails defensive and defensible strategies.

F. List examples of how alternative data may be used defensively for understanding risk.

G. Discuss the applications of alternative data to risk measurement and management for institutional investors.

H. Describe the operational alpha gains by institutional investors through the use of alternative data sets.

I. Describe types of alternative data sets in terms of the origins of a data set.

J. Discuss why the volume, veracity, and velocity of big data may not determine the value of alternative data for institutional investors.

K. Describe the six-dimensional characterization of alternative data.

L. Discuss external asset managers and alternative data providers as methods of accessing alternative data.

M. Discuss the consequences of the increased use of alternative data on risk for institutional investors.

N. Compare accessing alternative data through external asset managers versus alternative data vendors.

O. Describe rivalry and excludability (limited or permanent) as determinants of alternative data set's value.

Reading 9.3 Simonian, J., C. Wu, D. Itano and V. Narayanan. (2019). A Machine Learning Approach to Risk Factors: A Case Study Using the Fama–French–Carhart Model. Journal of Financial Data Science, 1(1), 32-44.

Learning Objectives *continued*

Keywords

Factors (p. 32)
Linear (p. 32)
Nonlinear (p. 32)
Random forest (p. 33)
CART (p. 34)
Binary recursive partitioning (p. 34)
Bagging (p. 34)
Out-of-bag data (p. 34)
Decision node (p. 34)
Supervised (p. 34)
Unsupervised (p. 34)
Terminal node (p. 34)
Mean decrease accuracy (p. 34)
Root node (p. 34)
Feature importance (p. 35)
Fama-French-Carhart (p. 37)
Probabilistic Sharpe ratio (p. 42)

Learning Objectives

Demonstrate proficiency in the following areas:

9.3.1 Applications of random forest regression algorithm to factor models

For example:

A. Discuss two shortcomings of parametric nonlinear factor models that are developed to address the shortcomings of linear models.

B. Discuss the ability of the random forest algorithm to overcome one shortcoming of linear models.

C. Discuss the ability of the random forest algorithm to overcome one shortcoming of parametric nonlinear models.

D. List four components of the decision tree when applied to the regression problem of factor models.

E. Describe how bagging is employed in an ensemble of decision trees (random forest).

F. Calculate the predicted value of a dependent (response) variable given a set of predictor values and the outputs of a binary regression decision tree algorithm.

G. Describe the role of out-of-bag observations in a random forest algorithm.

H. Discuss the mean decrease accuracy approach to estimating feature importance in a random forest algorithm.

I. Recognize and apply the probabilistic Sharpe ratio. (The formula appearing in the text is incorrect). The correct formula appears below:

$$PSR\left[SR^*\right] = Z\left[\frac{\left(\hat{SR} - SR^*\right)\sqrt{T-1}}{\sqrt{1-\hat{\gamma}_3\hat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\hat{SR}^2}}\right]$$

Learning Objectives *continued*

Reading 9.4 Rasekhschaffe, K., and R. Jones. Machine Learning for Stock Selection. Financial Analyst Journal, 13 May 2019 Volume 75 Issue 3 (pp. 1-14, 24).

Keywords

Machine learning (p. 3)
Signal-to-noise ratio (p. 4)
Overfitting (p. 5)
Ensemble algorithms (p. 6)
Feature engineering (p. 8)
Forecast horizon (p. 10)
Bagging (p. 11)
Boosting (p. 11)
Technical factors (p. 11)
Fundamental factors (p. 13)

Learning Objectives

Demonstrate proficiency in the following areas:

9.4.1 The applications of machine learning algorithms to stock selection

For example:

A. Describe the role of signal to noise ratio in creating overfitted models.

B. Discuss the implications of the paper's findings concerning in-sample versus out-of-sample errors as the number of boosting iterations increases.

C. Describe the four different approaches to bagging and boosting employed by the paper to avoid overfitting.

D. Explain the importance of feature engineering and the role of subject matter expertise in mitigating the overfitting problem.

E. Describe the three decisions that must be made about the forecasting goals of the machine learning algorithms.

F. Describe the bias versus variance tradeoff.

G. Explain the role of bagging and boosting in affecting the bias versus variance tradeoff.

H. Understand the terms appearing in the glossary: activation function, artificial neural networks, and deep learning, bagging, boosting, dropout, gradient boosted trees, random forest.

Reading 9.5 Gu, S., B. Kelly, and D. Xiu. (2018). Empirical Asset Pricing via Machine Learning. (pp. 1-24)

Learning Objectives *continued*

Keywords

Machine learning (p. 1)
Regularization (p. 2)
Mean squared error (p. 9)
Ordinary least squares (p. 9)
Penalized linear models (p. 11)
Loss function (p. 11)
Penalty function (p. 11)
Net elastic (p. 11)
Heavy tails (p. 11)
Huber loss function (p. 12)
Hyperparameters (p. 12)
Tuning parameters (p. 12)
LASSO (p. 13)
Principal components (p. 13)
Partial least squares (p. 13)
Approximation error (p. 15)
Estimation error (p. 15)
Intrinsic error (p. 15)
Terminal nodes (p. 16)
Impurity (p. 16)
Weak learners (p. 17)
Boosting (p. 18)
Random forest (p. 18)
Gradient boosted regression trees (p. 18)
Neural network (p. 19)
Hidden layers (p. 19)
Feed-forward networks (p. 19)
Input layer (p. 20)
Output layer (p. 20)
ReLU function (p. 21)
Stochastic gradient descent (p. 21)
Early stopping (p. 22)

Learning Objectives

Demonstrate proficiency in the following areas:

9.5.1 Applications of machine learning algorithms to empirical asset pricing

For example:

A. Describe the three components of the definition of machine learning.

B. Describe the three aspects of empirical asset pricing models that make them attractive for the applications of machine learning algorithms.

C. Compare and contrast the overall performance of linear versus nonlinear models in predicting individual stock returns and portfolio returns.

D. Explain one potential shortcoming of machine learning algorithms when used to predict asset returns.

E. Describe the roles of “training” set, “validation” set, and “testing” set in using machine learning algorithms to predict stock returns.

F. Recognize the Huber loss function.

G. Describe the benefit of using the Huber loss function as opposed to the standard least-squares method to the estimation of linear models.

H. Recognize the “elastic net” approach for modeling penalized linear models.

I. Compare and contrast the “elastic net” penalty versus LASSO and Ridge Regression.

J. Compare and contrast the principle components regression versus partial least squares.

Learning Objectives *continued*

K. Recognize and describe the three sources of a model's forecast errors (decomposition of forecast errors).

L. Describe the boosting regularization method in the context of regression trees.

M. Describe the random forest regularization method in the context of regression trees.

N. Describe the dropout method in the context of random forest regression trees.

O. Recognize rectified linear unit (ReLU) activation function in the context of neural networks.

Reading 9.6 López de Prado, M. (2018). The 10 Reasons Most Machine Learning Funds Fail. The Journal of Portfolio Management, 44 (6) 120-133.

Keywords

Backtesting (p. 122)
Volume clock (p. 123)
Dollar bars (p. 123)
Stationary (p. 123)
Integer differentiation (p. 123)
Fractional differentiation (p. 124)
Triple barrier method (p. 127)
Precision (p. 128)
Recall (p. 128)
F1-score (p. 128)
Walk-forward approach (p. 129)
Leakage (p. 129)
Deflated Sharpe ratio (p. 132)
Probabilistic Sharpe ratio (p. 132)

Learning Objectives

Demonstrate proficiency in the following areas:

9.6.1 The most common errors made when machine learning techniques are applied to financial data sets

For example:

A. Compare and contrast the silo approach in discretionary strategies versus the meta-strategy in machine learning strategies.

B. Compare and contrast repeated backtesting using machine learning versus examining feature importance of the results from a machine learning application.

C. Describe the two problems with data samples generated using time bars.

D. Describe the advantages of dollar bars over time bars in creating data for machine learning algorithms.

E. Describe the benefit of using fractional differentiation in generating stationary series while preserving memory.

F. Explain the triple-barrier method for labeling observed returns.

G. Describe the definitions of precision, recall, and F1-score as features of machine learning algorithms.

Learning Objectives *continued*

H. Explain the role of non-independent identically distributed returns in the failure of k-fold cross-validation in finance.

I. Describe walk forward (WF) approach to backtesting of trading strategies.

J. Describe the advantages and disadvantages of walk forward approach.

K. Explain the relationship between the maximum Sharpe ratio obtained from several backtested strategies and the return volatility of those strategies.

L. Describe the concept of probabilistic Sharpe ratio.

M. List the impacts of nonnormalized Sharpe ratio, length of track record, skewness, and kurtosis on the probabilistic Sharpe ratio.

Reading 9.7 Harvey, C. R. and Y. Liu. (2014). Evaluating Trading Strategies. [Special 40th Anniversary Issue]. The Journal of Portfolio Management, 40(5), 108-118.

Keywords

T-statistics (p. 110)
Family-wise error rate (p. 111)
False discovery rate (p. 111)
Holm test (p. 112)
BHY hurdle (p. 112)
Bonferroni test (p. 112)
Type I error (p. 113)
Type II error (p. 113)

Learning Objectives

Demonstrate proficiency in the following areas:

9.7.1 Using statistical techniques to evaluate trading strategies in the presence of multiple tests.

For example:

A. Describe why standard statistical tools, such as p-values and t-statistics, can lead to false discoveries in the presence of multiple tests.

B. Calculate the t-statistic based on the reported Sharpe ratio for testing a single trading strategy.

C. Describe and apply Bonferroni tests in the context of the family-wise error rate (FWER) approach to adjusting p-values for multiple tests.

D. Describe the Holm method in the context of the false discovery rate (FDR) approach to adjusting p-values for multiple tests.

E. Recognize and apply the Holm function to calculate adjusted p-values.

F. Understand the process of accepting and rejecting tests using the Holm method.

G. Describe the false discovery approach to adjusting p-values in the presence of multiple tests.

Learning Objectives *continued*

H. Recognize and apply the BHY formula to calculate adjusted p-values.

I. Understand the process of accepting and rejecting tests using the BHY method.

J. Explain the relationship between avoiding false discoveries and missing on profitable opportunities.

Reading 9.8 Raman, J., and R. Lam (2019). Artificial Intelligence Applications in Financial Services https://www.oliverwyman.com/our-expertise/insights/2019/dec/artificial-intelligence-applications-in-financial-services.html **PDF:** https://www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2019/dec/ai-app-in-fs.pdf

Keywords

Digitisation (Digitization) (p. 6)
Big data (p. 6)
Automation (p. 6)
Chatbots (p. 17)
Underwriting (p. 19)

Learning Objectives

Demonstrate proficiency in the following areas:

9.8.1 Artificial intelligence (AI) and technology trends.

For example:

A. Describe the pitfalls that an organization might face if prudence is not exercised as it employs AI.

B. Explain why boards of organizations employing AI must have sufficient understanding of AI

C. Describe the relationship between digitization, big data, and AI.

D. Describe examples of “known knowns”, “known unknowns” and “unknown unknowns” in the context of big data.

E. Describe two key aspects of cybersecurity in the context of AI.

9.8.2 Applications of Artificial intelligence (AI) in financial services.

For example:

A. Explain the key benefit of AI in risk management with the asset management industry.

B. Explain the key benefit of AI in alpha generation using big data.

C. Describe the relevance of AI for long-term and short-term investors.

D. Describe applications of AI in various areas of banking, such as customer service, underwriting, cross-selling of products, and fraud detection.

E. Describe the risks insurance companies could face when using AI.

Learning Objectives *continued*

Reading 9.9 Zappa, D., M. Borrelli, G.P. Clemente, N. Savelli. Text Mining In Insurance: From Unstructured Data To Meaning https://www.variancejournal.org/articlespress/articles/Text_Mining-Zappa-Borrelli-Clemente-Savelli.pdf

Keywords

Text mining (p. 1)
N-Grams (p. 3)
Bag-of-words (p. 5)
Tokenization (p. 7)
Stemming (p. 7)
Lemmatization (p. 7)
Document term matrix (p. 8)
Term document matrix (p. 8)
Term frequency (p. 8)
Inverse document frequency (p. 8)
Natural language processing (p. 12)
Continuous bag of words (p. 13)
Part-of-speech tagging (p. 17)

Learning Objectives

Demonstrate proficiency in the following areas:

9.9.1 Text mining and its applications in the insurance industry.

For example:

A. Describe an example of a competitive advantage that an insurance company can gain through text mining.
B. Describe and apply N-gram to analyze a document.
C. Describe and apply the chain rule to calculate the joint probability of words in a text.
D. Describe and apply the chain rule of probability calculation when applying an N-gram language model.
E. Describe the tokenization process when applied to a document.
F. Describe the use of stop-words in reducing the number of features of a document.
G. Describe the stemming pre-processing approach when applied to a document.
H. Describe the lemmatization pre-processing approach when applied to a document.
I. Describe, interpret, and apply term frequency and inverse term frequency algorithms.
J. Explain the objective of the simplest version of a continuous bag of words algorithm.
K. Describe the part-of-speech tagging pre-processing approach when applied to a document.

Action Words

In each of the above learning objectives, action words are used to direct your study focus. Below is a list of all action words used in this study guide, along with definitions and two examples of usage, in a question example and a description. Should you not understand what is required for any learning objective, we suggest you refer to the table below for clarification.
NOTE: The question examples in this table are NOT sample questions for the current exam.

Term	Definition	Question Example	Example of Term Use
Analyze	Study the interrelations	George has identified an opportunity for a convertible arbitrage reverse hedge. What risks are associated with this hedge? A. The convertible may remain overvalued, causing the positive cash flow to harm the position's return profile. B. The short convertible may be called in and the position must be delivered, forcing the hedge to be unwound at an inopportune time. C. The implied volatility may decrease, lowering the bond's value.	You have to **analyze** the positions and factors impacting them. ***Correct Answer: B***
Apply	Make use of **Note:** If you are asked to apply a model to data, you will be expected to have the appropriate equation memorized, unless the question also contains the action word "recognize"..	Alicia Weeks, CFA, Real Estate Investment Advisor, works in an Asian country where there are no securities laws or regulations. According to CFA Institute Standard I, Fundamental Responsibilities, Alicia: A. Must adhere to the standards as defined in a neighboring country that has the strictest laws and regulations. B. Need not concern herself with ethics codes and standards. C. Must adhere to the CFA Institute's codes and standards.	You have to **apply** CFA Institute Standard I to find the correct answer. ***Correct Answer: C***
Argue	Prove by reason or by presenting the associated pros and cons; debate	Why did the shape of the supply curve for venture capital funds change after 1979?	You have to describe how the curve has changed AND **argue** why it changed by providing reasons and supporting the reasons with statements of facts (e.g., change in regulations).
Assess	Determine importance, size, or value	How are lower capital gains taxes expected to impact firm commitments? A. Through increased supply of capital, firm commitments are expected to rise. B. Through decreased supply of capital, firm commitments are expected to rise. C. Through decreased after-tax return on venture investments, firm commitments are expected to rise.	You must **assess** the significance of the change in the tax rate for firm commitments. ***Correct Answer: A***

Action Words *continued*

Term	Definition	Question Example	Example of Term Use
Calculate	Determine a value mathematically **Note:** You will be expected to have the appropriate equation memorized, unless the question also contains the action word "recognize".	Consider a set of 100 people. Eighty percent have feature A and twenty percent do not have feature A. What is the entropy for this set? A. 0.72 B. 0.88 C. 0.93	You have to **calculate** entropy based on the given probabilities. ***Correct Answer: B***
Compare	Describe similarities and differences	Which of the following least accurately compares the Sharpe and Treynor ratios? A. Both ratios contain excess return in the numerator. B. Both ratios express a measure of return per unit of some measure of risk. C. The Sharpe ratio is the inverse of the Treynor ratio	You must **compare** the ratios based on their most important similarities and their most important differences. ***Correct Answer: C***
Compare and Contrast	Examine in order to note similarities or differences	A comparison of monthly payments and loan balances of a constant payment mortgage with a constant amortization mortgage with the same loan terms will show that: A. The initial payment will be the same. B. The payments of the constant payment mortgage are initially greater than those of the constant amortization mortgage, but at some point the payments of the constant payment mortgage become less. C. The present value of the payment streams of the two loan types are the same.	You must **compare** indices to arrive at the answer. ***Correct Answer: C***
Construct	Make or form by combining or arranging parts or elements	A reverse convertible arbitrage hedge consists of a: A. Short convertible position plus a long position in the stock. B. Short convertible position plus a put option on the stock. C. Long convertible position plus a put option on the stock.	You must **combine** positions to construct the hedge. ***Correct Answer: A***
Contrast	Expound on the differences	Which of the following best characterizes a difference between value at risk (VaR) and modified VaR? A. Modified VaR is expressed as a percent while VaR is a dollar value. B. Modified VaR uses a user defined confidence interval while VaR uses a 99% interval. C. Modified VaR incorporates non-normality while traditional VaR assumes normality	You have to **contrast** the assumptions of the first model to those of the second model so that the differences are clear. ***Correct Answer: C***

Action Words *continued*

Term	Definition	Question Example	Example of Term Use
Define	State the precise meaning	The interest rate charged by banks with excess reserves at a Federal Reserve Bank to banks needing overnight loans to meet reserve requirements is called the: A. Prime rate. B. Discount rate. C. Federal funds rate.	You must **define**, in this case, the federal funds rate. ***Correct Answer: C***
Describe	Convey or characterize an idea	Which of the following words best describes expected return? A. Spread B. Average C. Spread squared	You need to choose the word that best **describes** the concept from a list. **Correct Answer: B**
Differentiate	Constitute the distinction between; distinguish	What type of convertible hedge entails shorting a convertible and going long in the underlying stock? A. Reverse hedge B. Call-option hedge C. Traditional convergence hedge	You must **differentiate** one type of hedge from another. ***Correct Answer: A***
Discuss	Examine or consider a subject	Discuss the limitations of private equity data.	You must present a **discussion** of a set of ideas in a list or paragraph.
Explain	Illustrate the meaning	1. Explain why return on assets (ROA) rather than return on equity (ROE) might be the preferred measure of performance in the case of hedge funds. or 2. Which of the following best explains risk from the standpoint of investment? A. Investors will lose money. B. Terminal wealth will be less than initial wealth. C. More than one outcome is possible.	1. You must place a series of thoughts together as an **explanation** of a term or issue. 2. You need to identify the term that best **explains** a term or issue. ***Correct Answer: C***
Identify	Establish the identity	The investments that have historically performed best during periods of recession are: A. Commodities. B. Treasury bills. C. Stocks and bonds.	You must **identify** the term that best meets the criterion of the question ***Correct Answer: C***
Illustrate	Clarify through examples or comparisons	For two types of convergence hedges, what situations present profitable opportunities, how are the hedges set up, and what are the associated risks?	You must provide an example for each hedge or compare the two to **illustrate** how they work.

Action Words *continued*

Term	Definition	Question Example	Example of Term Use
Interpret	Explain the meaning	Your certificate of deposit will mature in one week, and you are considering how to invest the proceeds. If you invest in a 30-day CD, the bank will pay you 4% interest. If you invest in a 2-year CD, the bank will pay you 6% interest. You should choose the: A. 2-year CD if you expect that interest rates will fall in the future B. 30-day CD, no matter what you expect interest rates to do in the future. C. 2-year CD, no matter what you expect interest rates to do in the future.	You must **interpret** the features of an investment scenario. ***Correct Answer: A***
List	Create a series of items	List the determinants of real interest rates.	You must differentiate from a **list** those items that are consistent with the question.
Outline	Summarize tersely	Which of the following best characterizes the steps in computing a geometric mean return based on a series of periodic returns from *T* time periods? A. Add one to each return, add them together, divide by T and subtract one. B. Add one to each return, multiply them together, take the Tth root and subtract one. C. Add one to each return, multiply them together, divide by T and subtract one.	You must **outline** the study's most important findings rather than explain them in detail. ***Correct Answer: B***
Recognize	Recall the purpose of a given equation or term, and its name when appropriate. **Note:** When the action word "recognize" is used and applied to an equation, the equation will be provided within the question stem or the correct answer choice.	What is the following equation called and used for in the context of artificial neural networks? $\sigma(z) \equiv \frac{1}{1+e^{-z}}$ A. It is called a neuron and used to make a NAND gate. B. It is called a sigmoid function and is used to model sigmoid neurons that better enable learning than perceptrons do. C. It is called a perceptron which is used to create a smoother function than a logistic function.	You must **recognize** that this is a sigmoid function, also referred to as a logistic function, which is used as the basis for a sigmoid neuron. ***Correct Answer: B.***
Relate	Show or establish logical or causal connection	Which of the following effects does NOT help to explain growth in the venture capital industry? A. Amendments to the prudent man rule B. The rise of limited partnerships as an organizational form C. Decline in the valuations of small capitalization stocks	You must **relate** effects or factors (e.g., the prudent man rule) to another result or concept (e.g.) growth in an industry.) ***Correct Answer: C***

FDP Editorial Staff

Hossein Kazemi, PhD, CFA, Senior Advisor, The CAIA Association
Mirjam Dekker, Project Manager, The FDP Institute
Kathryn Wilkens, PhD, CAIA, Curriculum Advisor, The FDP Institute
Keith Black, Ph.D., CFA, CAIA, FDP, Managing Director Content Strategy at CAIA Association
Don Chambers, Ph.D., CAIA, Associate Director of Programs at CAIA Association
Hossein Pishro-Nik, Ph.D.,Associate Professor at University of Massachusetts Amherst