FDP INSTITUTE® by CAIA

FDP CHARTER CANDIDATE STUDY GUIDE

April 8 - April 22, 2024

Learning objectives and keywords to facilitate your exam study

Brought to you by: CAIA ASSOCIATION®

INTRODUCTION TO THE FINANCIAL DATA PROFESSIONAL (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 financial data science, an increasingly important part of the financial services industry.

The digital revolution has disrupted the financial industry in recent years. 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 finance professionals with an efficient way to learn about financial data science's essential aspects. 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 financial industry segments. Candidates will earn their FDP Charter once they pass the FDP exam and fulfill the online class requirements, which can be completed 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 data science applications to the financial industry.

Passing the FDP examination is an important accomplishment that will require significant preparation. All Candidates will need to study and become familiar with the FDP curriculum material to develop the knowledge and skills necessary to succeed on the examination day. A limited set of questions is available to anyone who joins the FDP Community by setting up a profile on the FDP website. These questions should help prospective Candidates understand the scope of the materials and type of questions before pursuing the Charter.

This study guide is organized to facilitate quick learning and easy retention. Each topic is structured around learning objectives (LOs) and keywords that define the content to be tested on the exam. The learning objectives and keywords are important ways for Candidates to organize their studies as they form the basis for examination questions. All learning objectives and keywords reflect the FDP curriculum content, and all exam questions are written to address the learning objectives or keywords directly.

To assist the FDP program Candidates, the Appendix contains a list of learning objectives, equations, and quantitative concepts that appear in FDP's curriculum readings. Consistent with the action words in the Study Guide, the Appendix shows which equations must be memorized and which must be recognized. Understanding and mastering these formulas are central to key algorithms and concepts of financial data science and are essential for completing the FDP exam.

A Candidate who has mastered all learning objectives and keywords in the study guide should be well-prepared for the exam. We believe that the FDP Institute has built a rigorous program with high standards while maintaining an awareness of the value of Candidates' time.

Candidates for the FDP Charter must complete the FDP exam and the online class requirements. Since the FDP program is designed for finance professionals, it is assumed that Candidates understand the central concepts of financial economics. Candidates are expected to have knowledge of various financial institutions and instruments' roles and characteristics and the financial models these institutions employ to value the instruments and measure their 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

FDP Candidates must complete the following two components with a passing score before obtaining their FDP Charters.

- **FDP exam.**
- **Online classes covering Python or R programming.**

The FDP exam will not contain any coding questions. However, FDP Candidates must demonstrate some Python or R programming language knowledge before they obtain their FDP charter. FDP Candidates who do not have a verifiable academic background in Python or R programming can demonstrate their understanding of these languages by completing the online classes listed below. The online classes can be completed before or after a Candidate completes the FDP exam.

The FDP Institute recommends DataCamp's (https://www.datacamp.com) or Udemy's (https://www.udemy.com/) introductory online courses for completing the FDP Charter's requirements. The list of acceptable online classes for the FDP Charter appears on the FDP Institute's website and in this Study Guide.

The approved online classes offered by DataCamp or Udemy are available as soon as a Candidate creates an account on DataCamp or Udemy. Limited free access to the classes is available.

The Candidate Handbook available on the FDP website describes the procedure for sending proof of successful completion of the online classes to the FDP Institute.

The classes listed below are recommended to complete the FDP Charter's programming knowledge requirement. These recommendations assume that a Candidate has no prior Python or R programming knowledge. If a Candidate has prior knowledge of these languages, the Candidate is encouraged to take more advanced Python or R programming classes at DataCamp or Udemy. If a Candidate has a verifiable academic background in Python or R, the Candidate can seek an exemption from the online classes. Approving the exemption of Python or R programming language is at the sole discretion of the FDP Institute. Please contact the FDP institute to learn more about this option.

FDP Candidates can satisfy the coding requirement of the FDP program by completing two Python or two R classes offered by DataCamp or by completing one Python or one R class offered by Udemy. All class options can be accessed through the website of the course providers. Candidates are responsible for the cost of classes offered at DataCamp or Udemy. Candidates are encouraged to take advantage of the limited free access to evaluate teaching methods of online course providers.

DataCamp: Python (both courses should be completed)

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

2.Intermediate Python
https://www.datacamp.com/courses/intermediate-python

DataCamp: R (both courses should be completed)

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

2.Intermediate R
https://www.datacamp.com/courses/intermediate-r

Udemy: Python (either one of the courses should be completed)

1. Python Programming For Beginners: Learn Python In 9 Days
https://www.udemy.com/course/python-programming-for-beginners-learn-python-in-9-days/

2.The Python Bible™ | Everything You Need to Program in Python
https://www.udemy.com/course/the-python-bible/

Udemy: R

1. R Programming - R Language for Absolute Beginners
https://www.udemy.com/course/r-for-absolute-beginners/

FDP EXAMINATION

The FDP examination is a four-hour, computer-administered examination offered at test centers worldwide. The FDP examination consists of 80 multiple-choice questions weighted 75% of the total points and two to four constructed response questions (multi-part essay type) weighted 25% of the total points. Approximately 30% to 40% of the total points will come from questions involving some calculations. The FDP exam will not contain any Python or R programming questions.

The FDP program is organized to facilitate quick learning and easy retention based on the study guide. 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 important ways for Candidates to organize their studies as they form the basis for examination questions. All learning objectives reflect the FDP curriculum content and all examination questions are written to address the learning objectives directly.

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

SAMPLE EXAM AND PRACTICE QUESTIONS

A sample exam is available for the Candidates to assist with their study efforts. This sample exam contains 80 multiple-choice questions and several multi-part constructed response questions. There is also a set of practice questions available to Candidates. The set of practice questions contains more questions than the number of questions in the actual exam. In addition to helping the Candidates learn the topic material, the questions can also help the Candidates get familiar with the style and conventions used. An example is a simplifying convention of using the natural logarithm to solve any problem requiring the calculation of logarithm on the exam. This convention is announced at the beginning of the sample exam and on the actual exam. This convention is also described in the Candidate Handbook.

OTHER STUDY TOOLS AND RESOURCES

In addition to this Study Guide and the Candidate Handbook, the FDP Institute website directs Candidates to the readings covered in the curriculum. The readings are detailed below by topic area and include textbooks, often used across topics, and individual articles that are usually topic specific. The textbooks can be purchased from Amazon or the publisher's website. Some of the individual articles are publicly available free of charge. These articles are also posted on the FDP Institute website.

For Candidates' convenience, three articles published by PMR Journals are provided in one collection titled "Alternative Data and Machine Learning in the Financial Industry: A Collection of Articles from PMR Journals." Free access is granted to all registered Candidates. This collection has two sets of page numbers: one corresponds to the collection's table of

contents, and the other corresponds to each article's page number in the original journal. The page numbers next to the keywords refer to the page numbers as they appeared in the original article.

Note: Check if your employer has a subscription to Portfolio Management Research (PMR) as this might provide free access to the six PMR readings.

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 completion for 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	5-10
2.	Linear and Logistic Regression, Support Vector Machines, Regularization, and Time Series	10-15
3.	Decision Trees, Supervised Segmentation, and Ensemble Methods	8-12
4.	Classification, Clustering, and Naïve Bayes	8-12
5.	Neural Networks and Reinforcement Learning	8-12
6.	Performance Evaluation, Back-Testing, and False Discoveries	5-10
7.	Textual Analysis and Large Language Models	10-15
8.	Ethics, Privacy, and Regulation	8-12
9.	Fintech Applications	15-25

THE FDP CURRICULUM: THE COMPLETE READING LIST

The following is a complete list of all April 2024 FDP exam curriculum readings.

Two of the four books listed below must be purchased. Please refer to the FDP website for links to the books available free of charge from the authors' website.

The "Alternative Data and Machine Learning in the Financial Industry: A Collection of Articles from PMR Journals" and the "Topics in Financial Data Science" articles are available free of charge to exam registrants. Candidates may access all materials from the authors' or the publishers' websites or via the FDP website. Please use the web link below to access all curriculum materials.

https://fdpinstitute.org/Curriculum-Materials

A. Books

1. Provost, F., and T. Fawcett (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapters 1-10. Candidates should visit the book's errata page.

2. Hull, J. C. (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapters 1-11.

3. James, G., D. Witten, T. Hastie, and R. Tibshirani (2021). An Introduction to Statistical Learning: With Applications in R. Springer, 2nd Edition. Chapters 1, 2 (sections 1, 2), Chapter 3 (sections 1-3), Chapter 6 (sections 1-3), and Chapter 8 (sections 1, 2). Candidates should visit the book's errata page. A pdf version of the book is available at https://hastie.su.domains/ISLR2/ISLRv2_website.pdf.

4. Wolfram, S. (2023). What is ChatGPT Doing … and Why Does It Work? Wolfram Media, Inc., 1st Edition. Pages 1-10, 41-74. An equivalent, no-cost version of the book can be found at https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

B. Alternative Data and Machine Learning in the Financial Industry: A Collection of Articles From the PMR Journals

1. Lo, A.W. and M. Singh. (2023). From ELIZA to ChatGPT: The Evolution of Natural Language Processing and Financial Applications. The Journal of Portfolio Management, 49 (7): Pages 201-235. Reading 7.4

2. Ekster, G. and Kolm P. N. (2021). Alternative Data in Investment Management: Usage, Challenges, and Valuation. The Journal of Financial Data Science, 3 (4): Pages 10-32. Reading 9.1

3. López de Prado, M. (2018). The 10 Reasons Most Machine Learning Funds Fail. The Journal of Portfolio Management, 44 (6): Pages 120-133. Reading 9.5

C. Topics in Financial Data Science

1. Das, S., and H. Kazemi (2022). Time Series: A Financial Perspective. The FDP Institute. This reading is provided by the FDP Institute free of charge. Reading 2.4

2. Colquhoun, D. (2014). An Investigation of the False Discovery Rate and the Misinterpretation of p-values. Royal Society Open Science, 1 (3): Pages 1-16. Reading 6.3

3. Jingwen, J., B. Kelly, and D. Xiu (2022). Expected Returns and Large Language Models. Available at SSRN: https://ssrn.com/abstract=4416687. Pages 5-14. Reading 7.5

4. Smith, G., and I. Rustagi (2020). Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook. Berkeley Haas Center for Equity, Gender, and Leadership. Reading 8.2

5. FinRegLab (2021). The Use of Machine Learning for Credit Underwriting: Market and Data Science Context. Reading 8.3

6. Recommendations for Regulating AI. Google (2023). Available at https://ai.google/static/documents/recommendations-for-regulating-ai.pdf. Reading 8.4

7. Bagattini, G., Z. Benetti, and C. Guagliano (2023). Artificial Intelligence in EU Securities Markets. European Securities and Markets Authority (ESMA). Reading 9.2

8. Harvey, C.R., Y. Liu and A. Saretto (2020). An Evaluation of Alternative Multiple Testing Methods for Finance Applications. Available at SSRN: https://ssrn.com/abstract=3480087 or http://dx.doi.org/10.2139/ssrn.3480087. Reading 9.3

9. Francis, L. A. (2006). Taming Text: An Introduction to Text Mining. Casualty Actuarial Society Forum, Pages 51-88. Reading 9.4

ACTION WORDS

In each learning objective that appears below, action words are used to direct Candidates' focus of study. The following table contains the list of all action words used in this Study Guide and their definitions.

Action Word	Meaning
Analyze	Examine the constitution or structure of the information or concept the LO covers methodically and in detail. This is similar to offering an explanation and an interpretation. It is used chiefly to explain relationships.
Apply	Use or employ a concept or a mathematical relationship (equation) to bring into action. If the LO is about an equation, the Candidate must memorize the equation (see Recognize below).
Calculate	It is similar to Apply but relates to a mathematical concept and equation. If the LO is about an equation, the Candidate must memorize the equation (see Recognize below).
Compare	Estimate, measure, or note the similarity or dissimilarity between two concepts or definitions.
Contrast	Similar to Compare. In this case, the emphasis is on the differences.
Define	A general action word. The Candidate is expected to state or describe precisely a concept's nature, scope, or meaning. If the LO is about a mathematical equation, the Candidate is not expected to memorize the exact equation but is expected to describe its essential aspects.
Describe	Similar to Define. The Candidate should give an account in words of concepts covered by the LO. The Candidate is expected to cover all the relevant characteristics, qualities, or relationships the LO covers. If the LO is about a mathematical equation, the Candidate is not expected to memorize the exact equation but is expected to describe its essential aspects.
Discuss	It is similar to Analyze. The Candidate is to provide details about a key word or concept. If the LO is about an equation, the Candidate does not need to memorize it but must know its uses and applications.

Action Word	Meaning
Explain	Similar to Describe. The Candidate is expected to clarify an idea, problem, or relationship by describing it in more detail or revealing relevant facts or ideas. If the LO is about a mathematical equation, the Candidate is not expected to memorize the exact equation but is expected to describe its essential aspects.
Identify	The Candidate is expected to recognize or establish as being a particular model, concept, or relationship. The LO may expect the Candidate to verify a given relationship or recognize a particular pattern. If the LO is about a mathematical equation, the Candidate is not expected to memorize the exact equation but is expected to describe its essential aspects.
Interpret	Similar to Explain. The Candidate is expected to give or provide an explanation for the observed pattern, relationship, or information. If the LO is about a mathematical equation, the Candidate is not expected to memorize the exact equation but is expected to describe its essential aspects.
List	The Candidate is expected to learn the list of related items or concepts the LO covers. The Candidate is not expected to describe the members of the list. A separate LO may state that some or all of the list's members must be explained.
Recognize	The Candidate is expected to identify an equation or model from the readings. The Candidate is not expected to memorize the equation. The Candidate is expected to apply the equation or make some calculations using the equation provided on the exam.

LEARNING OBJECTIVES

Topic 1. Introduction to Data Science

Reading 1.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapters 1 and 2.

Keywords

Data mining (p. 2)
Data Science (p. 2)
Data-driven decision making (p.5)
Big data (p. 8)
Classification (p. 20)
Regression (p. 21)
Similarity matching (p. 21)
Clustering (p. 21)
Co-occurrence grouping (p. 21)
Profiling (p. 22)
Link prediction (p. 22)
Data reduction (p. 22)
Causal modeling (p. 23)
Unsupervised learning (p. 24)
Supervised learning (p. 24)
Leak (p. 30)

Learning Objectives

Demonstrate proficiency in the following areas:

1.1.1 Data Analytic Thinking (Ch. 1)

For example:

A. List data mining examples in finance, marketing, and customer relationship management.

B. Contrast data science with data mining.

C. Describe the two types of decisions that can benefit from data-driven decision making.

D. Describe the reason for the finance and telecommunications industries' early adoption of automated decision-making.

E. Contrast data science with data processing.

F. Describe the usage of big data.

G. Explain why appropriate data and data scientists are required to extract useful knowledge from data.

H. Explain why it is necessary to understand data science even if someone will not use it directly.

I. List and describe the four fundamental concepts of data science.

1.1.2 Business Problems and Data Science Solutions (Ch. 2)

For example:

A. Describe when each type of data mining algorithm should be used, such as classification, regression, similarity matching, clustering, co-occurrence grouping, profiling, link-prediction, data reduction, and causal modeling.

B. Explain the differences between regression and classification.

C. Contrast supervised learning with unsupervised learning.

D. List the algorithms that can be used for supervised and unsupervised learning.

E. Contrast data mining with the use of data mining results.

F. List and describe the steps used in the Cross Industry Standard Process for Data Mining (CRISP-DM).

G. Explain the reason for having an iterative process involved in CRISP-DM.

H. Describe the characteristics of credit card and Medicare fraud.

I. List the reasons for deploying the data mining system itself rather than the models produced by a data mining system.

Reading 1.2 Hull, J. C. (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 1.

Keywords

Machine learning (p. 1)
Artificial intelligence (p. 1)
Features (p. 6)
Labels (p. 6)
Semi-supervised learning (p. 7)
Training set (p. 8)
Root-mean squared error (p. 9)
Bias-variance tradeoff (p. 15)
Numerical feature (p. 16)
Categorical feature (p. 16)
Outliers (p. 17)
Bayes' Theorem (p. 18)

Learning Objectives

Demonstrate proficiency in the areas of:

1.2.1 Introduction

For example:

A. List the advantages for society of replacing human decision-making with machines.

B. Contrast machine learning to statistics.

C. Describe a training set, validation set, and test set.

D. Define instances.

E. Analyze the relationship between model error and model complexity.

F. Define bias and variance in the context of machine learning.

G. List the usage of the training set, validation set, and test set.

H. List and explain different data cleaning issues.

I. List types of models that are least and most affected by outliers.

J. Calculate types conditional probability using Bayes' Theorem.

Reading 1.3 James, G., D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: With Applications in R. Springer, 2nd Edition. Chapters 1, 2.1, and 2.2.

Keywords

Statistical learning (p. 1)
Flexible (p. 22)
Thin plate spline (p. 23)
Classification problems (p. 28)
Quantitative variables (p. 28)
Qualitative response (p. 28)
Binary response (p. 28)
Predictors (p. 29)
Mean squared error (MSE) (p. 29)
Test MSE (p. 30)
Test data (p. 30)
Training MSE (p. 30)
Degrees of freedom (p. 31)
Expected test MSE (p. 34)
Bias (p. 35)
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)

Learning Objectives

Demonstrate proficiency in the areas of:

1.3.1 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.

1.3.2 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.

B. Explain various error terms (reducible and irreducible), the expected value of error squared, and the variance of error terms.

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

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

E. Explain when a supervised learning model is preferable to unsupervised or semi-supervised learning models.

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

1.3.3 Assessing Model Accuracy (Ch. 2.2)

For example:

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

B. Explain the goal of measuring the fit quality 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. Calculate the Bayes error rate.

G. Explain and apply the Bayesian classifier.

H. Explain how the K-nearest neighbors (KNN) classifier relates to the Bayes classifier and how the choice of K impacts results.

I. Calculate the conditional probability of a point belonging to a particular class.

J. Analyze the relationship between the value of K and the bias-variance tradeoff for a KNN classifier.

K. Explain what happens to the decision boundary as K increases in a KNN classifier.

Topic 2. Linear and Logistic Regression, Support Vector Machines, Regularization, and Time Series

Reading 2.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapter 4.

Keywords

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 (p. 97)
Log-odds (p. 99)
Logistic function (p. 101)
Nonlinear SVM (p. 107)
Neural networks (p. 108)

Learning Objectives

Demonstrate proficiency in the areas of:

2.1.1 Classification via Mathematical Functions

For example:

A. Apply the equation of a straight line using slope and intercept.
B. Describe, apply, and interpret a linear discriminant.
C. Calculate the best value for the parameters of a linear discriminant for a set of instances.
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 the general idea behind optimizing the objective function for a linear discriminant for a particular data set.
G. Describe how linear discriminant functions can be used for scoring and ranking instances.
H. Analyze the relationship between the distance from the decision boundary of a linear discriminant and the likelihood of response.
I. Describe the important idea behind the Support Vector Machine (SVM).
J. Describe the objective function of the SVM.
K. Explain how the objective function used in SVM utilizes the concept of the hinge-loss function.
L. Describe the reason for not using a squared loss function in classification problems.

2.1.2 Regression via Mathematical Functions

For example:

A. Describe the major drawback of least-squares regression.
B. Calculate odds and log odds.
C. List the important features of logistic regression.
D. Calculate the log-odds linear function.
E. Calculate class probability using the logistic function.
F. Describe the shape of the logistic function.
G. Describe the decision boundary for the logistic regression.
H. Describe how an objective function is formed in the logistic regression.
I. Compare and contrast classification trees with linear classifiers.
J. Explain the basic idea behind nonlinear SVMs and neural networks.

Reading 2.2 Hull, J. C. (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapters 3 and 5.

Keywords

Polynomial regression (p. 53)
One-hot encoding (p. 54)
Dummy variable trap (p. 55)
Regularization (p. 56)
Logistic regression (p. 69)
Sigmoid function (p. 70)
Balanced data set (p. 108)
Support vectors (p. 110)
Hard margin classification (p. 114)
Soft margin classification (p. 114)
Gaussian radial basis function (p. 118)
SVM regression (p. 119)

Learning Objectives

Demonstrate proficiency in the following areas:

2.2.1 Supervised Learning (Ch. 3)

For example:

A. List the conditions that must be satisfied for linear regression to be valid.
B. List the steps used in the gradient descent method.
C. Calculate the probability of a positive outcome using the sigmoid function.
D. Recognize the cost function for the logistic regression.
E. Analyze the effect of using different types of regularization on logistic regression.

Note that this chapter contains many other topics with no learning objectives specified in this section. Candidates are still encouraged to read these sections to understand subsequent material better. Questions from these missing topics will primarily be asked from the book "An Introduction to Statistical Learning: With Applications in R" by James, G., D. Witten, T. Hastie, and R. Tibshirani (Reading 2.3).

2.2.2 Support Vector Machines (Ch. 5)

For example:

A. List the advantages and disadvantages of using support vector machines (SVM).

B. Describe the reason for normalizing data before using it in SVM.

C. Calculate the dimension of a separating hyperplane.

D. Recognize the equation of a separating hyperplane with m features.

E. Describe the reasons for using regularization in SVM.

F. Recognize the objective function used in creating SVM with m features.

G. Recognize the objective function for a soft margin classification.

H. Describe the type of regularization used in soft margin classification.

I. Describe how violations and misclassifications are measured in soft margin classification.

J. Analyze the relationship between the hyperparameter, C, and the pathway width for soft margin classification.

K. Describe the general approach to finding a non-linear boundary when using a linear model.

L. Recognize the Gaussian radial basis function (RBF) for an observation.

M. Explain the effect of the parameter gamma on RBF.

2.2.3 SVM Regression (Ch. 5)

For example:

A. Describe how an error is calculated in SVM regression.

B. Recognize the equations of hyperplanes in SVM regression.

C. Recognize the objective function used in SVM regression.

D. Describe the interaction of the two terms in the objective function of an SVM regression.

E. Contrast simple linear regression with SVM.

Reading 2.3 James, G., D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: With Applications in R. Springer, 2nd Edition. Chapters 3.1, 3.2, 3.3, 6.1, 6.2, and 6.3.

Keywords

Residual (p.61)
Residual sum of squares (p. 63)
Population regression line (p. 63)
Least squares line (p. 63)
Bias (p. 65)
Unbiased (p. 65)
Standard error (pg. 65)
Residual standard error (p. 66)
Confidence interval (p. 66)
Null hypothesis (p. 67)
Alternative hypothesis (p. 67)
t-statistic (p. 67)
R^2 statistic (p. 68)
Total sum of squares (p. 70)
F-statistic (p. 75)
Forward selection (p. 79)
Backward selection (p. 79)
Mixed selection (p. 79)
Prediction interval (p. 82)
Dummy variable (p. 83)
Additive (p. 87)
Linear (p. 87)
Hierarchical principle (p. 89)
Residual plot (p. 93)
Heteroscedasticity (p. 96)
Outlier (p. 97)
Collinearity (p. 99)
Power (p. 101)
Multicollinearity (p. 102)
Variance inflation factor (p. 102)
Feature selection (p. 226)
Variable selection (p. 226)
Best subset selection (p. 227)
Deviance (p. 228)
Forward stepwise selection (p. 229)
Backward stepwise selection (p. 231)
C_p (p. 233)
Akaike information criterion (AIC) (p. 233)
Bayesian information criterion (BIC) (p. 233)
Adjusted R^2 (p. 233)
Ridge regression (p. 237)
Tuning parameter (p. 237)
Shrinkage penalty (p. 237)
ℓ_2 norm (p. 238)
Scale equivariant (p. 239)
Lasso (p. 241)
ℓ_1 norm (p. 241)
Sparse (p. 242)
Soft-thresholding (p. 248)
Signal and noise variables (p. 250)
Dimension reduction methods (p. 251)
Linear combination (p. 251)
Principal component analysis (p. 252)
Principal component scores (p. 254)
Orthogonal (p. 256)
Principal component regression (p. 252)
Partial least squares (p. 260)

Learning Objectives

Demonstrate proficiency in the following areas:

2.3.1 Simple Linear Regression (Ch. 3.1)

For example:

A. Calculate the value of RSS.

B. Calculate the least-squares coefficient estimates.
C. Interpret least-squares coefficients.
D. Recognize the standard error of a statistic.
E. Apply standard errors of linear regression.
F. Calculate the 95% confidence interval.
G. Calculate the t-statistic.
H. Explain the rules for rejecting the null hypothesis using p-values.
I. Explain the accuracy of linear regression.
J. Calculate and interpret the R^2 statistic.
K. Describe the advantages of the R^2 statistic over the RSE.
L. Calculate the correlation from R^2 for the simple linear regression.

2.3.2 Multiple Linear Regression (Ch. 3.2)

For example:

A. Interpret the coefficients of multiple linear regression.
B. Describe how a multiple linear regression tests the relationship between responses and predictors.
C. Recognize the F-statistic given TSS, RSS, n, and p.
D. Explain how the F-statistic can be used for hypothesis testing.
E. Explain why the t-statistic value can be a misleading indicator of variable importance in multiple regression.
F. Describe how to determine the importance of variables in a multiple regression.
G. Describe the tools used to examine model fit for multiple regression.
H. Calculate RSE given the values of RSS, n, and p.

2.3.3 Considerations in the Regression Model (Ch. 3.3)

For example:

A. Apply dummy variables.
B. Describe using 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. Describe the interaction effect.
F. Interpret the coefficients of an interaction term.
G. Explain when an interaction term should be added to a multiple regression model.
H. Describe the potential problems related to non-linearity, correlation of error terms, the non-constant variance of error terms, outliers, high-leverage points, and collinearity for a linear regression model.
I. Explain what happens to standard errors and confidence intervals in the presence of correlated errors.

J. Explain how heteroscedasticity can be mitigated using data transformation.

K. Describe high leverage points and the leverage statistic.

L. Explain how high leverage points can be detected using the leverage statistic.

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

N. Recognize the variance inflation factor.

2.3.4 Subset Selection (Ch. 6.1)

For example:

A. Define the best subset selection.

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

C. Analyze the relationship between the number of variables and RSS (or R^2) for multiple linear regression.

D. Explain the effect of low RSS (or high R^2) on training and test error.

E. Explain the role of deviance in a logistic regression model.

F. Analyze the relationship between the value of deviance and the fit of a model.

G. Describe the key drawback of using the best subset selection.

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

I. Explain the advantage of forward stepwise regression over the best subset selection method.

J. Describe a disadvantage of forward stepwise regression and backward stepwise regression relative to the best subset selection model.

K. Describe a key requirement for the number of samples and predictors when using the backward stepwise regression.

L. Describe the hybrid approach of using forward and backward stepwise regression together.

M. List the two common approaches to selecting the best model concerning test error.

N. Explain the reason for not using the training set RSS and training set R^2 for selecting the best model from a set of models with different predictors.

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

P. Describe the decision rule for selecting a model based on C_p, AIC, and BIC.

Q. Analyze the interaction between the RSS and the penalty term in C_p, AIC, and BIC.

R. Recognize the adjusted R^2.

2.3.5 Ridge Regression (Ch. 6.2)

For example:

A. Recognize the objective function of ridge regression.

B. Explain the effect of the tuning parameter on the coefficients in ridge regression.

C. Explain when the ridge regression is equivalent to the least-squares regression model.

D. Explain when the ridge regression is equivalent to the null model.

E. Calculate the ℓ_2 norm.

F. Explain the effect of multiplying a predictor by a factor before using it in the ridge regression.

G. Describe standardizing the predictors.

H. Explain what happens to the bias-variance trade-off as the tuning parameter changes in ridge regression.

I. Describe what happens to the least-squares coefficients when the number of variables is as large as the number of observations.

J. Describe when ridge regression can be used but least-squares regression cannot be used.

K. Describe the advantage of ridge regression over best subset selection.

2.3.6 The Lasso (Ch. 6.2)

For example:

A. Describe the key disadvantage of ridge regression.

B. Describe the advantage of Lasso over ridge regression.

C. Recognize the objective function of Lasso.

D. Calculate the ℓ_1 norm.

E. Describe the variable selection property of Lasso.

F. Explain the effect of the tuning parameter on the coefficients in Lasso.

G. Recognize the alternative formulation of the objective function for Lasso and ridge regression.

H. Analyze the impact of the size of the budget in estimating Lasso and ridge regression.

I. Explain the graphical interpretation of Lasso and ridge regression when there are two features.

J. Describe the geometric shape of the constraint for Lasso and ridge regression in two or more dimensions.

K. List the key advantage of Lasso over ridge regression.

L. Describe when Lasso is expected to perform better than ridge regression and when ridge regression is expected to perform better than Lasso.

M. Explain the relationship between best subset selection and Lasso or ridge regression.

N. Explain the type of shrinkage done by Lasso and ridge regression.

O. Describe the rule governing the tuning parameter selection for Lasso and ridge regression.

P. Analyze the expected values of coefficients for the signal and noise variables for a robust regression model.

2.3.7 Principal Component Analysis (Ch. 6.3)

For example:

A. List the two ways of controlling variance.

B. Describe the relationship between the number of features and the number of parameters estimated in a dimension reduction method.

C. List the two major steps used in a dimension reduction method.

D. Explain the characteristics of the first principal component.

E. Explain the meaning of projecting a point on a line.

F. Describe the constraint that must be used to find the loadings for the principal components.

G. Describe an alternative interpretation for principal component analysis (PCA).

H. Explain the information content of the first principal component.

I. Explain the effect of zero correlation between the first and the second principal component.

J. Explain orthogonal properties of principal components.

K. Explain the expected information content of the second principal component when there are two predictors.

L. Analyze the relationship between the number of principal components and the number of features.

M. Describe the key assumption behind using principal component regression (PCR).

N. Explain the problem mitigated by using PCR provided the assumptions underlying PCR holds.

O. Explain when PCR is expected to perform better than linear regression with all features.

P. Explain why PCR is not a feature selection method.

Q. Describe the equivalence between the PCR and ridge regression.

R. Explain the process of selecting the number of principal components.

S. Describe when to standardize features before using PCR and when not to standardize the features for using them in PCR.

2.3.8 Partial Least Squares (Ch. 6.3)

For example:

A. List the key drawback of principal component regression (PCR).

B. Describe the key difference between PCR and partial least squares (PLS).

C. Describe the way the first PLS is found.

D. Analyze the impact of least-squares coefficients from the simple linear regression of each feature on the weight of the first PLS.

E. Describe the process of finding the second PLS.

Reading 2.4 Das, S. and H. Kazemi (2022). Time Series: A Financial Perspective. The FDP Institute.

Keywords

Strictly stationary (p. 3)
Weakly stationary (p. 4)
Gaussian White noise (p. 4)
Random walk (p. 4)
Simple moving average (p. 6)
Weighted moving average (p. 7)
Exponentially weighted moving average (p. 7)
Autoregressive model (p. 10)
Moving average process (p. 18)
Autoregressive moving average models (p. 21)
Homoskedasticity (p. 21)
Volatility clustering (p. 25)
Engle's ARCH test (p. 27)
Persistence Parameter (p. 28)

Learning Objectives

Demonstrate proficiency in the following areas:

2.4.1 Stationary Time Series and Moving Average Methods

For example:

A. Explain the reason for making a time series stationary before analyzing it.

B. Describe how stock prices can be converted to a stationary series.

C. List the conditions that are satisfied by a strictly stationary time series.

D. Describe the process that can be used to detect the presence of stationarity.

E. Describe the characteristics of the autocorrelation function of a stationary series and a non-stationary series.

F. List the conditions satisfied by Gaussian white noise.

G. Analyze the effect of window size on a simple moving average (SMA).

H. List a key advantage of using an SMA.

I. Describe the most common range of values for the weighting parameter in the exponentially weighted moving average (EWMA).

J. Calculate the value of EWMA for a series.

K. Explain the impact of the weighting parameter on EWMA.

L. Describe the choice of the weighting parameter that makes EWMA equivalent to SMA.

2.4.2 Autoregressive Models

For example:

A. Describe one of the key differences between autoregressive models and moving average methods.

B. List the conditions required for an AR(1) process to be stationary.

C. Analyze the process that pulls a stationary AR(1) process close to its mean.

D. Calculate the mean, variance, autocovariance, and autocorrelation of a stationary AR(1) process.

E. Analyze the effect of the coefficients of a stationary AR(1) process on autocorrelation.

F. Calculate the value of an AR(1) process.

G. Explain how a shock affects a stationary AR(1) process.

H. Calculate the mean and the variance of a random walk.

I. Explain how a shock affects a random walk.

J. Explain why some financial time series cannot be modeled as a random walk process.

K. Calculate conditional forecast of mean and variance for a stationary AR(1) model.

2.4.3 Moving Average Models

For example:

A. Define the order of a moving average model.

B. Recognize the unconditional mean, variance, and autocovariance of a moving average model.

C. Calculate the conditional value of the mean of forecast for an MA(1) model.

D. Recognize the conditional value of variance of forecast error of an MA(1) model.

E. Recognize the unconditional mean of an ARMA(p, q) model.

F. Explain the characteristics of the autocorrelation function for an ARMA(p, q) model.

2.4.4 Volatility Models

For example:

A. Analyze the effect of heteroskedasticity on the standard errors and confidence intervals for least-squares regression.

B. Describe the advantages of ARCH and GARCH models.

C. List the stylized facts that indicate financial time series error terms are not homoscedastic.

D. List the challenges in creating a time-varying volatility model.

E. Explain how an ARCH(1) model satisfies the challenges of creating a time-varying volatility model.

F. Calculate the conditional and unconditional variance for the error term when an ARCH(1) model is used.

G. List the conditions that must be satisfied by the parameters of an ARCH(1) model for the model to be stationary.

H. Describe the weaknesses of the ARCH model.

I. Describe the restrictions imposed on GARCH(1,1) model parameters.

J. Calculate the long-term mean of volatility for a GARCH(1,1) model.

K. Explain the effect of the persistent parameter on a GARCH(1,1) model.

L. Analyze the equivalence between an ARCH(1) and a GARCH(1,1) model.

M. Calculate the forecasted value of volatility using a GARCH(1,1) model.

Topic 3. Decision Trees, Supervised Segmentation, and Ensemble Methods

Reading 3.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapters 3 and 5.

Keywords

Information (p. 43)
Tree induction (p. 44)
Predictive model (p. 45)
Instance (p. 46)
Descriptive modeling (p. 46)
Feature vector (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)
Entropy (p. 51)
Information gain (p. 51)
Parent set (p. 52)
Children set (p. 52)
Variance (p. 56)
Entropy graph/chart (p. 58)
Leaf (p. 63)
Decision nodes (p. 63)
Classification tree (p. 63)
Regression tree (p. 64)
Probability estimation tree (p. 64)
Decision line (p. 69)
Decision surface or boundary (p. 69)
Frequency-based estimation of class partitions (p.70)
Membership probability (p. 72)
Laplace correction (p. 73)
Generalization (p. 112)
Generalization performance (p. 113)
Overfitting (p. 113)
Fitting graph (p. 113)
Holdout data (p. 113)
Test set (p. 114)
Base rate (p. 115)
Sweet spot (p. 117)
Cross-validation (p. 126)
Folds (p. 127)
Learning curve (p. 131)
Complexity (p. 131)
Sub-training set (p. 134)
Pruning (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)
Penalty function (p. 138)

Learning Objectives

Demonstrate proficiency in the following areas:

3.1.1 Models, Induction and Prediction (Ch. 3)

For example:

A. Define prediction in the context of data science.

B. Compare and contrast predictive modeling with descriptive modeling.

C. Compare and contrast induction with deduction.

3.1.2 Supervised Segmentation (Ch. 3)

For example:

A. List the complications arising from selecting informative attributes.

B. Calculate the value of entropy.

C. Recognize and apply entropy with the maximum and minimum disorder.

D. Contrast the parent set with the children set.

E. Calculate information gain for children sets from a parent set.

F. Discuss the issues with numerical variables for supervised segmentation.

G. Discuss the application of variance to numeric variables for supervised segmentation.

H. Describe how entropy and an entropy chart can be used to select an informative variable.

3.1.3 Visualizing Segmentations and Probability Estimation (Ch. 3)

For example:

A. Describe the relationship between the decision surface and the number of variables.

B. Define frequency-based estimation of class membership probability.

C. Calculate the probability at each node of a decision tree.

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

E. Calculate the value of the Laplace correction.

F. Explain how one can determine the predictive power of each attribute.

3.1.4 Generalization, Overfitting, and Its Avoidance (Ch. 5)

For example:

A. Apply the graph fitting concept to find the optimal tree induction model.

B. Describe the relationship between complexity and error rates.

C. Describe the relationship between tree size and accuracy.

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. Compare and contrast a learning curve with a fitting graph.

H. Describe the shape of learning curves for logistic regression and tree induction.

I. List and describe strategies that can be used to avoid overfitting in tree induction.

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

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

L. Explain nested cross-validation.

M. Describe the main idea behind regularization.

N. Analyze the relationship between overfitting and multiple comparisons.

Reading 3.2 Hull, J. C. (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 4.

Keywords

Gini measure (p. 88)
Naïve Bayesian classifier (p. 94)
Bagging (p. 94)
Random forest (p. 95)
Boosting (p. 95)
Ensemble learning (p. 102)

Learning Objectives

Demonstrate proficiency in the following areas:

3.2.1 Decision Trees

For example:

A. Describe the advantages of decision trees over linear or logistic regression.

B. Describe and calculate entropy, information gain, and Gini measures.

C. Describe and calculate the confusion matrix for a decision tree.

D. Describe and calculate various points of an ROC curve given various confusion matrices.

3.2.2 The Naïve Bayes Classifier

For example:

A. Describe and apply Bayes' theorem.

B. Calculate conditional probabilities using Bayes' formula.

C. Explain the conditions under which the Naïve Bayes classifier can be applied.

D. Apply Naïve Bayes classifier to a decision tree problem.

E. Describe the criterion for determining the optimal feature choice and its threshold when the target is a continuous variable.

3.2.3 Ensemble Learning

For example:

A. Describe the primary idea behind ensemble learning.

B. Describe bagging with or without replacement.

C. Describe the random forest approach.

D. Describe boosting.

Reading 3.3 James, G., D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: With Applications in R. Springer, 2nd Edition. Chapters 8.1 and 8.2.

Keywords

Tree based method (p. 327)
Terminal nodes or leaves (p. 329)
Internal nodes (p. 329)
Stratification (p. 330)
Top-down approach (p. 330)
Bottom-up approach (p. 330)
Recursive binary splitting (p. 330)
Subtree (p. 331)
Cost complexity (p. 332)
Weakest link (p. 332)
Classification error rate (p. 335)
Gini index (p. 336)
Weak learner (p. 340)
Majority vote (p. 341)
Out-of-bag observations (p. 342)
Variable importance (p. 343)
Stump (p. 347)
Interaction depth (p. 347)
Bayesian additive regression trees (p. 348)

Learning Objectives

Demonstrate proficiency in the following areas:

3.3.1 The Basics of Decision Trees (Ch. 8.1)

For example:

A. Apply and interpret a decision tree's predictions.
B. Explain and apply a regression tree and a partition.
C. Recognize and interpret RSS for a given partition (box).
D. Recognize RSS to perform recursive binary splitting.
E. Describe tree pruning, specifically cost complexity (weakest link) pruning.
F. Compare regression and classification trees.
G. Describe the construction of classification trees using the classification error rate, Gini index, and entropy.
H. Calculate the Gini Index.
I. Contrast tree-based methods and linear models.
J. Describe the advantages and disadvantages of trees.

3.3.2 Bagging, Random Forests, Boosting, and Bayesian Additive Regression Tree (Ch. 8.2)

For example:

A. Describe bagging and out-of-bag error estimation.
B. Explain how low-variance procedures can be created from high-variance ones.
C. Describe how qualitative targets are predicted using bagging.
D. Describe the out-of-bag error and its importance.
E. Describe how variable importance measures can be created using the Gini index.

F. Describe how random forest attempts to decorrelate trees.

G. Compare and contrast random forests to bagging.

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

I. Explain why boosting is described as a slow learner.

J. Describe the key difference between BART and other ensemble methods, such as random forest and boosting.

Topic 4. Classification, Clustering, and Naïve Bayes

Reading 4.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapters 6 and 9.

Keywords

Euclidean distance (p. 143)
Nearest neighbors (p. 144)
Combining function (p. 147)
Weighted voting (p. 150)
Similarity moderate voting (p. 150)
Complexity parameter (p. 152)
Classification boundaries (153)
Intelligibility (p. 154)
Feature selection (p. 156)
Domain knowledge (p. 156)
Manhattan distance (p. 158)
Jaccard distance (p.159)
Levenshtein metric (p. 160)
Hierarchical clustering (p. 164)
Dendrogram (p. 164)
Linkage function (p. 166)
Centroids (p. 169)
Clusters' distortion (p. 172)
CRISP process (p. 183)
Joint probability (p. 236)
Independent events (p. 236)
Unconditional probability (p. 237)
Bayes' Rule (p. 237)
Prior (p. 238)
Posterior probability (p. 238)
Likelihood (p. 240)
Conditional independence (p. 241)
Naïve Bayes equation (p. 241)
Lift (p. 244)

Learning Objectives

Demonstrate proficiency in the following areas:

4.1.1 Calculating and Interpreting Similarity and Distance (Ch. 6)

For example:

A. Calculate the Euclidean distance.
B. Explain how combining functions can be used for classification.
C. Calculate the probability of belonging to a class based on the nearest neighbor classification.
D. Explain weighted voting (scoring) or similarity moderated voting (scoring).
E. Calculate contributions and class probabilities using weighted voting.
F. Explain how k in k-NN (Nearest-Neighbor) can address overfitting.
G. Describe issues with nearest-neighbor methods focusing on intelligibility, dimensionality, domain knowledge, and computational efficiency.
H. Describe two aspects of intelligibility.
I. Explain how the curse of dimensionality could be fixed using domain knowledge.
J. Interpret and calculate Manhattan and Cosine distance.

K. Describe and interpret combining functions.

L. Describe the primary idea behind clustering.

M. Describe the primary idea behind hierarchical clustering.

N. Describe the general approach to k-means clustering using centroids.

O. Explain the role of supervised learning in interpreting cluster analysis results.

4.1.2 Combining Evidence Probabilistically (Ch. 9)

For example:

A. Calculate joint probability for independent and dependent events.

B. Explain and apply Bayes' Rule with the help of an example.

C. Calculate posterior probability, prior, and likelihood.

D. Explain and apply the Naïve Bayes classifier.

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

F. List the advantages and disadvantages of the Naïve Bayes classifier.

G. Explain and calculate lift in the context of the Naïve Bayes method.

H. Define the generative model and Naïve-Naïve Bayes.

Reading 4.2 Hull, J. C. (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 2.

Keywords

Scaled feature (p. 24)
Z-score (p. 24)
Min-max scaling (p. 24)
k-means (p. 25)
Inertia (p. 30)
Elbow method (p. 30)
Silhouette method (p. 31)
Gap statistic (p. 32)
Curse of dimensionality (p. 33)
Cosine function (p. 33)
Principal component (p. 41)
Factor loading (p. 42)

Learning Objectives

Demonstrate proficiency in the following areas:

4.2.1 Unsupervised Learning

For example:

A. Calculate and interpret feature scaling using Z-score and mini-max.

B. Interpret the Euclidean distance.

C. Calculate and interpret the centroid of a cluster.

D. Describe the primary features and the process of implementing the k-means algorithm.

E. Calculate and interpret inertia as a measure of the clustering algorithm.
F. Describe the elbow method for selecting the number of clusters.
G. Describe and apply the silhouette method for selecting the number of clusters.
H. Describe and apply the gap statistic for selecting the number of clusters.
I. Describe the primary features of the hierarchical clustering method.
J. Describe the primary features of principal component analysis and how it relates to cluster analysis.

Topic 5. Neural Networks and Reinforcement Learning

Reading 5.1 Hull, J. C. (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapters 6, 7, and 8.

Keywords

Artificial neural network (ANN) (p. 125)
Multi-layer perceptrons (p. 125)
Hidden layer (p. 125)
Input layer (p. 126)
Output layer (p. 126)
Bias (p. 126)
Activation function (p. 126)
Cost function (p. 128)
Universal approximation theorem (p. 130)
ReLU activation function (p. 132)
Leaky ReLU activation function (p. 132)
Hyperbolic tangent activation function (p. 132)
Learning rate (p. 134)
Gradient descent algorithm (p. 134)
Backpropagation (p. 139)
L1 regularization (p. 140)
L2 regularization (p. 140)
Epoch (p. 140)
Mini-batch stochastic gradient descent (p. 140)
Gradient descent with momentum (p. 140)
Gradient descent with adaptive learning rate (p. 140)
Learning rate decay (p. 141)
Gradient descent with dropouts (p. 141)
Adam (p. 141)
Stopping rule (p. 141)
Implied volatility (p. 148)
Moneyness (p. 148)
Delta (p. 148)
Volatility surface (p. 148)
Autoencoders (p. 155)
Latent variables (p. 156)
Encoder (p. 157)
Decoder (p. 157)
Variational autoencoders (p. 160)
Kullback-Leibler divergence (p. 161)
Generative adversarial networks (p. 161)
Recurrent neural network (RNN) (p. 163)
Long short-term memory (LSTM) (p. 165)
Convolutional neural networks (CNN) (p. 165)
Feature map (p. 165)
Receptive field (p. 166)
Filter (p. 167)
Pooling (p. 167)
Flattening (p. 167)
Stride (p. 167)
Padding (p. 167)
Temporal convolutional network (TCN) (p. 168)
Reinforcement learning (p. 171)
Rewards (p. 171)
Exploitation choice (p. 172)
Exploration choice (p. 172)
Greedy action (p. 172)
Non-greedy action (p. 172)
Temporal difference learning (p. 182)
n-step bootstrapping (p. 185)
Deep reinforcement learning or deep Q-learning (p. 186)

Learning Objectives

Demonstrate proficiency in the following areas:

5.1.1 ANNs and Activation Functions (Ch. 6)

For example:

A. Describe an artificial neural network (ANN) with a single hidden layer.

B. Explain the downside of having a linear (or identity) activation function.

C. Calculate the value of a sigmoid function from weights and bias.

D. Explain the reason for using the linear activation function for numerical output values.

E. Explain when using the sigmoid function for the output layer is appropriate.

F. Recognize the number of parameters to be estimated for an ANN with a single hidden layer.

G. Recognize the cost function for an ANN.

H. Recognize the outputs of ReLU and leaky ReLU activation functions.

I. Calculate the output of hyperbolic tangent activation functions.

J. Identify the shapes of sigmoid, ReLU, leaky ReLU, and hyperbolic tangent activation functions.

5.1.2 Gradient Descent Algorithm (Ch. 6)

For example:

A. Calculate the change in the value of a function using the learning rate.

B. Explain the reason for requiring a good value for the learning rate.

C. Explain the reason for scaling all variables before using them in the gradient descent algorithm.

D. Calculate the gradient of a function with multiple features.

E. Calculate the relationship between a function and its scaled version.

F. Explain the reason for using backpropagation.

G. Describe how the partial derivative of an objective function can be calculated using backpropagation.

H. Describe the usage of L1 and L2 regularization in the objective function of neural networks.

I. Analyze the effect of L1 and L2 regularization in the objective function of neural networks.

J. Describe mini-batch stochastic gradient descent.

K. Explain the way Adam selects the learning rate.

L. Analyze the relationship between a neural network's learning rate and the different iteration stages.

M. Explain the adjustment required when gradient descent with dropouts is used.

N. Explain the reason for not minimizing the cost function with many parameters for the training set.

O. Describe the most commonly used stopping rule.

5.1.3 Volatility Modeling (Ch. 6)

For example:

A. Describe the advantage of neural networks over Monte Carlo simulation.

B. List the advantages and disadvantages of using neural networks for derivative pricing.

C. Explain the reason for observing many variations in the pattern of implied volatility.

D. List the reasons for the need to understand movement in the volatility surface.

5.1.4 Applications of Neural Networks (Ch. 7)

For example:

A. Describe the objective of an autoencoder.

B. Explain how the number of neurons in the hidden layer is determined in an autoencoder.

C. Recognize the objective function of an autoencoder.

D. List the advantages of PCA and autoencoders.

E. Describe variational autoencoder (VAE) and its key objective.

F. Contrast an autoencoder with a VAE.

G. Describe the two components of the objective function of a VAE.

H. Analyze the effect of the single hyperparameter on the twin objective of a VAE.

I. Describe a generative adversarial network (GAN) and the two types of networks.

J. Describe the objective of a GAN.

K. Explain what happens to the maximum likelihood function when the GAN gets everything correct and when it does not.

L. Describe the key difference between a recurrent neural network (RNN) and an ANN.

M. List applications of RNN.

N. Describe the relationship between RNN and an exponentially weighted moving average.

O. Explain the key problem of RNN that is overcome by the Long Short-term Memory (LSTM) network.

P. Describe the key difference between an ANN and a convolutional neural network (CNN).

Q. Analyze the effect of applying a filter to a feature map.

R. List the key advantages of the architecture used in a CNN.

S. Describe how stride reduces the size of feature maps.

T. Explain the reason for using padding.

U. Describe the temporal convolutional network (TCN).

5.1.5 Reinforcement Learning (Ch. 8)

For example:

A. Describe the objective of a reinforcement learning algorithm.

B. Analyze the relationship between the probability of exploration and the number of trials.

C. Recognize the probability of exploration using a decay factor.

D. List the two quantities that are needed for updating expected rewards.

E. Explain what happens when a low number is chosen for the decay factor in the multi-arms bandit problem.

F. Recognize the objective function having discount factor for reinforcement learning with changing environment.

G. Describe the characteristics of a state in reinforcement learning with a changing environment.

H. Explain the reason for assigning more weights to later trials for reinforcement learning with changing environments.

I. Describe the key concept of dynamic programming.

J. Recognize the updated values of reward using temporal difference updating.

K. Identify the reason for using an artificial neural network with reinforcement learning.

L. Explain the process of converting the Q-values as the probability of winning.

M. List applications of reinforcement learning.

N. List the problems faced in using reinforcement learning to treat a patient.

O. Explain the application of reinforcement learning to portfolio management and hedging a derivatives portfolio.

P. Describe how reinforcement learning can be used when limited data is available.

Topic 6. Performance Evaluation, Back-Testing, and False Discoveries

Reading 6.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapters 7 and 8.

Keywords

Accuracy (p. 189)
Confusion matrix (p. 189)
False positive (p. 190)
False negative (p. 190)
True positive (p. 200)
True negative (p. 200)
Class prior (p. 201)
Precision (p.203)
Recall (p.203)
F-measure (p. 204)
Sensitivity (p. 204)
Specificity (p. 204)
Majority Classifier (p. 205)
Ranking classifier (p.210)
Profit curve (p. 212)
ROC graph (p. 215)
Hit rate (p. 216)
False alarm rate (p. 216)
Conservative classifier (p. 216)
Permissive classifier (p. 217)
AUC (p. 219)
Lift curve (p. 219)
Cumulative response curve (p. 219)

Learning Objectives

Demonstrate proficiency in the following areas:

6.1.1 Describing and Evaluating Classifiers (Ch. 7)

For example:

A. Calculate accuracy and error rate.
B. Identify false positives and false negatives within a confusion matrix.
C. Describe unbalanced data and the problems associated with unbalanced data.
D. Calculate the accuracy of a model developed using a balanced dataset but applied to an unbalanced dataset.
E. Discuss the problems with unequal costs and benefits of errors.

6.1.2 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. Calculate the minimum probability of response for which a customer should be targeted.
D. Describe how the expected value can be used to frame classifier evaluation.
E. Recognize the expected profit for a classifier with and without using priors.
F. Describe the two common pitfalls to formulating cost-benefit analysis.

G. Calculate true positive, false positive, true negative, and false negative rates for a confusion matrix.
H. Calculate and interpret precision and recall.
I. Calculate the value of the F-measure.
J. Calculate specificity and sensitivity.
K. Describe the reasons for the need to have a baseline model.

6.1.3 Visualizing Model Performance (Ch. 8)

For example:

A. Describe how thresholding can create different confusion matrices.
B. Calculate a confusion matrix using a threshold.
C. List the variables used on both axes of a profit curve.
D. Describe the properties of a profit curve.
E. Recognize points on a profit curve.
F. Calculate the proportion of sample data that can be targeted when a fixed budget is available.
G. List the two critical conditions that must be met for using the profit curve.
H. Describe the ROC graph, including the variables used on the x-axis and the y-axis.
I. Calculate points on an ROC graph using data from a confusion matrix.
J. Describe the four corners and the diagonal of the ROC graph.
K. Analyze the behavior of a random classifier on the ROC graph.
L. Describe how to use the ROC space to evaluate classifiers.
M. Describe a key advantage of using the ROC graph.
N. Explain the equivalence between the AUC and the Gini Index.
O. List the variables used on the x-axis and the y-axis for the cumulative response curve.
P. Explain the equivalence between the lift curve and the cumulative response curve.
Q. Describe the key assumption in creating the lift curve or the cumulative response curve.
R. Calculate points on a cumulative response curve.

Reading 6.2 Hull, J. C. (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 10.

Keywords

Model interpretability (p. 214)
White boxes (p. 215)
Black boxes (p. 215)
Partial dependence plot (p. 223)
Shapley values (p. 223)
Local interpretable model-agnostic explanations (LIME) (p. 226)

Learning Objectives

Demonstrate proficiency in the following areas:

6.2.1 Model Interpretability

For example:

A. Explain the reason for the need to understand how predictions are made.

B. List examples of black boxes and white boxes.

C. Interpret the value of weights in linear regression.

D. Interpret the value of bias in a linear regression when the features are measured as the difference from their means.

E. Calculate confidence limits for sensitivities using the t-statistic.

F. Explain the impact of a particular feature when the difference from the mean of the feature is used in a linear regression.

G. List an important reason for using regularization.

H. Calculate the combined impact of all features in a linear regression when the difference from the mean is used as features.

I. Calculate the probability of a positive and negative outcome for logistic regression.

J. Recognize the probability of an increase in positive outcomes in a logistic regression for small changes in the value of a continuous or categorical feature.

K. Recognize the odds against a given probability.

L. Calculate probabilities from odds on or odds against.

M. List the steps used in creating an expected conditional prediction to understand the role of a particular feature in the prediction.

N. Describe the shape of a partial dependence plot for the linear regression.

O. Explain the difficulty in measuring the combined effect of all features for a non-linear model.

P. Calculate the contribution of features using Shapley values.

Q. List the properties illustrated by the use of Shapley values.

R. List the limitations of Shapley values.

S. List the steps used in LIME.

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

Positive predictive power (p.2) *Inflation effect (p. 9)*

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. Describe the false discovery rate with the help of a tree diagram.

B. Calculate the probability of real effect given a result is significant.

C. Recognize the false discovery rate.

D. Describe an underpowered study.

E. Describe the inflation effect in the context of false discovery.

F. Describe what happens when we consider p=0.05 rather than p<=0.05.

G. Describe Berger's approach.

H. Calculate the false discovery rate using conditional probabilities.

I. Calculate the conditional probability of the real effect.

J. Calculate the odds ratio using the Bayes approach.

Topic 7. Textual Analysis and Large Language Models

Reading 7.1 Provost, F. and T. Fawcett (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media Inc., 1st Edition. Chapter 10.

Keywords

Linguistic structure (p. 250)
Dirty (p. 250)
Document (p. 251)
Corpus (p. 251)
Tokens (p. 251)
Terms (p. 251)
Bag of words (p. 252)
Term frequency (p. 252)
Inverse document frequency (p. 254)
TFIDF (p. 256)
N-grams (p. 263)
Bi-grams (p. 263)
Named entity extraction (p. 264)
Topic models (p. 264)
Latent information model (p. 266)
Information triage (p. 274)

Learning Objectives

Demonstrate proficiency in the following areas:

7.1.1 Broad Issues Involved in Mining Text

For example:

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

7.1.2 Text Representation

For example:

A. Describe the meaning of "terms" (or "tokens") when used in information retrieval.
B. List the steps used in converting a document to a term frequency representation.
C. Calculate term frequency (TF), inverse document frequency (IDF), and term frequency inverse document frequency (TFIDF).
D. Describe the treatment for rare and common words when deciding the weight of a term.
E. Identify the general shape of IDF when plotted against the number of documents containing the term.
F. Describe the relationship between a corpus and IDF.
G. Describe the relationship between a document and TFIDF.
H. List the drawbacks of the "bag of words" approach.
I. Calculate IDF using the probability of a term in a set of documents.
J. Calculate the entropy of a term using IDF.

7.1.3 Additional Text Representation Approaches Beyond "Bag of Words"

For example:

A. Explain the term "bag of n-grams up to three."

B. Describe when n-gram sequences would be more useful than their component words.

C. List the main disadvantage of n-gram sequences.

D. Describe key requirements for using the named entity extraction.

E. Contrast topic models with the "bag of words" approach.

F. Describe the process used to learn about topics in topic models.

G. Compare the topic model to the latent information model.

7.1.4 Mining News Stories to Predict Stock Price Movement

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 Hull, J. C. (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 9.

Keywords

Sentiment analysis (p. 196)
Web scraping (p. 197)
Tokenization (p. 199)
Stop words (p. 199)
Stemming (p. 199)
Lemmatization (p. 200)
Laplace smoothing (p. 205)
Word vectors (p. 209)
Word embedding (p. 209)

Learning Objectives

Demonstrate proficiency in the following areas:

7.2.1 Natural Language Processing (NLP)

For example:

A. List the reasons that make it difficult to develop NLP applications.

B. List applications of NLP.

C. Explain why one should not rush into developing a trading strategy based on NLP.

D. Describe the best approach to creating labeled data for NLP.

E. Describe the steps used in tokenization.

F. Describe a common approach to creating a list of stop words.
G. Contrast stemming from lemmatization.
H. Describe the treatment for rarely occurring words and abbreviations during pre-processing.
I. Describe how a bag-of-words approach can convert a sentence to a numerical array.
J. Identify the drawbacks of the bag-of-words approach.
K. Calculate the number of n-grams that can be created from a sentence.
L. Discuss the key assumption made in using the Naïve Bayes classifier.
M. Calculate the conditional probability of a document having a particular sentiment.
N. Explain the key drawback of the Naïve Bayes classifier.
O. Calculate the conditional probability of a document having a particular label using Laplace smoothing.
P. List the advantages of the logistic regression and SVM over the Naïve Bayes classifier
Q. List applications of word sequences.
R. List some of the algorithms used in translating from one language to another.

Reading 7.3 Wolfram, S. (2023). What is ChatGPT Doing … and Why Does It Work? Wolfram Media, Inc., 1st Edition. Pages 1-10 and 41-74 from the book are to be used for the FDP Exam.

Keywords

Temperature (p. 2) *Embeddings (p. 41)*

Learning Objectives

Demonstrate proficiency in the following areas:

7.3.1 It's Just Adding One Word at a Time

For example:

A. Explain the meaning of "reasonable" in the context of what ChatGPT is trying to do.
B. Describe the end results produced by ChatGPT.
C. Explain how ChatGPT incorporates creativity in its selection of words.
D. Explain what happens to the output of ChatGPT in a "zero temperature" environment.

7.3.2 Where Do the Probabilities Come From?

For example:

A. Interpret a "2-gram" plot for typical English text.
B. Explain what happens to the generated words when "2-gram" probabilities are used instead of probabilities of single words.

C. Describe how a meaningful sequence of words can be produced by using probabilities for pairs or longer n-grams of words.

D. Describe the key problem in using n-grams to generate meaningful text.

E. Describe the role of "large language models" (LLM) in ChatGPT.

7.3.3 The Concept of Embeddings

For example:

A. Explain what happens to the 2D projection of actual embeddings used in ChatGPT.

B. Describe the characteristics for embeddings of images for similar items.

C. Describe what can be used as embedding for images of digits.

D. Explain the difference between the "signature" of a digit and that of another digit.

E. Describe the task of "word prediction".

F. Explain how neural networks can be used for "word prediction".

G. List the common characteristics of different systems, such as word2vec, GloVe, BERT, and ChatGPT.

H. Describe the advantage of using "tokens" rather than words.

7.3.4 Inside ChatGPT

For example:

A. Describe the advantage of a convolutional neural network over a fully connected neural network for dealing with images.

B. List the three basic stages of ChatGPT.

C. Describe the role of two pathways in the embedding module of GPT-2.

D. Describe the role of "attention heads" for a transformer.

E. Describe the role of "attention blocks" for a transformer.

F. Explain the net effect of the transformer in ChatGPT.

G. Describe the key difference between ChatGPT and a typical computational system, such as a Turing machine.

H. Explain the reason why it takes a while to generate a long piece of text with ChatGPT.

7.3.5 The Training of ChatGPT

For example:

A. Describe the process of determining the weights for the neurons in ChatGPT.

B. Explain the effect of "back propagating" from the error on the weights of a neural net.

C. Describe the relationship between the "size of the network" and the "size of the training data".

7.3.6 Beyond Basic Training

For example:

A. List the steps required for training ChatGPT after it has been fed with large amounts of existing text.

B. Describe the type of rules the neural network in ChatGPT can learn easily.

7.3.7 What Really Lets ChatGPT Work?

For example:

A. Discuss why ChatGPT can capture the essence of human language.

B. Describe what is discovered during the training of ChatGPT.

7.3.8 Meaning Space and Semantic Laws of Motion

For example:

A. Describe the key feature of any linguistic feature space.

B. Explain how different meanings of the same word can be teased out from the linguistic feature space.

C. Describe the trajectory a sequence of words can take inside ChatGPT.

7.3.9 Semantic Grammar and the Power of Computational Language

For example:

A. Describe the important scientific fact revealed by ChatGPT.

B. Compare and contrast human language to computational language.

Reading 7.4 Lo, A.W. and M. Singh. (2023). From ELIZA to ChatGPT: The Evolution of Natural Language Processing and Financial Applications. The Journal of Portfolio Management, 49 (7) 201-235. Pages 201-216 are to be used for the FDP Exam.

Keywords

Parse trees (p. 203)
Hidden Markov model (HMM) (p. 205)
Maximum entropy model (MEM) (p. 205)
Lexicon-based model (p. 206)
Topic model (p. 206)
Count-based method (p. 206)
Predictive methods (p. 206)
Self-attention mechanism (p. 208)
Impact investing (p. 211)

Learning Objectives

Demonstrate proficiency in the following areas:

7.4.1 Evolution of NLP Models

For example:

A. Describe how grammar-based parsers and dictionaries can be used to analyze the structure of a text.

B. List the limitations of early chatbots, such as ELIZA.
C. List the first statistical NLP models used for NLP tasks.
D. List the limitations of HMMs and MEMs.
E. List the limitations of Lexicon-based models and topic models.
F. List the methods that can be applied to count-based methods to obtain word embeddings.
G. List the tasks that can be accomplished using word embeddings.
H. List the disadvantages of using embedding-based methods.
I. Describe the requirements for using recurrent neural networks and convolutional neural networks.
J. Describe the role of attention mechanisms, pretraining, and fine-tuning in ChatGPT.
K. List examples of AI applications for ChatGPT.
L. Describe the limitations of ChatGPT.

7.4.2 Applications

For example:

A. Identify the goal of financial risk management.
B. Explain how NLP can be used in regulatory compliance, credit risk, and fraud detection.
C. Describe the ways NLP models can be used in impact investing.
D. Describe the major concern regarding the use of deep learning-based language models for impact investing.
E. List the tasks in asset management that can use NLP.
F. Describe how sentiment analysis, news analysis involving NLP, and earnings call analysis using NLP are used in asset management.

7.4.3 What's Next

For example:

A. List the limitations for the use of NLP in Finance.
B. Explain how each of the limitations affects the use of NLP in Finance.

Reading 7.5 Jingwen, J., B. Kelly, and D. Xiu (2022). Expected Returns and Large Language Models. Available at SSRN: https://ssrn.com/abstract=4416687. Pages 5-14 are to be used for the FDP Exam.

Keywords

Self-Supervised learning (p. 7)
Transfer learning (p. 7)
Next token prediction or autoregressive language modeling (p. 10)
Feature extraction approach or probing (p. 10)

Learning Objectives
Demonstrate proficiency in the following areas:

7.5.1 A Tale of Two Objectives

For example:

A. List the two objectives of the paper.
B. Describe how sentiment analysis is treated as a classification problem.
C. Explain the relationship between the quantitative value of an article and its sentiment.
D. Describe the requirements for the training sample needed for sentiment labeling.
E. Explain the reason for creating a sentiment label based on three-day returns.
F. Describe the process of using sentiment analysis for predicting returns.

7.5.2 Large Language Models

For example:

A. List the key characteristics of Bidirectional Encoder Representations from Transformers (BERT).
B. List examples of tokens.
C. Describe a feature of Large Language Models (LLMs) that helps alleviate data sparsity.
D. Apply the BERT and BPE tokenizers to a given string.
E. List the key features of the transformer approach used in LLMs.
F. List the advantages of using the transformer approach used in LLMs.
G. Describe the features of BERT, RoBERTa, and OPT that allow them to be used for specific applications.
H. List where BERT, RoBERTa, or OPT can be used.
I. List the main features of a language learned during the pre-training of an LLM.
J. Describe the adaptation stage of an LLM.
K. Explain the advantages of using the feature extraction approach.
L. Explain the process of creating article-level representation from tokens.

7.5.3 Word Embeddings

For example:

A. Describe the key limitations of the Word2Vec model.

7.5.4 Why not ChatGPT?

For example:

A. List the drawbacks of using ChatGPT as a sentiment analyzer.

Topic 8. Ethics, Privacy, and Regulation

Reading 8.1 Hull, J. C. (2021). Machine Learning in Business: An Introduction to the World of Data Science. Independently Published by GFS Press, 3rd Edition. Chapter 11.

Keywords

Global Data Protection Regulation (GDPR) (p. 230)
Trolley problem (p. 232)
Adversarial machine learning (p. 233)
Spoofing (p. 233)
Four industrial revolutions (p. 235)

Learning Objectives

Demonstrate proficiency in the following areas:

8.1.1 Data Privacy

For example:

A. Discuss the Global Data Protection Regulation (GDPR) and list its requirements.

B. List the consequences of violating the GDPR.

8.1.2 Biases

For example:

A. Discuss biases, including representativeness and data availability.

B. Discuss how biases can arise from cleaning data, which models are used, and how models are interpreted.

C. Discuss what constitutes informed consent.

8.1.3 Ethics

For example:

A. Discuss whether machine learning models and their applications, such as warfare, can be ethical or unethical.

B. Explain the trolley problem and how it applies to algorithms used for driverless cars.

C. Explain Microsoft's "Thinking About You" and how decisions in the model building can lead to unexpected results.

8.1.4 Transparency

For example:

A. Discuss the importance of making machine learning algorithms transparent.

8.1.5 Adversarial Machine Learning

For example:

A. List an example of adversarial machine learning.

B. List approaches to limiting adversarial machine learning.

8.1.6 Legal Issues

For example:

A. List the potential legal liabilities of algorithms, including ownership and use of data, biased algorithms, and assignment of liability for actions of autonomous systems.

8.1.7 Man vs. Machine

For example:

A. Discuss the four industrial revolutions, including concerns and benefits, and the implications for job markets.

B. Discuss the skill of monitoring machine learning algorithms.

Reading 8.2 Smith, G., and I. Rustagi (2020). Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook. Berkeley Haas Center for Equity, Gender, and Leadership.

Keywords

Biased AI (p. 20)
Fairness (p. 20)
Proxy (p. 32)
Audit (p. 33)
Diversity (p. 40)
White box model (p. 42)

Learning Objectives

Demonstrate proficiency in the following areas:

8.2.1 The Bias in AI Map (p. 21-39)

For example:

A. Identify the source of bias in various examples illustrating the use of biased datasets.

B. Identify the source of bias in various examples illustrating the use of biased algorithms.

C. Identify the source of bias occurring in the context, alteration, or interpretation of an AI system.

8.2.2 Challenges (p. 40-43)

For example:

A. Describe challenges to mitigating bias at the organizational level.

B. Describe challenges to mitigating bias at the industry level.

C. Describe challenges to mitigating bias at the societal level.

8.2.3 Executing Strategic Plays (p. 47-50)

For example:

A. Describe how actions in the Teams playbook bucket can overcome the challenges of mitigating bias.

B. Describe how actions in the AI Models playbook bucket can overcome the challenges of mitigating bias.

C. Describe how actions in the Corporate Governance & Leadership playbook bucket can overcome the challenges of mitigating bias.

Reading 8.3 FinRegLab (2021). The Use of Machine Learning for Credit Underwriting: Market & Data Science Context.

Keywords

Inherently interpretable model (p. 36)
Monotonicity (p. 40)
Regularization (p. 40, 95)
Shapley (SHAP) additive explanation (p. 43)
Local interpretable model-agnostic explanations (LIME) (p. 41)
Integrated gradients (p. 44)
Partial dependence plots (p. 45)
Individual conditional expectation (ICE) plots (p. 46)
Accumulated local effects (ALE) plots (p. 47)
Counterfactual explanations (p. 49)
Adversarial examples (p. 49)
Auto ML (p. 52)
Weight of evidence (WoE) method (p. 70)
Reject inference (p. 70)
Simple augmentation (p. 71)
Fuzzy parceling (p. 71)
Credit scorecards (p. 71)
Representation bias (p. 76)
Historical bias (p. 76)
Omitted variable bias (p. 76)
Selection bias (p. 76)
Aggregation bias (p. 78)
Measurement bias (p. 78)
Statistical parity or demographic parity (p. 83)
Conditional statistical parity (p. 84)
Predictive parity (p. 85)
Equalized odds (p. 85)
Counterfactual fairness (p. 86)
Calibration (p. 87)
Fairness through unawareness (p. 88)
Fairness through awareness (p. 89)
Adversarial debiasing (p. 95)

Learning Objectives

Demonstrate proficiency in the following areas:

8.3.1 Model Transparency (p. 34 to 38)

For example:

A. Discuss the debate about interpretable and explainable machine learning.

B. Identify an algorithm or scorecard as inherently interpretable, moderately interpretable, or uninterpretable.

8.3.2 Options for Enabling Transparency (p. 38 to 52)

For example:

A. List common constraints that can be used during model selection and training to produce more interpretable models.

B. Explain how different post hoc methods, such as surrogate models, feature importance explainability methods, and example-based explainability techniques can be used to provide additional information about how a model arrived at its predictions.

C. Describe the strengths and weaknesses of different post hoc explainability methods, such as LIME, SHAP, and counterfactual explanations.

D. List sources of explanation errors.

8.3.3 Emerging Model Diagnostic Tools (p. 52 to 53)

For example:

A. Explain how emerging model diagnostic tools can be used to identify potential sources of bias or discrimination in machine learning models.

B. List three diagnostic tool capabilities.

8.3.4 Data Selection and Preparation (p. 65 to 70)

For example:

A. Explain the importance of data selection and preparation in developing machine learning models for credit underwriting.

B. Describe the different types and forms of data available for selection by credit model developers.

C. Describe the impact choices made for dealing with missing values and during coarse classing may have on fairness, model performance, and reliability of post hoc explainability techniques.

8.3.5 Reject Inference and Credit Scorecards (p. 70 to 72)

For example:

A. Explain the concept of "reject inference."

B. Describe two reject inference methods for reducing bias.

C. Describe the benefits of using credit scorecards.

8.3.6 Fairness and Bias and Sources of Bias (p. 73 to 79)

For example:

A. Identify the stage of the data, algorithm, and user interaction feedback loop where particular biases can be introduced into a model.

B. Compare and contrast the different sources of bias in machine learning models used for credit underwriting.

C. Explain how model choices regarding training rates, model pruning, and privacy-enhancing technologies can produce biases.

8.3.7 Measuring Fairness (p. 79 to 90)

For example:

A. Explain the concept of the "impossibility theorem of fairness" and its implications for using fairness metrics.

B. Calculate the eight fairness metrics presented, given the required model inputs, model outputs, protected class features, and actual outcomes.

C. Discuss the advantages and disadvantages of each of the eight statistical and similarity-based fairness measures presented.

8.3.8 Model Debiasing and Debiasing Approaches (p. 90 to 96)

For example:

A. Explain the three main approaches to the pre-processing debiasing methods.

B. Explain the two main approaches to the in-processing debiasing methods.

C. Explain why post-processing debiasing methods are often not relevant to financial services like credit underwriting.

Reading 8.4 Recommendations for Regulating AI. Google (2023).

Keywords

Model cards (p. 10)

Learning Objectives

Demonstrate proficiency in the following areas:

8.4.1 General Approaches to Regulating AI

For example:

A. Explain the benefits of taking a sectoral approach to regulating AI.

B. Discuss the importance of adopting a proportionate, risk-based framework for regulating AI.

C. Explain why promoting an interoperable approach is important for ensuring the responsible development and use of AI.

D. Discuss strategies for addressing challenges of ensuring parity in expectations between non-AI and AI systems.

E. Discuss the role of transparency in regulating AI.

8.4.2 Practical Implications of Regulating AI

For example:

A. Discuss how organizations can clarify expectations for conducting risk assessments.

B. Describe three general principles for setting disclosure standards.

C. Explain why workable standards for explainability and reproducibility require compromise.

D. Explain why auditing should center on processes.

E. Discuss strategies for defining appropriate fairness benchmarks for AI systems.

F. Explain why mandating techniques for testing AI systems for robustness may undermine longer-term robustness.

G. Discuss the benefits and limitations of human oversight for verifying and monitoring AI systems.

Topic 9. Fintech Applications

Reading 9.1 Ekster, G. and Kolm, P. N. (2021). Alternative Data in Investment Management: Usage, Challenges, and Valuation. The Journal of Financial Data Science, 3(4): 10-32.

Keywords

Alternative data (Alt-data) (p. 2)
Originators (p. 3)
Intermediaries (p. 3)
Data curators (p. 4)
Alpha decay (p. 4)
Entity mapping (p. 5)
Ticker tagging (p. 5)
Panel (p. 6)
Unbalanced panel (p. 6)
Balanced panel (p. 6)
Panel stabilization (p. 6)
Debiasing (p. 7)
Golden triangle event study methodology (p. 8)
Public information test (p. 8)
Market reaction test (p. 8)
Report card (p. 9)
Leave-one-out (LOO) cross-validation (p. 14)

Learning Objectives

Demonstrate proficiency in the following areas:

9.1.1 Background

For example:

A. Discuss properties of alt-data.

B. List examples of alt-datasets.

9.1.2 The Alternative Data Ecosystem

For example:

A. List and discuss the constituents in the alt-data ecosystem, including originators, intermediaries, and investment professionals.

B. Identify sources of raw data.

C. List the intermediary dynamics that should be kept in mind by the buyers of alt-data.

D. Discuss the misalignment of incentives created between data intermediaries and buy-side clients.

E. Explain alpha decay and the types of data that have less alpha potential.

F. Compare the use of alt-data in fundamental funds vs. quantitative funds.

G. List the drawbacks of alt-datasets for using them in quantitative funds.

9.1.3 Challenges With Alternative Data

For example:

A. Explain entity mapping, ticker tagging, panel stabilization, and debiasing.

B. List the desirable properties of a practical entity mapping solution.

C. Describe the advantages and disadvantages of imputation of missing data.

D. Explain the difficulty in identifying bias in an alt-dataset.

9.1.4 The Value of Alternative Data

For example:

A. List the purpose of alt-data valuation.

B. List the two fundamental methods of evaluating alt-datasets.

C. Describe the method that can be used to measure the impact of a factor constructed from an alt-dataset.

D. List and discuss the three steps of the golden triangle event study methodology.

E. List the factors that can determine the value of an alt-dataset to a buy-side fund manager.

F. Explain how report cards can be used to determine the value of an alternative dataset.

G. Explain the relationship between a dataset's structure and investment performance.

H. Describe how a raw and unstructured dataset can differ from a corresponding aggregated and structured dataset.

I. List the two approaches that can be used as a trade-off between the cost of analyzing and exploring alt-data and the uniqueness of any investment insights.

9.1.5 Issues in Processing Data

For example:

A. Explain outlier detection and resolution, and imputation error estimation.

9.1.6 Trends in the Alternative Data Space

For example:

A. Discuss the cost-benefit analysis of intermediaries vs. originators.

Reading 9.2 Bagattini, G., Z. Benetti, and C. Guagliano (2023). Artificial Intelligence in EU Securities Markets. European Securities and Markets Authority (ESMA) Report on Trends, Risks and Vulnerabilities Risk Analysis.

Keywords

Robo-advisors (p. 9) *Explainability (p. 12)*

Learning Objectives

Demonstrate proficiency in the following areas:

9.2.1 Asset Management

For example:

A. Discuss how AI techniques can be used by discretionary and systematic fund managers.

B. List the reasons that are driving AI adoption by funds.

C. Describe the distinctive trait of AI that allows for using it to extract economically meaningful information.

D. List the characteristics of financial data that make them difficult to use in ML applications.

E. Describe the key reasons that may deter investors from investing in funds using AI tools.

F. Explain the reasons for a fund avoiding references to AI in marketing the fund.

G. List the key reasons for AI not being an integral part of robo-advisors.

9.2.2 Trading

For example:

A. Discuss how AI models can be used before a trade is executed, to reduce market impact of a trade, and for securities lending transactions.

B. Explain why market impact costs are hard to estimate.

C. Describe how ML methods can be used in post-trade processing.

9.2.3 Other Entities

For example:

A. List applications of AI in the credit rating agencies.

B. Explain why AI tools are not currently being used for credit rating assessment.

C. Describe the challenges expected in the future to increase use of AI tools in the credit rating industry.

D. Discuss the different ways proxy advisory firms are using AI.

9.2.4 Potential Risks

For example:

A. List the risks associated with using AI in the financial system.

B. Explain the effect of each risk associated with using AI in the financial system.

Reading 9.3 Harvey, C. R., Y. Liu, and A. Saretto. An Evaluation of Alternative Multiple Testing Methods for Finance Applications. October 2019 (Updated Feb 2, 2020). https://ssrn.com/abstract=3480087**. Pages 1-20 are to be used for the FDP Exam.**

Keywords

Multiple testing (p. 1)
p-value (p. 1)
False positive (p. 2)
True positive (p. 2)
True Negative (p. 2)
False Negative (p. 2)
False discovery proportion (FDP) (p. 2)
False nondiscovery proportion (FNDP) (p. 2)
False Omission Rate (p. 2)
Family-wise error rate (FWER) (p. 3)
Bonferroni method (p. 4)
Holm method (p. 4)
bootstrap reality check (BRC) (p. 5)
StepM (p. 7)
Stepwise model selection (p. 7)
k-FWER (p. 9)
False discovery rate (FDR) (p. 11)
Benjamini and Hochberg (BH) (p. 13)
Benjamini and Yekutieli (BY) (p. 13)

Learning Objectives
Demonstrate proficiency in the following areas:

9.3.1 False discovery

For example:

A. Discuss why the traditional use of p-values is no longer valid when multiple testing exists.

B. Explain and calculate the False Positive Proportion (FDP).

C. Explain and calculate False Nondiscovery Proportion (FNDP) or False Omission Rate.

D. Explain the relationship between the number of multiple tests and the number of False Positives (FP).

9.3.2 Multiple hypothesis method (MHT)

For example:

A. Explain the concept of the family-wise error rate (FWER).

B. Explain what happens to FWER as multiple tests increase or decrease.

C. Compare the differences between FWER and k-FWER.

D. Calculate and apply the Bonferroni method.

E. Calculate and apply the Holm method.

F. Describe the basic ideas behind the bootstrap reality check (BRC) method.

G. Compare the differences between the Bonferroni method and the BRC method.

H. Describe the concept of controlling the false discovery proportion (FDP).

I. Discuss the basic ideas behind the StepM method and how it compares to the BRC method.

J. Describe the basic ideas behind the stepwise model selection (SMS) method.

K. Describe the basic ideas behind the false discovery proportion (FDP) and how it relates to FWER.

L. Describe the basic ideas behind controlling the false discovery rate (FDR) and how it relates to the FDP.

M. Calculate, and apply the Benjamini and Hochberg (BH) method for controlling the FDR.

N. Recognize and apply the Benjamini and Yekutieli (BY) for controlling the FDR.

9.3.3 Finance applications of multiple hypothesis method

For example:

A. Explain how having an unbalanced data set of historical returns on mutual funds could impact the results of studies of mutual funds' performance.

B. Explain why studies of long-short trading strategies involving factors are less likely to suffer from the lack of power than studies of mutual funds' performance.

C. Describe the three key challenges researchers face when applying MHT to financial problems.

Reading 9.4 Francis, L. A. (2006). Taming Text: An Introduction to Text Mining. Casualty Actuarial Society Forum, 51-88.

Keywords

Structured data (p. 51)
Unstructured data (p. 52)
Text mining (p. 52)
Term extraction (p. 55)
Feature creation (p. 55)
Parsing (p. 56)
Sparse (p. 58)
Cluster analysis (p. 60)
Dimensions (p. 60)
Factor analysis (p. 61)
Simple matching dissimilarity measure (p. 65)
Rogers and Tanimoto dissimilarity measure (p. 65)
Proximity matrix (p. 68)
Main effects model (p. 79)
ANOVA (p. 80)

Learning Objectives

Demonstrate proficiency in the following areas:

9.4.1 Introduction

For example:

A. Explain the reasons for not using unstructured data in analysis.

9.4.2 Research Context

For example:

A. List examples of text mining.

9.4.3 Background and Methods

For example:

A. Describe the process of creating indicator variables from text description using parsing.

B. Explain how factor analysis can be used to reduce dimensions.

C. Describe the objective of the K-means clustering algorithm when it is applied to text mining.

D. Recognize simple matching dissimilarity measure and Rogers and Tanimoto dissimilarity measure.

E. Explain when tf-idf is most appropriate to use.

F. Explain the equivalence between a cluster's frequency for a term and the proportion of records containing the term in that cluster when K-means clustering is used.

G. Describe the effect of having too many or too few clusters in a K-means clustering algorithm.

H. Describe the most common way to implement hierarchical clustering.
I. Explain how proximity matrix can be used to form clusters in hierarchical clustering.
J. Describe the insights offered by hierarchical clustering.
K. List the different ways of determining the number of clusters.
L. Explain how stepwise regression can find the optimal number of clusters.
M. Explain how the BIC statistic can be used to find the optimal number of clusters.
N. Describe the outcome of cluster analysis performed on a single descriptive field.
O. Explain how crosstabulation can be used to find important words in defining a cluster.
P. Explain how the frequency of terms by cluster can be used to find words that are important in defining a cluster.
Q. Describe the logistic regression process to categorize observations into high and low-importance observations.
R. Describe the requirements for using the simple analysis of variance (ANOVA).

Reading 9.5 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)
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:

9.5.1 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 a machine learning application results.

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 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.

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

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

J. Describe the advantages and disadvantages of the 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 the probabilistic Sharpe ratio.

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

FDP EDITORIAL STAFF

Hossein Kazemi, Ph.D., CFA, Senior Advisor, CAIA Association
Satya Das, CFA, Senior Advisor, Curriculum and Exams, FDP Institute
Kim Durand, Project & General Operations Manager, FDP Institute
Kathryn Wilkens, Ph.D., CAIA, Consultant, FDP Institute

Design: Gabriele Villamena, Sichtwerk, Inc.

APPENDIX A: SUMMARY OF FORMULAS AND QUANTITATIVE CONCEPTS

1. Introduction to Data Science

1.2.1.J Calculate conditional probability using Bayes' Theorem.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$P(X)$ and $P(Y)$ are the unconditional probabilities of events X and Y, $P(Y|X)$ is the probability of event Y conditional on the occurrence of event X, and $P(X|Y)$ is the probability of event X conditional on the occurrence of event Y.

1.3.3.F Calculate the Bayes error rate.

$$Bayes\ error\ rate\ = 1 - E(\max_j P(Y = j|X))$$

$P(Y = j|X)$ is the probability of $Y = j$ conditional on X and $E()$ denotes the expectation operator.

1.3.3.I Calculate the conditional probability of a point belonging to a particular class.

$$P(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in \mathcal{N}_0} I(y_i = j)$$

K is a positive integer, x_0 is the test observation, $\mathcal{N}_0$ are the K points closest to x_0, Y is the response variable, j is the class of response variable of interest, and $I()$ is an indicator function equal to 1 when $y_i = j$ and 0 otherwise.

2. Linear and Logistic Regression, Support Vector Machines, Regularization, and Time Series

2.1.1.A Apply the equation of a straight-line using slope and intercept.

$$y = mx + b$$

m is the slope and b is the intercept of the straight line.

2.1.1.C Calculate the best value for the parameters of a linear discriminant for a set of instances.

Use the sample data points on different versions of the linear discriminant to find the classification for the sample data points. The version of the linear discriminant that provides classification consistent with the label of sample data points should be the best linear discriminant.

2.1.2.B Calculate odds and log odds.

$$Odds = \frac{p}{1-p}$$

$$log\text{-}odds = ln\left(\frac{p}{1-p}\right)$$

p is the probability of an event occurring and ln is the natural logarithm.

2.1.2.D Calculate the log-odds linear function.

$$ln\left(\frac{p_+(\boldsymbol{x})}{1-p_+(\boldsymbol{x})}\right) = f(\boldsymbol{x}) = w_o + w_1x_1 + w_2x_2 + \cdots$$

$p_+(\boldsymbol{x})$ represents the model's estimate of the probability of class membership of a data item represented by feature vector $\boldsymbol{x}$ and ln is the natural logarithm.

2.1.2.E Calculate class probability using the logistic function.

$$p_+(\boldsymbol{x}) = \frac{1}{1+e^{-f(x)}}$$

e is the base of the natural logarithm.

2.2.1.C Calculate the probability of a positive outcome using the sigmoid function.

$$p = \frac{1}{1 + e^{-a-\sum_{j=1}^{m} b_j X_j}}$$

a is the bias and b_j is the estimated weights for j^{th} factor.

2.2.1.D Recognize the cost function for the logistic regression.

$$Cost\ function = \frac{1}{n}\left[-\sum_{\substack{positive \\ outcomes}} \ln(Q) - \sum_{\substack{negative \\ outcomes}} \ln(1-Q)\right]$$

Q is the probability of a positive outcome and n is the number of observations.

2.2.2.C Calculate the dimension of a separating hyperplane.

$$Dimension\ of\ a\ separating\ hyperplane\ =\ m-1$$

m is the number of features.

2.2.2.D Recognize the equation of a separating hyperplane with m features.

$$\sum_{j=1}^{m} w_j x_j = b$$

m is the number of features, w_j is the weight for feature j, x_j is the value of feature j, and b is the bias.

2.2.2.F Recognize the objective function used in creating SVM with m features.

$$\min_{w} \sum_{j=1}^{m} w_j x_{ij}$$

$$subject\ to \sum_{j=1}^{m} w_j x_{ij} \geq b+1\ \ for\ positive\ outcomes\ and$$

$$\sum_{j=1}^{m} w_j x_{ij} \leq b-1\ \ for\ negative\ outcomes$$

2.2.2.G Recognize the objective function for a soft margin classification.

$$\min_{w} C\sum_{i=1}^{n} z_i + \sum_{j=1}^{m} w_j^2$$

$$subject\ to\ z_i = \max\left(b + 1 - \sum_{j=1}^{m} w_j x_{ij}\ , 0\right)\ for\ positive\ outcomes\ and$$

$$z_i = \max\left(\sum_{j=1}^{m} w_j x_{ij} - b + 1, 0\right)\ for\ negative\ outcomes$$

C is a hyperparameter.

2.2.2.L Recognize the Gaussian radial basis function (RBF) for an observation.

$$Radial\ basis\ function\ =\ e^{-\gamma D^2}$$

γ is the parameter that determines the rate at which the function declines as the distance from a landmark increase and D is the Euclidean distance from a landmark.

2.2.3.B Recognize the equations of hyperplanes in SVM regression.

$$y\ = \sum_{j=1}^{m} w_j x_j + b \pm e$$

$2e$ is the distance between the two hyperplanes forming the pathway, m is the number of features, and b is the bias.

2.2.3.C Recognize the objective function used in SVM regression.

$$C\sum_{i=1}^{n} z_i + \sum_{j=1}^{m} w_j^2$$

z_i is the vertical distance between a data point and the edge of the pathway, C is a constant that determines the degree of regularization, n is the sample size, and m is the number of features.

2.3.1.A Calculate the value of RSS.

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the coefficients of a simple linear regression, e_i represents the error associated with the i^{th} data point, and n is the number of data points.

2.3.1.B Calculate the least-squares coefficient estimates.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \frac{Covariance(x,y)}{Variance(x)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\bar{y}$ and $\bar{x}$ are the sample mean of y and x.

2.3.1.D Recognize the standard error of a statistic.

$$SE\left(\hat{\beta}_0\right)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$SE\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

σ^2 is the variance of error term and is estimated using $\frac{RSS}{n-2}$.

2.3.1.F Calculate the 95% confidence interval.

$$Approximate\ 95\%\ confidence\ interval\ for\ \beta_1 = \hat{\beta}_1 \pm 2 * SE(\hat{\beta}_1)$$

$$Approximate\ 95\%\ confidence\ interval\ for\ \beta_0 = \hat{\beta}_0 \pm 2 * SE(\hat{\beta}_0)$$

2.3.1.G Calculate the t-statistic.

$$t\text{-}statistic = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

2.3.1.J Calculate and interpret the R^2 statistic.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$TSS = \sum(y_i - \bar{y})^2$ is the total sum of squares and RSS is as defined before.

2.3.1.L Calculate the correlation from R^2 for the simple linear regression.

$$Correlation = \sqrt{R^2}$$

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

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

$TSS = \sum(y_i - \bar{y})^2$ and $RSS = \sum(y_i - \hat{y}_i)^2$, n is the number of sample data points, and p is the number of predictors.

2.3.2.H Calculate RSE given the values of RSS, n, and p.

$$RSE = \sqrt{\frac{RSS}{n - p - 1}}$$

$RSS = \sum(y_i - \hat{y}_i)^2$, n is the number of sample data points, and p is the number of predictors.

2.3.3.N Recognize the variance inflation factor.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

$R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all the other predictors.

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

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

$$AIC = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

$$BIC = \frac{1}{n}(RSS + ln(n)d\hat{\sigma}^2)$$

$\hat{\sigma}^2$ is an estimate of the variance of the error estimated using the full model containing all predictors, d is the number of predictors, n is the number of sample data points, and ln is the natural logarithm.

2.3.4.R Recognize the adjusted R².

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

$TSS = \sum(y_i - \bar{y})^2$ and $RSS = \sum(y_i - \hat{y}_i)^2$, n is the number of sample data points, and p is the number of predictors.

2.3.5.A Recognize the objective function of ridge regression.

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\} = \min_{\beta}\left\{RSS + \lambda\sum_{j=1}^{p}\beta_j^2\right\}$$

λ is the tuning parameter that determines the degree of regularization, $RSS = \sum(y_i - \hat{y}_i)^2$, n is the number of sample data points, and p is the number of predictors.

2.3.5.E Calculate the ℓ_2 norm.

$$\ell_2\ norm = \sqrt{\sum_{j=1}^{p}\beta_j^2}$$

2.3.6.C Recognize the objective function of Lasso.

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} = \min_{\beta}\left\{RSS + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

λ is the tuning parameter that determines the degree of regularization, $RSS = \sum(y_i - \hat{y}_i)^2$, n is the number of sample data points, and p is the number of predictors.

2.3.6.D Calculate the ℓ_1 norm.

$$\ell_1\ norm\ = \sum_{j=1}^{p} |\beta_j|$$

2.3.6.G Recognize the alternative formulation of the objective function for Lasso and ridge regression.

$$Lasso\colon \min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}\ subject\ to\ \sum_{j=1}^{p} |\beta_j| \le s$$

$$Ridge\colon \min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}\ subject\ to\ \sum_{j=1}^{p} \beta_j^2 \le s$$

2.4.1.J Calculate the value of EMWA for a series.

$$S_t = \lambda Y_t + (1-\lambda) S_{t-1} = \sum_{i=0}^{k} \lambda(1-\lambda)^i Y_{t-i}\ + (1-\lambda)^{k+1} S_{t-(k+1)} \approx \sum_{i=0}^{k} \lambda(1-\lambda)^i Y_{t-i}$$

λ is the weighting parameter.

2.4.2.D Calculate the mean, variance, autocovariance, and autocorrelation of a stationary AR(1) process.

$$AR(1)\ process\colon Y_t = \alpha + \phi_1 Y_{t-1} + \varepsilon_t$$

$$Mean, E[Y_t] = \frac{\alpha}{(1-\phi_1)}$$

$$Variance, \sigma_Y^2 = \frac{\sigma_\varepsilon^2}{1-\phi_1^2}$$

$$Covariance[Y_t, Y_{t-1}] = \phi_1 \sigma_Y^2$$

$$Correlation[Y_t, Y_{t-k}] = \phi_1^k$$

σ_ε^2 is the variance of ε_t.

2.4.2.F Calculate the value of an AR(1) process.

$$Y_t = \alpha + \phi_1 Y_{t-1} + \varepsilon_t$$

2.4.2.H Calculate the mean and variance of a random walk.

$$Random\ walk: Y_t = \alpha + Y_{t-1} + \varepsilon_t$$

$$Mean, E[Y_t] = t \times \alpha$$

$$Variance, \sigma_Y^2 = t \times \sigma_\varepsilon^2$$

σ_ε^2 is the variance of ε_t.

2.4.2.K Calculate the conditional forecast of mean and variance for a stationary AR(1) model.

$$AR(1)\ model: Y_{t+1} = \alpha + \phi_1 Y_t + \varepsilon_{t+1}$$

$$Conditional\ mean, E[Y_{t+1}|Y_t] = \alpha + \phi_1 Y_t$$

$$Conditional\ variance, Var[Y_{t+1}|Y_t] = \sigma_\varepsilon^2$$

2.4.3.B Recognize the unconditional mean, variance, and autocovariance of a moving average model.

$$Moving\ average\ model: Y_t = \mu + \psi_0 \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_1 \varepsilon_{t-2} + \cdots + \psi_q \varepsilon_{t-q}$$

$$Mean, E[Y_t] = \mu$$

$$Variance, \sigma_Y^2 = \sigma_\varepsilon^2 \sum_{i=0}^{q} \psi_i^2$$

$$Covariance[Y_t, Y_{t+k}] = \begin{cases} \sigma_\varepsilon^2 \sum_{j=0}^{q-k} \psi_j \psi_{j+k} & if\ k = 0, 1, \cdots, q \\ 0 & k > q \end{cases}$$

2.4.3.C Calculate the conditional value of the mean of forecast for an MA(1) model.

$$MA(1)\ model: Y_t = \mu + \psi_0 \varepsilon_t + \psi_1 \varepsilon_{t-1}$$

Conditional mean for the one-step ahead forecast, $\hat{Y}_{t+1} = E[Y_{t+1}|Y_t] = \mu + \psi_1 \varepsilon_t$

Conditional mean for the two-step ahead forecast, $\hat{Y}_{t+2} = E[Y_{t+2}|Y_t] = \mu$

2.4.3.D Recognize the conditional value of variance of forecast error of an MA(1) model.

$$MA(1)\ model: Y_t = \mu + \psi_0\varepsilon_t + \psi_1\varepsilon_{t-1}$$

The variance of the one-step ahead forecast error is $\psi_0^2\sigma_\varepsilon^2$

The variance of the one-step ahead forecast error is $(\psi_0^2 + \psi_1^2)\sigma_\varepsilon^2$

2.4.3.E Recognize the unconditional mean of an ARMA(p, q) model.

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots \phi_2 Y_{t-p} + \psi_1 Y_{t-1} + \psi_2 Y_{t-2} + \cdots + \psi_2 Y_{t-q}$$

$$Unconditional\ mean, E[Y_t] = \mu = \frac{\alpha}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}$$

2.4.4.F Calculate the conditional and unconditional variance for the error term when an ARCH(1) model is used.

Consider the following ARCH(1) process.

$$Y_t = Y_{t-1} + \varepsilon_t$$

$$\varepsilon_t = \sigma_t\mu_t \text{ where } \mu_t \sim N(0,1)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2$$

$$Conditional\ variance, Var[\varepsilon_t|\varepsilon_{t-1}] = \sigma_t^2$$

$$Unconditional\ variance,\ \sigma_\varepsilon^2 = \frac{\alpha_0}{1 - \alpha_1}$$

2.4.4.J. Calculate the long-term mean of volatility for a GARCH(1,1) model.

Consider the following GARCH(1, 1) process.

$$Y_t = \phi_1 Y_{t-1} + \varepsilon_t$$

$$\varepsilon_t = \sigma_t\mu_t \text{ where } \mu_t \sim N(0,1)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \beta_1\sigma_{t-1}^2$$

$$Long\text{-}term\ mean\ of\ volatility, E[\sigma_t^2] = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}$$

2.4.4.M. Calculate the forecasted value of volatility using a GARCH(1, 1) model.

$$\sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \beta_1\sigma_{t-1}^2$$

3. Decision Trees, Supervised Segmentation, and Ensemble Methods

3.1.2. B Calculate the value of entropy.

$$entropy = -p_1\, ln(p_1) - p_2\, ln(p_2) - \cdots$$

Each p_i is the probability of property i within the set andln represents the natural logarithm.

3.1.2.E Calculate information gain for children sets from a parent set.

$$Information\ gain\ (parent, children) = entropy(parent) - [p(c_1) * entropy(c_1) + p(c_2) * entropy(c_2) + \ldots]$$

$entropy(c_i)$ is the entropy of child i and $p(c_i)$ is the probability of an instance belonging to child i

3.1.3.C Calculate the probability at each node of a decision tree.

$$p(c) = \frac{n}{n+m}$$

$p(c)$ is the probability of a binary class, n is the number of examples in a leaf belonging to class c, and m is the number of examples in a leaf not belonging to class c.

3.1.3.E Calculate the value of the Laplace correction.

$$p(c) = \frac{n+1}{n+m+2}$$

$p(c)$ is the probability of a binary class with Laplace correction, n is the number of examples in a leaf belonging to class c, and m is the number of examples in a leaf not belonging to class c.

3.2.2.B Calculate conditional probabilities using Bayes' formula.

Please refer to the formula in *1.2.1.J*.

3.3.1.C Recognize and interpret RSS for a given partition (box).

$$RSS_J = \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

$\hat{y}_{R_j}$ is the mean response for the observations within the j^{th} partition and y_i is the target value of observation i in the j^{th} partition.

3.3.1.D Recognize RSS to perform recursive binary splitting.

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

$\hat{y}_{R_j}$ is the mean response for the observations within the j^{th} partition, y_i is the target value of observation i in the j^{th} partition, and J is the total number of partitions.

3.3.1.H Calculate the Gini Index.

$$Gini\ index = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) = 1 - \sum_{k=1}^{K} \hat{p}_{mk}^2$$

$\hat{p}_{mk}$ is the proportion of training observations in the m^{th} region that are from the k^{th} class and K is the total number of classes.

4. Classification, Clustering, and Naïve Bayes

4.1.1.A Calculate the Euclidean distance.

$$Euclidean\ distance\ = \sqrt{(d_{1,A} - d_{1,B})^2\ + (d_{2,A} - d_{2,B})^2 + \cdots +\ (d_{n,A} - d_{n,B})^2}$$

$(d_{1,A}, d_{2,A}, \ldots, d_{n,A})$ are the features of object A and $(d_{1,B}, d_{2,B}, \ldots, d_{n,B})$ are the features of object B.

4.1.1.C Calculate the probability of belonging to a class based on the nearest neighbor classification.

$$Probability\ of\ class\ = \frac{n}{m}$$

n is the number of target instances and m is the total number of instances.

4.1.1.E Calculate contributions and class probabilities using weighted voting.

$$Contribution\ = \frac{Similarity\ weight\ of\ an\ instance}{Sum\ of\ similarity\ weights}$$

$$Class\ probability\ =\ Sum\ of\ contribitions\ from\ a\ particular\ type\ of\ class$$

4.1.2.A Calculate joint probability for independent and dependent events.

$$Joint\ probability\ of\ independent\ events, P(AB)\ =\ P(A) * P(B)$$

$$Joint\ probability\ using\ conditional\ probability, P(AB)\ =\ P(A) * P(B|A)$$

4.1.2.C Calculate posterior probability, prior, and likelihood.

$$p(C = c|E) = \frac{P(E|C = c) * p(C = c)}{P(E)}$$

C is the target variable, c is the class of interest, and E is the evidence. $p(C = c)$ is the prior, $p(C = c|E)$ is the posterior probability, $P(E)$ is the likelihood of evidence E, and $P(E|C = c)$ is the likelihood of seeing the evidence E given class $C = c$.

4.2.1A Calculate and interpret feature scaling using Z-score and mini-max.

$$Z\text{-}score\ scaling\ = \frac{V - \mu}{\sigma}$$

V is the feature value for an observation and μ and σ are the mean and standard deviation of the feature.

$$Min\text{-}Max\ scaling\ = \frac{V - min}{max - min}$$

V is the feature value for an observation and max and min are the maximum and minimum values of a feature.

4.2.1.C Calculate and interpret the centroid of a cluster.

$$Centroid\ coordinate\ for\ a\ feature\ = \frac{\sum_{i=1}^{n} f_i}{n}$$

f_i is the value of a feature for observation i and n is the number of observations in the cluster.

4.2.1.E Calculate and interpret inertia as a measure of the clustering algorithm.

$$Inertia = \sum_{i=1}^{N} d_i^2$$

d_i is the distance of observation i from the center of the cluster to which it belongs and N is the number of observations.

5. Neural Networks and Reinforcement Learning

5.1.1.C Calculate the value of a sigmoid function from weights and bias.

$$Value\ of\ a\ sigmoid\ activation\ function\ = \frac{1}{1 + e^{-a-\sum_{i=1}^{n} w_i x_i}}$$

a is the bias, w_i is the weight of feature i, and x_i is the value of feature i. n is the number of input features to a node and e is the base of natural logarithm.

5.1.1.F Recognize the number of parameters to be estimated for an ANN with a single hidden layer.

$$Number\ of\ parameters\ =\ n(x + 2) + 1$$

n is the number of nodes in the single hidden layer and x is the number of input features.

5.1.1.G Recognize the cost function for an ANN.

$$Cost\ function = -\sum_{\substack{Positive \\ outcomes}} \ln(Q) - \sum_{\substack{Negative \\ outcomes}} \ln(1 - Q)$$

Q is the value of the sigmoid (logistic) function for a data point.

5.1.1.H Recognize the outputs of ReLU and leaky ReLU activation functions.

$$Output\ of\ ReLU\ =\ max(x, 0)$$

$$Output\ of\ leaky\ ReLU\ =\ x\ if\ x \geq 0, \alpha x\ if\ x < 0$$

x is the input to the activation function.

5.1.1.I Calculate the output of hyperbolic tangent activation functions.

$$Output\ of\ hyperbolic\ tangent\ = \frac{e^{2x} - 1}{e^{2x} + 1}$$

x is the input to the activation function.

5.1.2.A Calculate the change in the value of a function using the learning rate.

$$Change\ in\ x_i = -learning\ rate * \Delta_i$$

$$New\ value\ of\ x_i = old\ value\ of\ x_i + Change\ in\ x_i$$

x_i is the feature and Δ_i is the gradient of the objective function.

5.1.2.D Calculate the gradient of a function with multiple features.

Requires application of derivate rules to a function.

5.1.2.E Calculate the relationship between a function and its scaled version.

$$Scaled\ value\ of\ a\ variable, x^* = \frac{x - \mu}{\sigma}$$

x is the original variable, μ is the mean of x, and σ is the standard deviation of x.

5.1.4.C Recognize the objective function of an autoencoder.

$$\frac{1}{mn}\sum_{i=1}^{n}\sum_{i=1}^{m}\left(r_{ij} - \hat{r}_{ij}\right)^2$$

m is the number of input nodes and n is the number of data points. r_{ij} is the value for the j^{th} node of the i^{th} observation and $\hat{r}_{ij}$ is the predicted value for the j^{th} node of the i^{th} observation.

5.1.5.C Recognize the probability of exploration using a decay factor.

$$Probability\ of\ exploration, \varepsilon = \beta^{t-1}$$

β is the decay factor and t is the trail number.

5.1.5.F Recognize the objective function having discount factor for reinforcement learning with changing environment.

$$G = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

G is the sum of the expected value of rewards at time t, R_{t+1} is the reward at time $t+1$, and $\gamma(< 1)$ is the discount factor.

5.1.5.J Recognize the updated values of reward using temporal difference updating.

$$Q^{new}(S,A) = Q^{old}(S,A) + \alpha[R + \gamma V(S') - Q^{old}(S,A)]$$

$Q(S,A)$ is the current estimate of the value of taking action A in state S, α is the weight of the current trial, R is the reward at the next step, γ is the discount factor, and $V(S')$ is the current value of the new state S' after action A is taken in state S.

6. Performance Evaluation, Back-Testing, and False Discoveries

6.1.1.A Calculate accuracy and error rate.

$$Accuracy = \frac{Number\ of\ correct\ decisions\ made}{Total\ number\ of\ decisions\ made}$$

$$Error\ rate = \frac{Number\ of\ incorrect\ decisions\ made}{Total\ number\ of\ decisions\ made}$$

6.1.2.A Calculate the expected value and expected benefit.

$$Expected\ value = p(o_1)v(o_1) + p(o_2)v(o_2) + p(o_3)v(o_3) + \cdots$$

o_i is a possible decision outcome, $P(o_i)$ is its probability, and $v(o_i)$ is its value.

$$Expected\ benefit\ of\ targeting\ consumer\ \boldsymbol{x} = p_R(\boldsymbol{x})v_R + [1 - p_R(\boldsymbol{x})]v_{NR}$$

v_R is the value from a response, v_{NR} is the value from a no response, and $p_R(\boldsymbol{x})$ is the probability of response from consumer $\boldsymbol{x}$.

6.1.2.C Calculate the minimum probability of response for which a customer should be targeted.

$$Solving\ p_R(\boldsymbol{x})v_R + [1 - p_R(\boldsymbol{x})]v_{NR} > 0\ for\ p_R(\boldsymbol{x})$$

v_R is the value from a response, v_{NR} is the value from a no response, and $p_R(\boldsymbol{x})$ is the probability of response from consumer $\boldsymbol{x}$.

6.1.2.E Recognize the expected profit for a classifier with and without using priors.

$$Expected\ profit\ without\ prior = p(\boldsymbol{Y},\boldsymbol{p})b(\boldsymbol{Y},\boldsymbol{p}) + p(\boldsymbol{N},\boldsymbol{p})b(\boldsymbol{N},\boldsymbol{p}) + p(\boldsymbol{N},\boldsymbol{n})b(\boldsymbol{N},\boldsymbol{n}) + p(\boldsymbol{Y},\boldsymbol{n})b(\boldsymbol{Y},\boldsymbol{n})$$

$p(\boldsymbol{Y},\boldsymbol{p})$ is the probability of responding when it is predicted to respond, $b(\boldsymbol{Y},\boldsymbol{p})$ is the benefit of responding when it is predicted to respond, $p(\boldsymbol{N},\boldsymbol{p})$ is the probability of not responding when it is predicted to respond, $b(\boldsymbol{N},\boldsymbol{p})$ is the benefit of not responding when it is predicted to respond, $p(\boldsymbol{N},\boldsymbol{n})$ is the probability of not responding when it is predicted to not respond, $b(\boldsymbol{N},\boldsymbol{n})$ is the benefit of not responding when it is predicted to not respond, $p(\boldsymbol{Y},\boldsymbol{n})$ is the probability of responding when it is predicted to not respond, and $b(\boldsymbol{Y},\boldsymbol{n})$ is the benefit of responding when it is predicted to not respond.

$$Expected\ profit\ with\ prior = p(\boldsymbol{p})[p(\boldsymbol{Y}|\ \boldsymbol{p})b(\boldsymbol{Y},\boldsymbol{p}) + p(\boldsymbol{N}|\boldsymbol{p})c(\boldsymbol{N},\boldsymbol{p})] + p(\boldsymbol{n})[p(\boldsymbol{N}|\boldsymbol{n})b(\boldsymbol{N},\boldsymbol{n}) + p(\boldsymbol{Y}|\boldsymbol{n})c(\boldsymbol{Y},\boldsymbol{n})]$$

$p(\boldsymbol{p})$ is the class prior of responding, $p(\boldsymbol{Y}|\ \boldsymbol{p})$ is the probability of responding given it is predicted to respond, $b(\boldsymbol{Y},\boldsymbol{p})$ is the benefit of responding when it is predicted to respond, $p(\boldsymbol{N}|\boldsymbol{p})$ is the probability of not responding given it is predicted to respond, $c(\boldsymbol{N},\boldsymbol{p})$ is the cost of not responding when it is predicted to respond, $p(\boldsymbol{n})$ is the class prior of not responding, $p(\boldsymbol{N}|\boldsymbol{n})$ is the probability of not responding given it is predicted to not respond, $b(\boldsymbol{N},\boldsymbol{n})$ is the benefit of not responding when it is predicted to not respond, $p(\boldsymbol{Y}|\boldsymbol{n})$ is the probability of responding given it is predicted to not respond, and $c(\boldsymbol{Y},\boldsymbol{n})$ is the cost of responding when it is predicted to not respond.

6.1.2.G Calculate true positive, false positive, true negative, and false negative rates for a confusion matrix.

Confusion matrix

	p	n
Y	TP = True positives	FP =False positives
N	FN = False negatives	TN = True negatives

$$True\ positive\ rate = \frac{TP}{TP + FN}$$

$$False\ positive\ rate = \frac{FP}{FP + TN}$$

$$True\ negative\ rate = \frac{TN}{TN + FP}$$

$$False\ negative\ rate = \frac{FN}{FN + TP}$$

6.1.2.H Calculate and interpret precision and recall.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

6.1.2.I Calculate the value of the F-measure.

$$F\text{-}measure\ = 2\frac{Precision * Recall}{Precision + Recall}$$

6.1.2.J Calculate specificity and sensitivity.

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

6.1.3.B Calculate a confusion matrix using a threshold.

Apply the threshold to the scores and count the number of instances in the four categories shown below.

	p	n
Y	Predicted to respond and responded	Predicted to respond but did not respond
N	Predicted not to respond but responded	Predicted not to respond and did not respond

6.1.3.E Recognize points on a profit curve.

$$Profit = \sum_{i=1}^{n} (p_i - c_i)\mathbf{1}_i$$

n is the number of instances determined from the percentage of population to be targeted after ordering the instances in descending order of probability of responding, p_i and c_i are the amounts of profit and cost for instance i, and $\mathbf{1}_i$ is equal to 1 for instances that are predicted to respond and 0 for instances that are predicted not to respond.

6.1.3.F Calculate the proportion of sample data that can be targeted when a fixed budget is available.

$$Proportion\ of\ sample\ data\ that\ can\ be\ targeted = \frac{Available\ Budget}{Cost\ per\ instance * Number\ of\ data\ points}$$

6.1.3.I Calculate points on an ROC graph using data from a confusion matrix.

Confusion matrix

	p	n
Y	TP = True positives	FP =False positives
N	FN = False negatives	TN = True negatives

$$value\ for\ x\text{-}axis = \frac{FP}{FP + TN}$$

$$value\ for\ y\text{-}axis = \frac{TP}{TP + FN}$$

6.1.3.R Calculate points on a cumulative response curve.

$$value\ for\ x\text{-}axis = Percentage\ of\ test\ instance\ (decreasing\ by\ score)$$

$$value\ for\ y\text{-}axis = Percentage\ of\ positives\ targeted$$

6.2.1.E Calculate confidence limits for sensitivities using the t-statistic.

$$\mu \pm Z * SE$$

μ is the estimated value of the parameter, Z is the value of standard normal distribution corresponding to the confidence interval used, and SE is the standard error.

6.2.1.H Calculate the combined impact of all features in a linear regression when the difference from the mean is used as features.

$$Combined\ impact\ for\ the\ j^{th} observation\ = \sum_{i=1}^{p} (X_{ij} - \bar{X}_i)\hat{\beta}_i$$

$\bar{X}_i$ and $\hat{\beta}_i$ are the average value and estimated coefficient of the i^{th} feature, p is the number of features, and X_{ij} is the value of the i^{th} feature for the j^{th} observation.

6.2.1.I Calculate the probability of a positive and negative outcome for logistic regression.

$$Prob(positive\ outcome) = \frac{1}{1 + e^{-(a+\sum_{i=1}^{p} b_i X_i)}}$$

$$Prob(negative\ outcome) = \frac{e^{-(a+\sum_{i=1}^{p} b_i X_i)}}{1 + e^{-(a+\sum_{i=1}^{p} b_i X_i)}}$$

X_i and b_i are the value and estimated weight for feature i, a is the bias, and e is the base of natural logarithm.

6.2.1.J Recognize the probability of an increase in positive outcomes in a logistic regression for small changes in the value of a continuous or categorical feature.

$$Change\ in\ the\ probability\ of\ an\ increase\ in\ positive\ outcome\ due\ to\ changes\ in\ feature\ i = Prob(positive\ outcome) * Prob(negative\ outcome) * b_i u$$

$Prob(positive\ outcome)$ and $Prob(negative\ outcome)$ are as shown in 6.2.1.I, b_i is the estimated weight for feature i, and u is the amount of increase in feature i.

6.2.1.K Recognize the odds against a given probability.

$$Odds\ against\ =\ e^{-(a+\sum_{i=1}^{p} b_i X_i)}$$

X_i and b_i are the value and estimated weight for feature i, a is the bias, and e is the base of natural logarithm.

6.2.1.L Calculate probabilities from odds on or odds against.

$$Probability\ =\frac{1}{1+odds\ against}$$

$$Probability\ =\frac{odds\ on}{1+odds\ on}$$

6.2.1.P Calculate the contribution of features using Shapley values.

Find contributions of each feature for going from one state to another for the given examples and then find the average of the contributions for each feature.

6.3.1.B Calculate the probability of real effect given a result is significant.

$$P(real|test\ sig)=\frac{P(test\ sig|real)P(real)}{P(test\ sig|real)P(real)+P(test\ sig|not\ real)P(not\ real)}$$
$$P(real|test\ sig)=1-False\ discovery\ rate$$

6.3.1.C Recognize the false discovery rate.

$$False\ discovery\ rate\ =\frac{(1-prevalance)*(1-specificity)}{prevlance*sensitivity+(1-prevalance)*(1-specificity)}$$

$$False\ discovery\ rate\ =\frac{(1-prevalance)*(1-significance\ level)}{prevlance*power+(1-prevalance)*(1-significance\ level)}$$

6.3.1.H Calculate the false discovery rate using conditional probabilities.

$$P(false\ discovery\ rate)=\frac{P(test\ sig|not\ real)P(not\ real)}{P(test\ sig|real)P(real)+P(test\ sig|not\ real)P(not\ real)}$$

6.3.1.I Calculate the conditional probability of the real effect.

$$P(real|test\ sig)=\frac{P(test\ sig|real)P(real)}{P(test\ sig|real)P(real)+P(test\ sig|not\ real)P(not\ real)}$$

6.3.1.J Calculate the odds ratio using the Bayes approach.

$$odds\ ratio = \frac{P(not\ real|test\ sig)}{P(real|test\ sig)} = \frac{P(H_0)}{P(H_1)}\frac{P(test\ sig|H_0)}{P(test\ sig|H_1)}$$

7. Text Mining

7.1.2.C Calculate term frequency (TF), inverse document frequency (IDF), and term frequency inverse document frequency (TFIDF).

$$Term\ frequency\ (TF)\ = \frac{Number\ of\ times\ a\ term\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

$$Inverse\ document\ frequency\ (IDF)\ = 1 + ln\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ a\ term}\right)$$

$$Term\ frequency\ inverse\ document\ frequency\ (TFIDF) = \ TF * IDF$$

ln represents natural logarithm. Note that a different formula for the IDF is used in some other sources. Any question for the FDP exam will be based on the above formula for calculations related to the IDF.

7.1.2.I Calculate IDF using the probability of a term in a set of documents.

$$IDF\ (t) = \ 1 + ln\ \left(\frac{1}{p(t)}\right) = 1 - ln(p(t))$$

$$p(t)\ = \frac{Number\ of\ documents\ containing\ t}{Total\ number\ of\ documents}$$

7.1.2.J Calculate the entropy of a term using IDF.

$$entropy\ (t) = p(t) * IDF(t) + (1 - p(t)) * IDF(not\_t)$$

7.2.1.K Calculate the number of n-grams that can be created from a sentence.

$$Number\ of\ n\text{-}grams\ in\ a\ text\ of\ T\ words\ = \ T + 1 - n$$

This can also be calculated by first finding all the different *n*-grams and then counting the number of *n*-grams.

7.2.1.M Calculate the conditional probability of a document having a particular sentiment.

$$Prob(Positive|words) = \frac{p_1 p_2 \dots p_m P}{p_1 p_2 \dots p_m P + q_1 q_2 \dots q_m Q + r_1 r_2 \dots r_m R}$$

$$Prob(Negative|words) = \frac{q_1 q_2 \dots q_m Q}{p_1 p_2 \dots p_m P + q_1 q_2 \dots q_m Q + r_1 r_2 \dots r_m R}$$

$$Prob(Neutral|words) = \frac{r_1 r_2 \dots r_m R}{p_1 p_2 \dots p_m P + q_1 q_2 \dots q_m Q + r_1 r_2 \dots r_m R}$$

P is the unconditional probability of a document being positive, Q is the unconditional probability of a document being negative, R is the unconditional probability of a document being neutral, p_j is the probability of word j appearing in documents labeled positive, q_j is the probability of word j appearing in documents labeled negative, and r_j is the probability of word j appearing in documents labeled neutral.

7.2.1.O Calculate the conditional probability of a document having a particular label using Laplace smoothing.

Recalculate P, Q, R, p_j, q_j, and r_j in the formuals shown in 7.2.1.M after adding m observations under each category (Positive, Negative, and Neutral) with each observation indicating presence of a particular word and absence of every other words.

9. Fintech Applications

9.3.2.D Calculate and apply the Bonferroni method.

$$\alpha^* = \frac{\alpha}{M}$$

α is the significance level for the hypotheses and M is the number of hypotheses. Null hypothesis is rejected if the calculated *p*-value is less than or equal to α^*.

9.3.2.E Calculate and apply the Holm method.

$$\alpha^* = \frac{\alpha}{M - m + 1}$$

α is the significance level for the hypotheses and M is the number of hypotheses. $m = 1, 2, \ldots, M$ when the tests are ordered from the most significant to the least significant. $m = 1$ for the most significant test and $m = M$ for the least significant test. Null hypothesis is rejected if the calculated *p*-value is less than or equal to α^*.

9.3.2.M Calculate and apply the Benjamini and Hochberg (BH) method for controlling FDR.

$$j^* = max\left\{j: p_j \leq \frac{j * \delta}{M}\right\}$$

j^* is the rank order of the least significant *p*-value that corresponds to a rejected hypothesis. *j* is the $j\text{-}th$ most significant hypothesis with *p*-value of p_j. δ is the significance level and M is the number of hypotheses.

9.3.2.N Recognize and apply the Benjamini and Yakuteli (BY) for controlling the FDR.

$$j^* = max\left\{j: p_j \leq \frac{j * \delta}{M * C_M}\right\}$$

j^* is the rank order of the least significant *p*-value that corresponds to a rejected hypothesis. j is the $j\text{-}th$ most significant hypothesis with *p*-value of p_j. δ is the significance level and M is the number of hypotheses. $C_M = \sum_{i=1}^{M} \frac{1}{i}$

9.4.3.D Recognize simple matching dissimilarity measure and Rogers and Tanimoto dissimilarity measure.

		Document 2	
		1	0
Document 1	1	a	b
	0	c	d

Crosstabulation of counts for two records

$$Simple\ matching\ dissimilarity\ measure\ = \frac{b+c}{a+b+c+d}$$

$$Rogers\ and\ Tanimoto\ dissimilarity\ measure\ = \frac{2(b+c)}{a+d+2(b+c)}$$