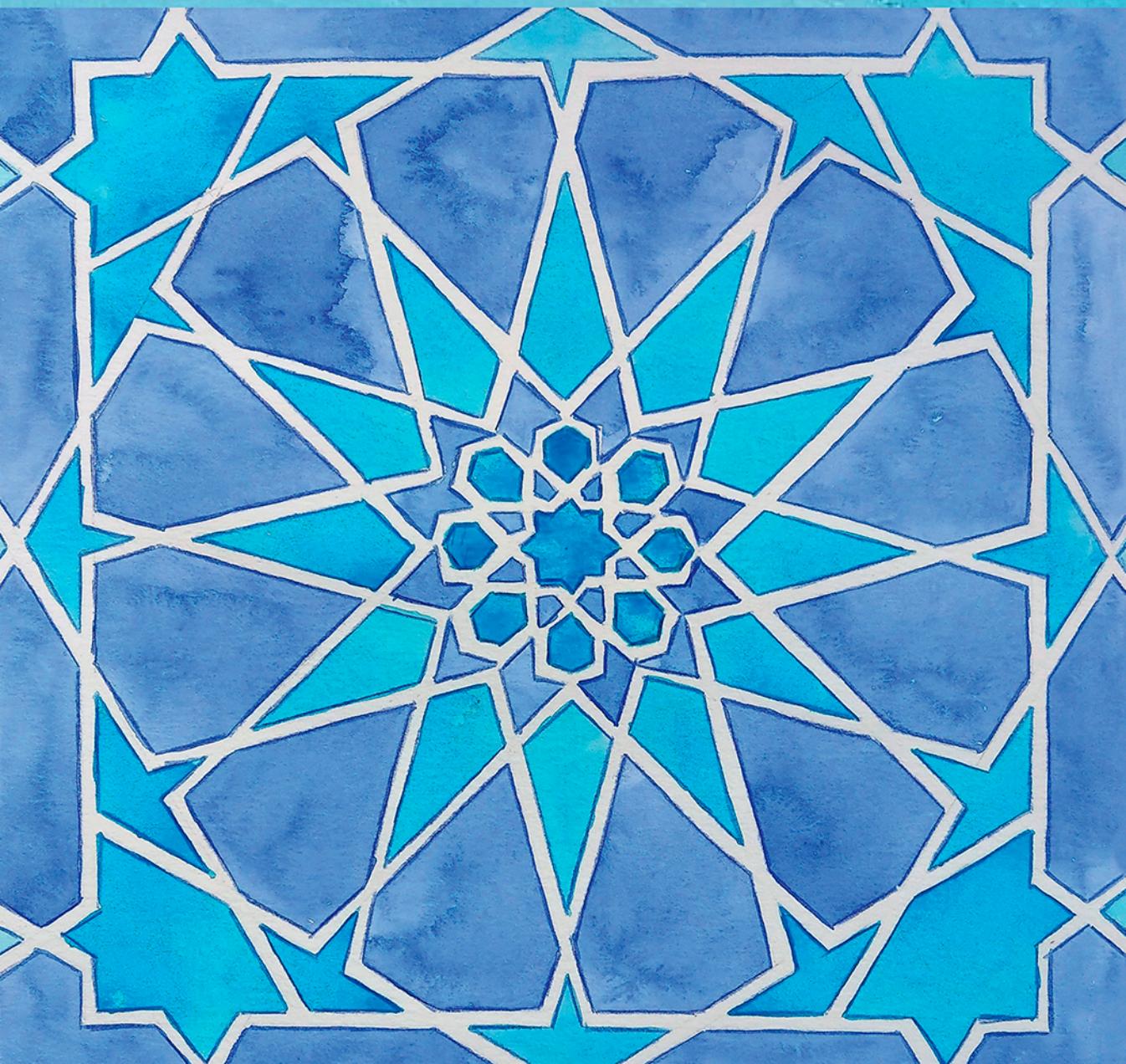


# Data Mining and Machine Learning

FUNDAMENTAL CONCEPTS  
AND ALGORITHMS

SECOND EDITION

MOHAMMED J. ZAKI  
WAGNER MEIRA, JR.





## **DATA MINING AND MACHINE LEARNING**

The fundamental algorithms in data mining and machine learning form the basis of data science, utilizing automated methods to analyze patterns and models for all kinds of data in applications ranging from scientific discovery to business analytics. This textbook for senior undergraduate and graduate courses provides a comprehensive, in-depth overview of data mining, machine learning and statistics, offering solid guidance for students, researchers, and practitioners. The book lays the foundations of data analysis, pattern mining, clustering, classification and regression, with a focus on the algorithms and the underlying algebraic, geometric, and probabilistic concepts. New to this second edition is an entire part devoted to regression methods, including neural networks and deep learning.

Mohammed J. Zaki is Professor of Computer Science at Rensselaer Polytechnic Institute, where he also serves as Associate Department Head and Graduate Program Director. He has more than 250 publications and is an Associate Editor for the journal Data Mining and Knowledge Discovery. He is on the Board of Directors for ACM SIGKDD. He has received the National Science Foundation CAREER Award, and the Department of Energy Early Career Principal Investigator Award. He is an ACM Distinguished Member, and IEEE Fellow.

Wagner Meira, Jr. is Professor of Computer Science at Universidade Federal de Minas Gerais, Brazil, where he is currently the chair of the department. He has published more than 230 papers on data mining and parallel and distributed systems. He was leader of the Knowledge Discovery research track of InWeb and is currently Vice-chair of INCT-Cyber. He is on the editorial board of the journal Data Mining and Knowledge Discovery and was the program chair of SDM'16 and ACM WebSci'19. He has been a CNPq researcher since 2002. He has received an IBM Faculty Award and several Google Faculty Research Awards.



# **DATA MINING AND MACHINE LEARNING**

## **Fundamental Concepts and Algorithms**

**MOHAMMED J. ZAKI**

Rensselaer Polytechnic Institute

**WAGNER MEIRA, JR.**

Universidade Federal de Minas Gerais



**CAMBRIDGE**  
UNIVERSITY PRESS

# CAMBRIDGE

## UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

[www.cambridge.org](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9781108473989](http://www.cambridge.org/9781108473989)

DOI: [10.1017/9781108564175](https://doi.org/10.1017/9781108564175)

© Mohammed J. Zaki and Wagner Meira, Jr. 2020

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First edition published 2014, second edition published 2020

Printed in the United Kingdom by TJ International Ltd., Padstow, Cornwall

*A catalog record for this publication is available from the British Library.*

*Library of Congress Cataloging in Publication Data*

Names: Zaki, Mohammed J., 1971- author. | Meira, Wagner, 1967- author.

Title: Data mining and machine learning : fundamental concepts and algorithms / Mohammed J. Zaki, Wagner Meira, Jr.

Other titles: Data mining and analysis

Description: Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2020. | Revised edition of: Data mining and analysis.

2014. | Includes bibliographical references and index.

Identifiers: LCCN 2019037293 (print) | LCCN 2019037294 (ebook) | ISBN 9781108473989 (hardback) | ISBN 9781108564175 (epub)

Subjects: LCSH: Data mining.

Classification: LCC QA76.9.D343 Z36 2020 (print) | LCC QA76.9.D343 (ebook) | DDC 006.3/12–dc23

LC record available at <https://lccn.loc.gov/2019037293>

LC ebook record available at <https://lccn.loc.gov/2019037294>

ISBN 978-1-108-47398-9 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate.

# Contents

|  |           |
|--|-----------|
| Preface                                    | xii       |
| <b>PART ONE: DATA ANALYSIS FOUNDATIONS</b> |           |
| <b>1 Data Matrix . . . . .</b>             | <b>3</b>  |
| 1.1 Data Matrix                            | 3         |
| 1.2 Attributes                             | 4         |
| 1.3 Data: Algebraic and Geometric View     | 5         |
| 1.4 Data: Probabilistic View               | 16        |
| 1.5 Further Reading                        | 28        |
| 1.6 Exercises                              | 28        |
| <b>2 Numeric Attributes . . . . .</b>      | <b>29</b> |
| 2.1 Univariate Analysis                    | 29        |
| 2.2 Bivariate Analysis                     | 40        |
| 2.3 Multivariate Analysis                  | 46        |
| 2.4 Data Normalization                     | 50        |
| 2.5 Normal Distribution                    | 52        |
| 2.6 Further Reading                        | 58        |
| 2.7 Exercises                              | 58        |
| <b>3 Categorical Attributes . . . . .</b>  | <b>61</b> |
| 3.1 Univariate Analysis                    | 61        |
| 3.2 Bivariate Analysis                     | 70        |
| 3.3 Multivariate Analysis                  | 81        |
| 3.4 Distance and Angle                     | 86        |
| 3.5 Discretization                         | 87        |
| 3.6 Further Reading                        | 89        |
| 3.7 Exercises                              | 90        |
| <b>4 Graph Data . . . . .</b>              | <b>92</b> |
| 4.1 Graph Concepts                         | 92        |
| 4.2 Topological Attributes                 | 96        |
| 4.3 Centrality Analysis                    | 101       |

|  |  |            |
|--|--|------------|
| 4.4                                      | Graph Models                                       | 111        |
| 4.5                                      | Further Reading                                    | 131        |
| 4.6                                      | Exercises  | 132        |
| <b>5</b>                                 | <b>Kernel Methods . . . . .</b>                    | <b>134</b> |
| 5.1                                      | Kernel Matrix                                      | 138        |
| 5.2                                      | Vector Kernels                                     | 144        |
| 5.3                                      | Basic Kernel Operations in Feature Space           | 149        |
| 5.4                                      | Kernels for Complex Objects                        | 155        |
| 5.5                                      | Further Reading                                    | 161        |
| 5.6                                      | Exercises  | 161        |
| <b>6</b>                                 | <b>High-dimensional Data . . . . .</b>             | <b>163</b> |
| 6.1                                      | High-dimensional Objects                           | 163        |
| 6.2                                      | High-dimensional Volumes                           | 167        |
| 6.3                                      | Hypersphere Inscribed within Hypercube             | 170        |
| 6.4                                      | Volume of Thin Hypersphere Shell                   | 171        |
| 6.5                                      | Diagonals in Hyperspace                            | 172        |
| 6.6                                      | Density of the Multivariate Normal                 | 173        |
| 6.7                                      | Appendix: Derivation of Hypersphere Volume         | 177        |
| 6.8                                      | Further Reading                                    | 181        |
| 6.9                                      | Exercises  | 181        |
| <b>7</b>                                 | <b>Dimensionality Reduction . . . . .</b>          | <b>184</b> |
| 7.1                                      | Background   | 184        |
| 7.2                                      | Principal Component Analysis                       | 188        |
| 7.3                                      | Kernel Principal Component Analysis                | 203        |
| 7.4                                      | Singular Value Decomposition                       | 210        |
| 7.5                                      | Further Reading                                    | 215        |
| 7.6                                      | Exercises  | 215        |
| <b>PART TWO: FREQUENT PATTERN MINING</b> |  | <b>217</b> |
| <b>8</b>                                 | <b>Itemset Mining . . . . .</b>                    | <b>219</b> |
| 8.1                                      | Frequent Itemsets and Association Rules            | 219        |
| 8.2                                      | Itemset Mining Algorithms                          | 223        |
| 8.3                                      | Generating Association Rules                       | 237        |
| 8.4                                      | Further Reading                                    | 238        |
| 8.5                                      | Exercises  | 239        |
| <b>9</b>                                 | <b>Summarizing Itemsets . . . . .</b>              | <b>244</b> |
| 9.1                                      | Maximal and Closed Frequent Itemsets               | 244        |
| 9.2                                      | Mining Maximal Frequent Itemsets: GenMax Algorithm | 247        |
| 9.3                                      | Mining Closed Frequent Itemsets: Charm Algorithm   | 250        |
| 9.4                                      | Nonderrivable Itemsets                             | 252        |
| 9.5                                      | Further Reading                                    | 258        |
| 9.6                                      | Exercises  | 258        |

|   |            |
|---|------------|
| <b>10 Sequence Mining . . . . .</b>                 | <b>261</b> |
| 10.1 Frequent Sequences                             | 261        |
| 10.2 Mining Frequent Sequences                      | 262        |
| 10.3 Substring Mining via Suffix Trees              | 269        |
| 10.4 Further Reading                                | 279        |
| 10.5 Exercises                                      | 279        |
| <b>11 Graph Pattern Mining . . . . .</b>            | <b>282</b> |
| 11.1 Isomorphism and Support                        | 282        |
| 11.2 Candidate Generation                           | 286        |
| 11.3 The gSpan Algorithm                            | 290        |
| 11.4 Further Reading                                | 298        |
| 11.5 Exercises                                      | 299        |
| <b>12 Pattern and Rule Assessment . . . . .</b>     | <b>303</b> |
| 12.1 Rule and Pattern Assessment Measures           | 303        |
| 12.2 Significance Testing and Confidence Intervals  | 318        |
| 12.3 Further Reading                                | 330        |
| 12.4 Exercises                                      | 330        |
| <b>PART THREE: CLUSTERING</b>                       | <b>332</b> |
| <b>13 Representative-based Clustering . . . . .</b> | <b>334</b> |
| 13.1 K-means Algorithm                              | 334        |
| 13.2 Kernel K-means                                 | 339        |
| 13.3 Expectation-Maximization Clustering            | 343        |
| 13.4 Further Reading                                | 360        |
| 13.5 Exercises                                      | 361        |
| <b>14 Hierarchical Clustering . . . . .</b>         | <b>364</b> |
| 14.1 Preliminaries                                  | 364        |
| 14.2 Agglomerative Hierarchical Clustering          | 366        |
| 14.3 Further Reading                                | 372        |
| 14.4 Exercises                                      | 373        |
| <b>15 Density-based Clustering . . . . .</b>        | <b>375</b> |
| 15.1 The DBSCAN Algorithm                           | 375        |
| 15.2 Kernel Density Estimation                      | 379        |
| 15.3 Density-based Clustering: DENCLUE              | 385        |
| 15.4 Further Reading                                | 390        |
| 15.5 Exercises                                      | 391        |
| <b>16 Spectral and Graph Clustering . . . . .</b>   | <b>394</b> |
| 16.1 Graphs and Matrices                            | 394        |
| 16.2 Clustering as Graph Cuts                       | 401        |
| 16.3 Markov Clustering                              | 417        |
| 16.4 Further Reading                                | 422        |
| 16.5 Exercises                                      | 424        |

|  |            |
|--|------------|
| <b>17 Clustering Validation . . . . .</b>          | <b>426</b> |
| 17.1 External Measures                             | 426        |
| 17.2 Internal Measures                             | 441        |
| 17.3 Relative Measures                             | 450        |
| 17.4 Further Reading                               | 464        |
| 17.5 Exercises                                     | 465        |
| <b>PART FOUR: CLASSIFICATION</b>                   | <b>467</b> |
| <b>18 Probabilistic Classification . . . . .</b>   | <b>469</b> |
| 18.1 Bayes Classifier                              | 469        |
| 18.2 Naive Bayes Classifier                        | 475        |
| 18.3 $K$ Nearest Neighbors Classifier              | 479        |
| 18.4 Further Reading                               | 480        |
| 18.5 Exercises                                     | 482        |
| <b>19 Decision Tree Classifier . . . . .</b>       | <b>483</b> |
| 19.1 Decision Trees                                | 485        |
| 19.2 Decision Tree Algorithm                       | 487        |
| 19.3 Further Reading                               | 498        |
| 19.4 Exercises                                     | 499        |
| <b>20 Linear Discriminant Analysis . . . . .</b>   | <b>501</b> |
| 20.1 Optimal Linear Discriminant                   | 501        |
| 20.2 Kernel Discriminant Analysis                  | 508        |
| 20.3 Further Reading                               | 515        |
| 20.4 Exercises                                     | 515        |
| <b>21 Support Vector Machines . . . . .</b>        | <b>517</b> |
| 21.1 Support Vectors and Margins                   | 517        |
| 21.2 SVM: Linear and Separable Case                | 523        |
| 21.3 Soft Margin SVM: Linear and Nonseparable Case | 527        |
| 21.4 Kernel SVM: Nonlinear Case                    | 533        |
| 21.5 SVM Training: Stochastic Gradient Ascent      | 537        |
| 21.6 Further Reading                               | 543        |
| 21.7 Exercises                                     | 544        |
| <b>22 Classification Assessment . . . . .</b>      | <b>546</b> |
| 22.1 Classification Performance Measures           | 546        |
| 22.2 Classifier Evaluation                         | 560        |
| 22.3 Bias-Variance Decomposition                   | 570        |
| 22.4 Ensemble Classifiers                          | 574        |
| 22.5 Further Reading                               | 584        |
| 22.6 Exercises                                     | 585        |
| <b>PART FIVE: REGRESSION</b>                       | <b>587</b> |
| <b>23 Linear Regression . . . . .</b>              | <b>589</b> |
| 23.1 Linear Regression Model                       | 589        |

|              |  |            |
|--------------|--|------------|
| 23.2         | Bivariate Regression                           | 590        |
| 23.3         | Multiple Regression                            | 596        |
| 23.4         | Ridge Regression                               | 606        |
| 23.5         | Kernel Regression                              | 611        |
| 23.6         | $L_1$ Regression: Lasso                        | 615        |
| 23.7         | Further Reading                                | 621        |
| 23.8         | Exercises                                      | 621        |
| <b>24</b>    | <b>Logistic Regression</b>                     | <b>623</b> |
| 24.1         | Binary Logistic Regression                     | 623        |
| 24.2         | Multiclass Logistic Regression                 | 630        |
| 24.3         | Further Reading                                | 635        |
| 24.4         | Exercises                                      | 635        |
| <b>25</b>    | <b>Neural Networks</b>                         | <b>637</b> |
| 25.1         | Artificial Neuron: Activation Functions        | 637        |
| 25.2         | Neural Networks: Regression and Classification | 642        |
| 25.3         | Multilayer Perceptron: One Hidden Layer        | 648        |
| 25.4         | Deep Multilayer Perceptrons                    | 660        |
| 25.5         | Further Reading                                | 670        |
| 25.6         | Exercises                                      | 670        |
| <b>26</b>    | <b>Deep Learning</b>                           | <b>672</b> |
| 26.1         | Recurrent Neural Networks                      | 672        |
| 26.2         | Gated RNNs: Long Short-Term Memory Networks    | 682        |
| 26.3         | Convolutional Neural Networks                  | 694        |
| 26.4         | Regularization                                 | 712        |
| 26.5         | Further Reading                                | 717        |
| 26.6         | Exercises                                      | 718        |
| <b>27</b>    | <b>Regression Evaluation</b>                   | <b>720</b> |
| 27.1         | Univariate Regression                          | 721        |
| 27.2         | Multiple Regression                            | 735        |
| 27.3         | Further Reading                                | 752        |
| 27.4         | Exercises                                      | 752        |
| <b>Index</b> |  | <b>755</b> |



## Preface

Data mining and machine learning enable one to gain fundamental insights and knowledge from data. They allow the discovery of insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data.

There are several good books in this area, but many of them are either too high-level or too advanced. This book is an introductory text that lays the foundations for the fundamental concepts and algorithms in machine learning and data mining. Important concepts are explained when first encountered, with detailed steps and derivations. A key goal of the book is to build the intuition behind the formulas via the interplay of geometric, (linear) algebraic and probabilistic interpretations of the data and the methods.

This second edition adds a whole new part on regression, including linear and logistic regression, neural networks, and deep learning. Content has also been updated in several other chapters and known errata have been fixed. The main parts of the book include data analysis foundations, frequent pattern mining, clustering, classification, and regression. These cover the core methods as well as cutting-edge topics such as deep learning, kernel methods, high-dimensional data analysis, and graph analysis.

The book includes many examples to illustrate the concepts and algorithms. It also has end-of-chapter exercises, which have been used in class. All of the algorithms in the book have been implemented by the authors. To aid practical understanding, we suggest that readers implement these algorithms on their own (using, for example, Python or R). Supplementary resources like slides, datasets and videos are available online at the book's companion site:

<http://dataminingbook.info>

The book can be used for both undergraduate and graduate courses in data mining, machine learning, and data science. A brief overview of the chapters is presented at the start of each part of the book. The chapters are mainly self contained (with important equations highlighted), but introductory courses would benefit by covering the basic foundations of data analysis in part one. For example, the kernel methods chapter in part one should be covered before other kernel-based algorithms that appear in later

parts. The different parts can be covered in a different order based on the emphasis of the course or the interest of the reader. Finally, we encourage you to contact us about errata or other suggestions via the book companion site.

Mohammed J. Zaki and Wagner Meira, Jr.

**PART ONE**

## DATA ANALYSIS FOUNDATIONS

This part lays the algebraic and probabilistic foundations for data analysis. It begins with basic statistical analysis of univariate and multivariate numeric data in [Chapter 2](#). We describe measures of central tendency such as mean, median, and mode, and then we consider measures of dispersion such as range, variance, and covariance. We emphasize the dual algebraic and probabilistic views, and highlight the geometric interpretation of the various measures. We especially focus on the multivariate normal distribution, which is widely used as the default parametric model for data in both classification and clustering. In [Chapter 3](#) we show how categorical data can be modeled via the multivariate binomial and the multinomial distributions. We describe the contingency table analysis approach to test for dependence between categorical attributes.

In [Chapter 4](#), we show how to analyze graph data in terms of the topological structure, with special focus on various graph centrality measures such as closeness, betweenness, prestige, PageRank, and so on. We also study basic topological properties of real-world networks such as the *small world property*, which states that real graphs have small average path length between pairs of nodes, the *clustering effect*, which indicates local clustering around nodes, and the *scale-free property*, which manifests itself in a *power-law* degree distribution. We describe models that can explain some of these characteristics of real-world graphs; these include the Erdős–Rényi random graph model, the Watts–Strogatz model, and the Barabási–Albert model.

Kernel methods are introduced in [Chapter 5](#), which provides new insights and connections between linear, nonlinear, graph, and complex data mining tasks. We briefly highlight the theory behind kernel functions, with the key concept being that a positive semidefinite kernel corresponds to a dot product in some high-dimensional feature space, and thus we can use familiar numeric analysis methods for nonlinear or complex object analysis, provided we can compute the pairwise kernel matrix of similarities between object instances. We describe various kernels for numeric or vector data, as well as sequence and graph data.

In [Chapter 6](#), we consider the peculiarities of high-dimensional space, colorfully referred to as *the curse of dimensionality*. In particular, we study the scattering effect, that is, the fact that data points lie along the boundaries and corners in high dimensions, with the “center” of the space being virtually empty. We show the proliferation of orthogonal axes and also the behavior of the multivariate normal distribution in high dimensions. Finally, in [Chapter 7](#), we describe widely used dimensionality

reduction methods such as principal component analysis (PCA) and singular value decomposition (SVD). PCA finds the optimal  $k$ -dimensional subspace that captures most of the variance in the data. We also show how kernel PCA can be used to find nonlinear directions that capture the most variance. We conclude with the powerful SVD spectral decomposition method, studying its geometry, and its relationship to PCA.

We begin this chapter by looking at basic properties of data modeled as a data matrix. We emphasize the geometric and algebraic views, as well as the probabilistic interpretation of data, which play a key role in machine learning and data mining.

## 1.1 DATA MATRIX

---

Data can often be represented or abstracted as an  $n \times d$  *data matrix*, with  $n$  rows and  $d$  columns, where rows correspond to entities in the dataset, and columns represent attributes or properties of interest. Each row in the data matrix records the observed attribute values for a given entity. The  $n \times d$  data matrix is given as

$$\mathbf{D} = \left( \begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array} \right)$$

where  $\mathbf{x}_i$  denotes the  $i$ th row, which is a  $d$ -tuple given as

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})$$

and  $X_j$  denotes the  $j$ th column, which is an  $n$ -tuple given as

$$X_j = (x_{1j}, x_{2j}, \dots, x_{nj})$$

Depending on the application domain, rows may also be referred to as *entities*, *instances*, *examples*, *records*, *transactions*, *objects*, *points*, *feature-vectors*, *tuples*, and so on. Likewise, columns may also be called *attributes*, *properties*, *features*, *dimensions*, *variables*, *fields*, and so on. The number of instances  $n$  is referred to as the *size* of the data, whereas the number of attributes  $d$  is called the *dimensionality* of the data. The analysis of a single attribute is referred to as *univariate analysis*, whereas the simultaneous analysis of two attributes is called *bivariate analysis* and the simultaneous analysis of more than two attributes is called *multivariate analysis*.

Table 1.1. Extract from the Iris dataset

|                    | Sepal<br>length | Sepal<br>width | Petal<br>length | Petal<br>width | Class           |
|--------------------|-----------------|----------------|-----------------|----------------|-----------------|
|                    | $X_1$           | $X_2$          | $X_3$           | $X_4$          | $X_5$           |
| $\mathbf{x}_1$     | 5.9             | 3.0            | 4.2             | 1.5            | Iris-versicolor |
| $\mathbf{x}_2$     | 6.9             | 3.1            | 4.9             | 1.5            | Iris-versicolor |
| $\mathbf{x}_3$     | 6.6             | 2.9            | 4.6             | 1.3            | Iris-versicolor |
| $\mathbf{x}_4$     | 4.6             | 3.2            | 1.4             | 0.2            | Iris-setosa     |
| $\mathbf{x}_5$     | 6.0             | 2.2            | 4.0             | 1.0            | Iris-versicolor |
| $\mathbf{x}_6$     | 4.7             | 3.2            | 1.3             | 0.2            | Iris-setosa     |
| $\mathbf{x}_7$     | 6.5             | 3.0            | 5.8             | 2.2            | Iris-virginica  |
| $\mathbf{x}_8$     | 5.8             | 2.7            | 5.1             | 1.9            | Iris-virginica  |
| $\vdots$           | $\vdots$        | $\vdots$       | $\vdots$        | $\vdots$       | $\vdots$        |
| $\mathbf{x}_{149}$ | 7.7             | 3.8            | 6.7             | 2.2            | Iris-virginica  |
| $\mathbf{x}_{150}$ | 5.1             | 3.4            | 1.5             | 0.2            | Iris-setosa     |

**Example 1.1.** Table 1.1 shows an extract of the Iris dataset; the complete data forms a  $150 \times 5$  data matrix. Each entity is an Iris flower, and the attributes include sepal length, sepal width, petal length, and petal width in centimeters, and the type or class of the Iris flower. The first row is given as the 5-tuple

$$\mathbf{x}_1 = (5.9, 3.0, 4.2, 1.5, \text{Iris-versicolor})$$

Not all datasets are in the form of a data matrix. For instance, more complex datasets can be in the form of sequences (e.g., DNA and protein sequences), text, time-series, images, audio, video, and so on, which may need special techniques for analysis. However, in many cases even if the raw data is not a data matrix it can usually be transformed into that form via feature extraction. For example, given a database of images, we can create a data matrix in which rows represent images and columns correspond to image features such as color, texture, and so on. Sometimes, certain attributes may have special semantics associated with them requiring special treatment. For instance, temporal or spatial attributes are often treated differently. It is also worth noting that traditional data analysis assumes that each entity or instance is independent. However, given the interconnected nature of the world we live in, this assumption may not always hold. Instances may be connected to other instances via various kinds of relationships, giving rise to a *data graph*, where a node represents an entity and an edge represents the relationship between two entities.

## 1.2 ATTRIBUTES

---

Attributes may be classified into two main types depending on their domain, that is, depending on the types of values they take on.

### Numeric Attributes

A *numeric* attribute is one that has a real-valued or integer-valued domain. For example, *Age* with  $\text{domain}(\text{Age}) = \mathbb{N}$ , where  $\mathbb{N}$  denotes the set of natural numbers (non-negative integers), is numeric, and so is *petal length* in [Table 1.1](#), with  $\text{domain}(\text{petal length}) = \mathbb{R}^+$  (the set of all positive real numbers). Numeric attributes that take on a finite or countably infinite set of values are called *discrete*, whereas those that can take on any real value are called *continuous*. As a special case of discrete, if an attribute has as its domain the set  $\{0, 1\}$ , it is called a *binary* attribute. Numeric attributes can be classified further into two types:

- *Interval-scaled*: For these kinds of attributes only differences (addition or subtraction) make sense. For example, attribute *temperature* measured in  $^{\circ}\text{C}$  or  $^{\circ}\text{F}$  is interval-scaled. If it is  $20\ ^{\circ}\text{C}$  on one day and  $10\ ^{\circ}\text{C}$  on the following day, it is meaningful to talk about a temperature drop of  $10\ ^{\circ}\text{C}$ , but it is not meaningful to say that it is twice as cold as the previous day.
- *Ratio-scaled*: Here one can compute both differences as well as ratios between values. For example, for attribute *Age*, we can say that someone who is 20 years old is twice as old as someone who is 10 years old.

### Categorical Attributes

A *categorical* attribute is one that has a set-valued domain composed of a set of symbols. For example, *Sex* and *Education* could be categorical attributes with their domains given as

$$\text{domain}(\text{Sex}) = \{\text{M}, \text{F}\}$$

$$\text{domain}(\text{Education}) = \{\text{HighSchool}, \text{BS}, \text{MS}, \text{PhD}\}$$

Categorical attributes may be of two types:

- *Nominal*: The attribute values in the domain are unordered, and thus only equality comparisons are meaningful. That is, we can check only whether the value of the attribute for two given instances is the same or not. For example, *Sex* is a nominal attribute. Also *class* in [Table 1.1](#) is a nominal attribute with  $\text{domain}(\text{class}) = \{\text{iris-setosa}, \text{iris-versicolor}, \text{iris-virginica}\}$ .
- *Ordinal*: The attribute values are ordered, and thus both equality comparisons (is one value equal to another?) and inequality comparisons (is one value less than or greater than another?) are allowed, though it may not be possible to quantify the difference between values. For example, *Education* is an ordinal attribute because its domain values are ordered by increasing educational qualification.

## 1.3 DATA: ALGEBRAIC AND GEOMETRIC VIEW

---

If the  $d$  attributes or dimensions in the data matrix  $\mathbf{D}$  are all numeric, then each row can be considered as a  $d$ -dimensional point:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id}) \in \mathbb{R}^d$$

or equivalently, each row may be considered as a  $d$ -dimensional column vector (all vectors are assumed to be column vectors by default):

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} = (x_{i1} \quad x_{i2} \quad \cdots \quad x_{id})^T \in \mathbb{R}^d$$

where  $T$  is the (*matrix*) *transpose* operator.

The  $d$ -dimensional Cartesian coordinate space is specified via the  $d$  unit vectors, called the standard basis vectors, along each of the axes. The  $j$ th *standard basis vector*  $\mathbf{e}_j$  is the  $d$ -dimensional unit vector whose  $j$ th component is 1 and the rest of the components are 0

$$\mathbf{e}_j = (0, \dots, 1_j, \dots, 0)^T$$

Any other vector in  $\mathbb{R}^d$  can be written as a *linear combination* of the standard basis vectors. For example, each of the points  $\mathbf{x}_i$  can be written as the linear combination

$$\mathbf{x}_i = x_{i1}\mathbf{e}_1 + x_{i2}\mathbf{e}_2 + \cdots + x_{id}\mathbf{e}_d = \sum_{j=1}^d x_{ij}\mathbf{e}_j$$

where the scalar value  $x_{ij}$  is the coordinate value along the  $j$ th axis or attribute.

**Example 1.2.** Consider the Iris data in [Table 1.1](#). If we *project* the entire data onto the first two attributes, then each row can be considered as a point or a vector in 2-dimensional space. For example, the projection of the 5-tuple  $\mathbf{x}_1 = (5.9, 3.0, 4.2, 1.5, \text{Iris-versicolor})$  on the first two attributes is shown in [Figure 1.1\(a\)](#). [Figure 1.2](#) shows the scatterplot of all the  $n = 150$  points in the 2-dimensional space spanned by the first two attributes. Likewise, [Figure 1.1\(b\)](#) shows  $\mathbf{x}_1$  as a point and vector in 3-dimensional space, by projecting the data onto the first three attributes. The point  $(5.9, 3.0, 4.2)$  can be seen as specifying the coefficients in the linear combination of the standard basis vectors in  $\mathbb{R}^3$ :

$$\mathbf{x}_1 = 5.9\mathbf{e}_1 + 3.0\mathbf{e}_2 + 4.2\mathbf{e}_3 = 5.9 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 3.0 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 4.2 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 5.9 \\ 3.0 \\ 4.2 \end{pmatrix}$$

Each numeric column or attribute can also be treated as a vector in an  $n$ -dimensional space  $\mathbb{R}^n$ :

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

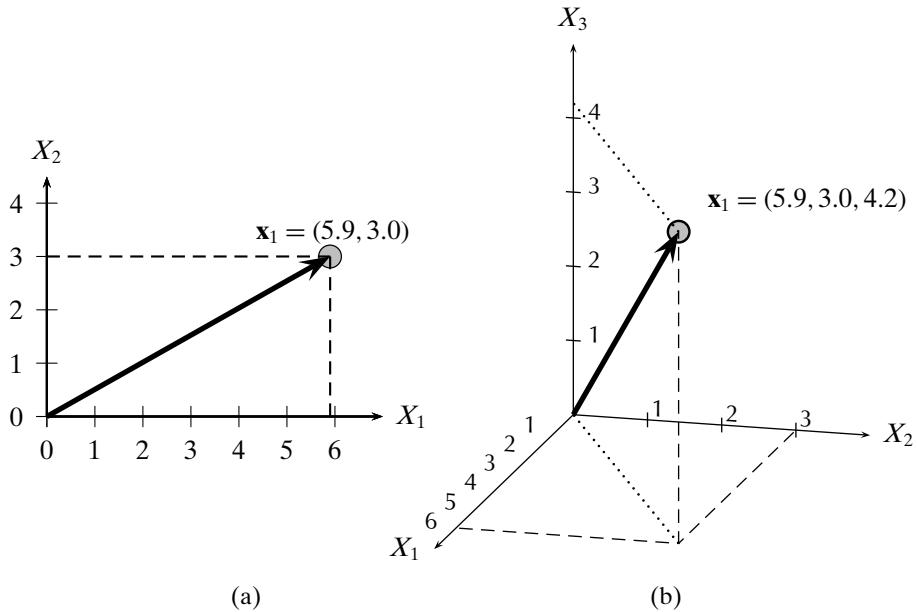


Figure 1.1. Row  $\mathbf{x}_1$  as a point and vector in (a)  $\mathbb{R}^2$  and (b)  $\mathbb{R}^3$ .

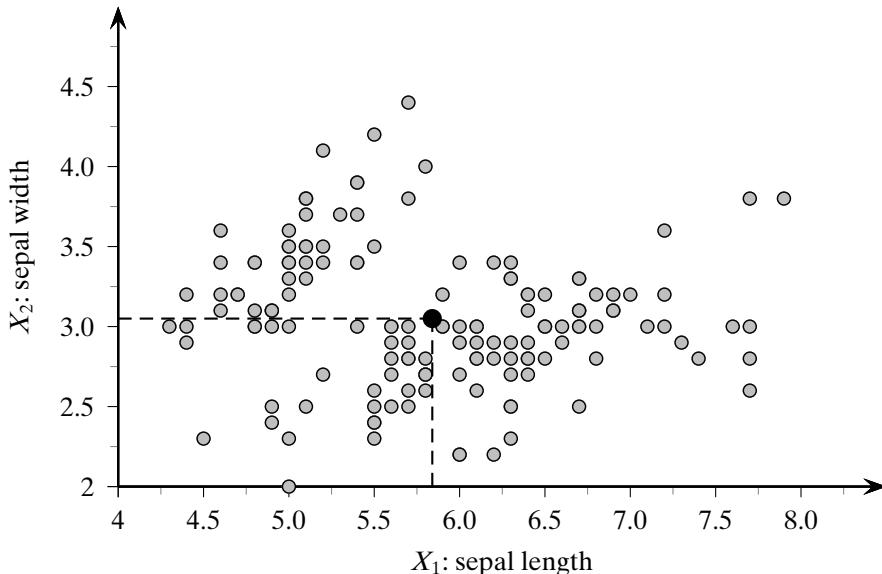


Figure 1.2. Scatterplot: sepal length versus sepal width. The solid circle shows the mean point.

If all attributes are numeric, then the data matrix  $\mathbf{D}$  is in fact an  $n \times d$  matrix, also written as  $\mathbf{D} \in \mathbb{R}^{n \times d}$ , given as

$$\mathbf{D} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} = \begin{pmatrix} -\mathbf{x}_1^T- \\ -\mathbf{x}_2^T- \\ \vdots \\ -\mathbf{x}_n^T- \end{pmatrix} = \begin{pmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & & | \end{pmatrix}$$

As we can see, we can consider the entire dataset as an  $n \times d$  matrix, or equivalently as a set of  $n$  row vectors  $\mathbf{x}_i^T \in \mathbb{R}^d$  or as a set of  $d$  column vectors  $X_j \in \mathbb{R}^n$ .

### 1.3.1 Distance and Angle

Treating data instances and attributes as vectors, and the entire dataset as a matrix, enables one to apply both geometric and algebraic methods to aid in the data mining and analysis tasks.

Let  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$  be two  $m$ -dimensional vectors given as

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

#### Dot Product

The *dot product* between  $\mathbf{a}$  and  $\mathbf{b}$  is defined as the scalar value

$$\mathbf{a}^T \mathbf{b} = (a_1 \ a_2 \ \cdots \ a_m) \times \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = a_1 b_1 + a_2 b_2 + \cdots + a_m b_m = \sum_{i=1}^m a_i b_i \quad (1.1)$$

#### Length

The *Euclidean norm* or *length* of a vector  $\mathbf{a} \in \mathbb{R}^m$  is defined as

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}^T \mathbf{a}} = \sqrt{a_1^2 + a_2^2 + \cdots + a_m^2} = \sqrt{\sum_{i=1}^m a_i^2} \quad (1.2)$$

The *unit vector* in the direction of  $\mathbf{a}$  is given as

$$\mathbf{u} = \frac{\mathbf{a}}{\|\mathbf{a}\|} = \left( \frac{1}{\|\mathbf{a}\|} \right) \mathbf{a}$$

By definition  $\mathbf{u}$  has length  $\|\mathbf{u}\| = 1$ , and it is also called a *normalized* vector, which can be used in lieu of  $\mathbf{a}$  in some analysis tasks.

The Euclidean norm is a special case of a general class of norms, known as  $L_p$ -norm, defined as

$$\|\mathbf{a}\|_p = \left( |a_1|^p + |a_2|^p + \cdots + |a_m|^p \right)^{\frac{1}{p}} = \left( \sum_{i=1}^m |a_i|^p \right)^{\frac{1}{p}} \quad (1.3)$$

for any  $p \neq 0$ . Thus, the Euclidean norm corresponds to the case when  $p = 2$ ; it is also called the  $L_2$ -norm.

### Distance

From the Euclidean norm we can define the *Euclidean distance* between  $\mathbf{a}$  and  $\mathbf{b}$ , as follows

$$\|\mathbf{a} - \mathbf{b}\| = \sqrt{(\mathbf{a} - \mathbf{b})^T(\mathbf{a} - \mathbf{b})} = \sqrt{\sum_{i=1}^m (a_i - b_i)^2} \quad (1.4)$$

Thus, the length of a vector is simply its distance from the zero vector  $\mathbf{0}$ , all of whose elements are 0, that is,  $\|\mathbf{a}\| = \|\mathbf{a} - \mathbf{0}\|$ .

From the general  $L_p$ -norm we can define the corresponding  $L_p$ -distance function, given as follows

$$\|\mathbf{a} - \mathbf{b}\|_p = \left( \sum_{i=1}^m |a_i - b_i|^p \right)^{\frac{1}{p}} \quad (1.5)$$

If  $p$  is unspecified, as in Eq. (1.4), it is assumed to be  $p = 2$  by default.

### Angle

The cosine of the smallest angle between vectors  $\mathbf{a}$  and  $\mathbf{b}$ , also called the *cosine similarity*, is given as

$$\cos\theta = \frac{\mathbf{a}^T \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} = \left( \frac{\mathbf{a}}{\|\mathbf{a}\|} \right)^T \left( \frac{\mathbf{b}}{\|\mathbf{b}\|} \right) \quad (1.6)$$

Thus, the cosine of the angle between  $\mathbf{a}$  and  $\mathbf{b}$  is given as the dot product of the unit vectors  $\frac{\mathbf{a}}{\|\mathbf{a}\|}$  and  $\frac{\mathbf{b}}{\|\mathbf{b}\|}$ .

The *Cauchy–Schwartz* inequality states that for any vectors  $\mathbf{a}$  and  $\mathbf{b}$  in  $\mathbb{R}^m$

$$|\mathbf{a}^T \mathbf{b}| \leq \|\mathbf{a}\| \cdot \|\mathbf{b}\|$$

It follows immediately from the Cauchy–Schwartz inequality that

$$-1 \leq \cos\theta \leq 1$$

Because the smallest angle  $\theta \in [0^\circ, 180^\circ]$  and because  $\cos\theta \in [-1, 1]$ , the cosine similarity value ranges from +1, corresponding to an angle of  $0^\circ$ , to -1, corresponding to an angle of  $180^\circ$  (or  $\pi$  radians).

### Orthogonality

Two vectors  $\mathbf{a}$  and  $\mathbf{b}$  are said to be *orthogonal* if and only if  $\mathbf{a}^T \mathbf{b} = 0$ , which in turn implies that  $\cos\theta = 0$ , that is, the angle between them is  $90^\circ$  or  $\frac{\pi}{2}$  radians. In this case, we say that they have no similarity.

**Example 1.3 (Distance and Angle).** Figure 1.3 shows the two vectors

$$\mathbf{a} = \begin{pmatrix} 5 \\ 3 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

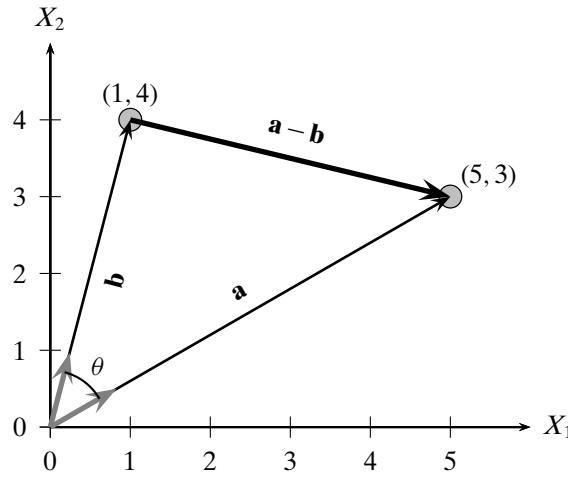


Figure 1.3. Distance and angle. Unit vectors are shown in gray.

Using Eq. (1.4), the Euclidean distance between them is given as

$$\|\mathbf{a} - \mathbf{b}\| = \sqrt{(5-1)^2 + (3-4)^2} = \sqrt{16+1} = \sqrt{17} = 4.12$$

The distance can also be computed as the magnitude of the vector:

$$\mathbf{a} - \mathbf{b} = \begin{pmatrix} 5 \\ 3 \end{pmatrix} - \begin{pmatrix} 1 \\ 4 \end{pmatrix} = \begin{pmatrix} 4 \\ -1 \end{pmatrix}$$

because  $\|\mathbf{a} - \mathbf{b}\| = \sqrt{4^2 + (-1)^2} = \sqrt{17} = 4.12$ .

The unit vector in the direction of  $\mathbf{a}$  is given as

$$\mathbf{u}_a = \frac{\mathbf{a}}{\|\mathbf{a}\|} = \frac{1}{\sqrt{5^2 + 3^2}} \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \frac{1}{\sqrt{34}} \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} 0.86 \\ 0.51 \end{pmatrix}$$

The unit vector in the direction of  $\mathbf{b}$  can be computed similarly:

$$\mathbf{u}_b = \begin{pmatrix} 0.24 \\ 0.97 \end{pmatrix}$$

These unit vectors are also shown in gray in Figure 1.3.

By Eq. (1.6) the cosine of the angle between  $\mathbf{a}$  and  $\mathbf{b}$  is given as

$$\cos \theta = \frac{\begin{pmatrix} 5 \\ 3 \end{pmatrix}^T \begin{pmatrix} 1 \\ 4 \end{pmatrix}}{\sqrt{5^2 + 3^2} \sqrt{1^2 + 4^2}} = \frac{17}{\sqrt{34 \times 17}} = \frac{1}{\sqrt{2}}$$

We can get the angle by computing the inverse of the cosine:

$$\theta = \cos^{-1}(1/\sqrt{2}) = 45^\circ$$

Let us consider the  $L_p$ -norm for  $\mathbf{a}$  with  $p = 3$ ; we get

$$\|\mathbf{a}\|_3 = (5^3 + 3^3)^{1/3} = (152)^{1/3} = 5.34$$

The distance between  $\mathbf{a}$  and  $\mathbf{b}$  using Eq. (1.5) for the  $L_p$ -norm with  $p = 3$  is given as

$$\|\mathbf{a} - \mathbf{b}\|_3 = \|(4, -1)^T\|_3 = (4^3 + |-1|^3)^{1/3} = (65)^{1/3} = 4.02$$

### 1.3.2 Mean and Total Variance

#### Mean

The *mean* of the data matrix  $\mathbf{D}$  is the vector obtained as the average of all the points:

$$\text{mean}(\mathbf{D}) = \boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad (1.7)$$

#### Total Variance

The *total variance* of the data matrix  $\mathbf{D}$  is the average squared distance of each point from the mean:

$$\text{var}(\mathbf{D}) = \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \quad (1.8)$$

Simplifying Eq. (1.8) we obtain

$$\begin{aligned} \text{var}(\mathbf{D}) &= \frac{1}{n} \sum_{i=1}^n (\|\mathbf{x}_i\|^2 - 2\mathbf{x}_i^T \boldsymbol{\mu} + \|\boldsymbol{\mu}\|^2) \\ &= \frac{1}{n} \left( \sum_{i=1}^n \|\mathbf{x}_i\|^2 - 2n\boldsymbol{\mu}^T \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \right) + n\|\boldsymbol{\mu}\|^2 \right) \\ &= \frac{1}{n} \left( \sum_{i=1}^n \|\mathbf{x}_i\|^2 - 2n\boldsymbol{\mu}^T \boldsymbol{\mu} + n\|\boldsymbol{\mu}\|^2 \right) \\ &= \frac{1}{n} \left( \sum_{i=1}^n \|\mathbf{x}_i\|^2 \right) - \|\boldsymbol{\mu}\|^2 \end{aligned}$$

The total variance is thus the difference between the average of the squared magnitude of the data points and the squared magnitude of the mean (average of the points).

#### Centered Data Matrix

Often we need to center the data matrix by making the mean coincide with the origin of the data space. The *centered data matrix* is obtained by subtracting the mean from

all the points:

$$\bar{\mathbf{D}} = \mathbf{D} - \mathbf{1} \cdot \boldsymbol{\mu}^T = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}^T \\ \boldsymbol{\mu}^T \\ \vdots \\ \boldsymbol{\mu}^T \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T - \boldsymbol{\mu}^T \\ \mathbf{x}_2^T - \boldsymbol{\mu}^T \\ \vdots \\ \mathbf{x}_n^T - \boldsymbol{\mu}^T \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{x}}_1^T \\ \bar{\mathbf{x}}_2^T \\ \vdots \\ \bar{\mathbf{x}}_n^T \end{pmatrix} \quad (1.9)$$

where  $\bar{\mathbf{x}}_i = \mathbf{x}_i - \boldsymbol{\mu}$  represents the centered point corresponding to  $\mathbf{x}_i$ , and  $\mathbf{1} \in \mathbb{R}^n$  is the  $n$ -dimensional vector all of whose elements have value 1. The mean of the centered data matrix  $\bar{\mathbf{D}}$  is  $\mathbf{0} \in \mathbb{R}^d$ , because we have subtracted the mean  $\boldsymbol{\mu}$  from all the points  $\mathbf{x}_i$ .

### 1.3.3 Orthogonal Projection

Often in data mining we need to project a point or vector onto another vector, for example, to obtain a new point after a change of the basis vectors. Let  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$  be two  $m$ -dimensional vectors. An *orthogonal decomposition* of the vector  $\mathbf{b}$  in the direction of another vector  $\mathbf{a}$ , illustrated in Figure 1.4, is given as

$$\mathbf{b} = \mathbf{b}_{\parallel} + \mathbf{b}_{\perp} = \mathbf{p} + \mathbf{r} \quad (1.10)$$

where  $\mathbf{p} = \mathbf{b}_{\parallel}$  is parallel to  $\mathbf{a}$ , and  $\mathbf{r} = \mathbf{b}_{\perp}$  is perpendicular or orthogonal to  $\mathbf{a}$ . The vector  $\mathbf{p}$  is called the *orthogonal projection* or simply projection of  $\mathbf{b}$  on the vector  $\mathbf{a}$ . Note that the point  $\mathbf{p} \in \mathbb{R}^m$  is the point closest to  $\mathbf{b}$  on the line passing through  $\mathbf{a}$ . Thus, the magnitude of the vector  $\mathbf{r} = \mathbf{b} - \mathbf{p}$  gives the *perpendicular distance* between  $\mathbf{b}$  and  $\mathbf{a}$ , which is often interpreted as the residual or error between the points  $\mathbf{b}$  and  $\mathbf{p}$ . The vector  $\mathbf{r}$  is also called the *error vector*.

We can derive an expression for  $\mathbf{p}$  by noting that  $\mathbf{p} = c\mathbf{a}$  for some scalar  $c$ , as  $\mathbf{p}$  is parallel to  $\mathbf{a}$ . Thus,  $\mathbf{r} = \mathbf{b} - \mathbf{p} = \mathbf{b} - c\mathbf{a}$ . Because  $\mathbf{p}$  and  $\mathbf{r}$  are orthogonal, we have

$$\mathbf{p}^T \mathbf{r} = (c\mathbf{a})^T (\mathbf{b} - c\mathbf{a}) = c\mathbf{a}^T \mathbf{b} - c^2 \mathbf{a}^T \mathbf{a} = 0$$

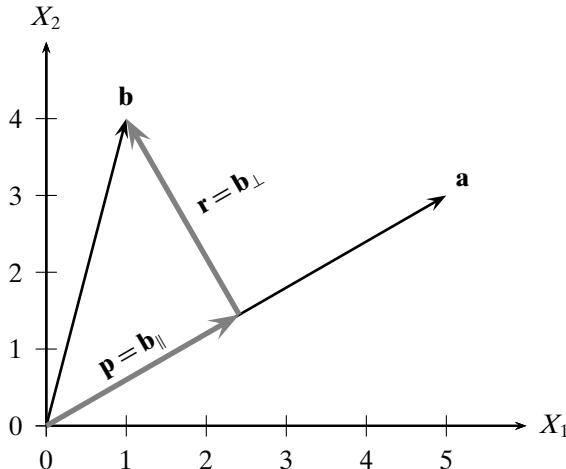


Figure 1.4. Orthogonal projection.

which implies that

$$c = \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}}$$

Therefore, the projection of  $\mathbf{b}$  on  $\mathbf{a}$  is given as

$$\mathbf{p} = c\mathbf{a} = \left( \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} \right) \mathbf{a} \quad (1.11)$$

The scalar offset  $c$  along  $\mathbf{a}$  is also called the *scalar projection* of  $\mathbf{b}$  on  $\mathbf{a}$ , denoted as

$$\text{proj}_{\mathbf{a}}(\mathbf{b}) = \left( \frac{\mathbf{b}^T \mathbf{a}}{\mathbf{a}^T \mathbf{a}} \right) \mathbf{a} \quad (1.12)$$

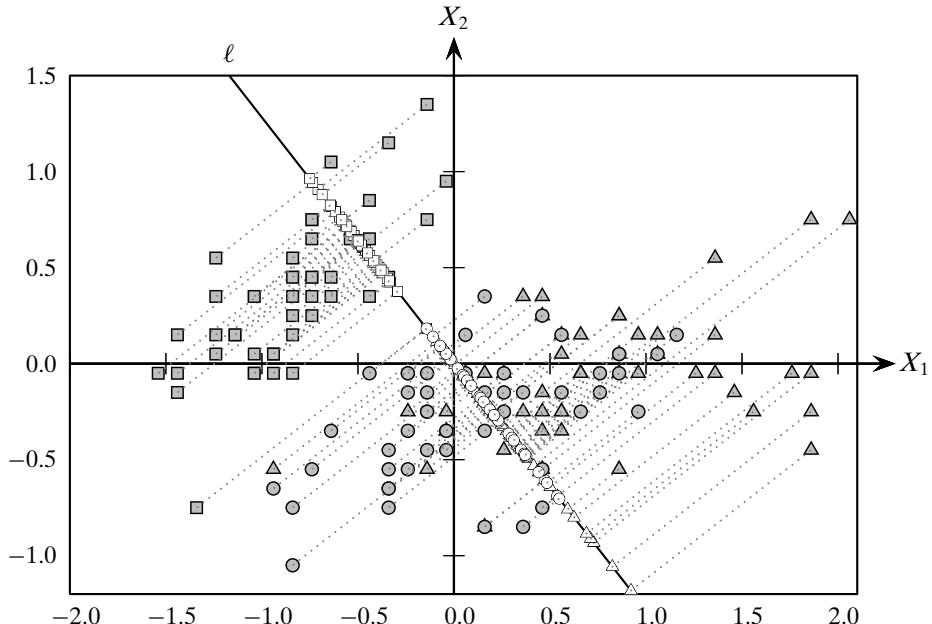
Therefore, the projection of  $\mathbf{b}$  on  $\mathbf{a}$  can also be written as

$$\mathbf{p} = \text{proj}_{\mathbf{a}}(\mathbf{b}) \cdot \mathbf{a}$$

**Example 1.4.** Restricting the Iris dataset to the first two dimensions, `sepal length` and `sepal width`, the mean point is given as

$$\text{mean}(\mathbf{D}) = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

which is shown as the black circle in [Figure 1.2](#). The corresponding centered data is shown in [Figure 1.5](#), and the total variance is  $\text{var}(\mathbf{D}) = 0.868$  (centering does not change this value).



[Figure 1.5.](#) Projecting the centered data onto the line  $\ell$ .

[Figure 1.5](#) shows the projection of each point onto the line  $\ell$ , which is the line that maximizes the separation between the class *iris-setosa* (squares) from the other two classes, namely *iris-versicolor* (circles) and *iris-virginica* (triangles). The line  $\ell$  is given as the set of all the points  $(x_1, x_2)^T$  satisfying the constraint  $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = c \begin{pmatrix} -2.15 \\ 2.75 \end{pmatrix}$  for all scalars  $c \in \mathbb{R}$ .

### 1.3.4 Linear Independence and Dimensionality

Given the data matrix

$$\mathbf{D} = (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_n)^T = (X_1 \quad X_2 \quad \cdots \quad X_d)$$

we are often interested in the linear combinations of the rows (points) or the columns (attributes). For instance, different linear combinations of the original  $d$  attributes yield new derived attributes, which play a key role in feature extraction and dimensionality reduction.

Given any set of vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$  in an  $m$ -dimensional vector space  $\mathbb{R}^m$ , their *linear combination* is given as

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k$$

where  $c_i \in \mathbb{R}$  are scalar values. The set of all possible linear combinations of the  $k$  vectors is called the *span*, denoted as  $\text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k)$ , which is itself a vector space being a *subspace* of  $\mathbb{R}^m$ . If  $\text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \mathbb{R}^m$ , then we say that  $\mathbf{v}_1, \dots, \mathbf{v}_k$  is a *spanning set* for  $\mathbb{R}^m$ .

#### Row and Column Space

There are several interesting vector spaces associated with the data matrix  $\mathbf{D}$ , two of which are the column space and row space of  $\mathbf{D}$ . The *column space* of  $\mathbf{D}$ , denoted  $\text{col}(\mathbf{D})$ , is the set of all linear combinations of the  $d$  attributes  $X_j \in \mathbb{R}^n$ , that is,

$$\text{col}(\mathbf{D}) = \text{span}(X_1, X_2, \dots, X_d)$$

By definition  $\text{col}(\mathbf{D})$  is a subspace of  $\mathbb{R}^n$ . The *row space* of  $\mathbf{D}$ , denoted  $\text{row}(\mathbf{D})$ , is the set of all linear combinations of the  $n$  points  $\mathbf{x}_i \in \mathbb{R}^d$ , that is,

$$\text{row}(\mathbf{D}) = \text{span}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$$

By definition  $\text{row}(\mathbf{D})$  is a subspace of  $\mathbb{R}^d$ . Note also that the row space of  $\mathbf{D}$  is the column space of  $\mathbf{D}^T$ :

$$\text{row}(\mathbf{D}) = \text{col}(\mathbf{D}^T)$$

#### Linear Independence

We say that the vectors  $\mathbf{v}_1, \dots, \mathbf{v}_k$  are *linearly dependent* if at least one vector can be written as a linear combination of the others. Alternatively, the  $k$  vectors are linearly

dependent if there are scalars  $c_1, c_2, \dots, c_k$ , at least one of which is not zero, such that

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = \mathbf{0}$$

On the other hand,  $\mathbf{v}_1, \dots, \mathbf{v}_k$  are *linearly independent* if and only if

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \text{ implies } c_1 = c_2 = \cdots = c_k = 0$$

Simply put, a set of vectors is linearly independent if none of them can be written as a linear combination of the other vectors in the set.

### Dimension and Rank

Let  $S$  be a subspace of  $\mathbb{R}^m$ . A *basis* for  $S$  is a set of vectors in  $S$ , say  $\mathbf{v}_1, \dots, \mathbf{v}_k$ , that are linearly independent and they span  $S$ , that is,  $\text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k) = S$ . In fact, a basis is a minimal spanning set. If the vectors in the basis are pairwise orthogonal, they are said to form an *orthogonal basis* for  $S$ . If, in addition, they are also normalized to be unit vectors, then they make up an *orthonormal basis* for  $S$ . For instance, the *standard basis* for  $\mathbb{R}^m$  is an orthonormal basis consisting of the vectors

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad \cdots \quad \mathbf{e}_m = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

Any two bases for  $S$  must have the same number of vectors, and the number of vectors in a basis for  $S$  is called the *dimension* of  $S$ , denoted as  $\dim(S)$ . Because  $S$  is a subspace of  $\mathbb{R}^m$ , we must have  $\dim(S) \leq m$ .

It is a remarkable fact that, for any matrix, the dimension of its row and column space is the same, and this dimension is also called the *rank* of the matrix. For the data matrix  $\mathbf{D} \in \mathbb{R}^{n \times d}$ , we have  $\text{rank}(\mathbf{D}) \leq \min(n, d)$ , which follows from the fact that the column space can have dimension at most  $d$ , and the row space can have dimension at most  $n$ . Thus, even though the data points are ostensibly in a  $d$  dimensional attribute space (the *extrinsic dimensionality*), if  $\text{rank}(\mathbf{D}) < d$ , then the data points reside in a lower dimensional subspace of  $\mathbb{R}^d$ , and in this case  $\text{rank}(\mathbf{D})$  gives an indication about the *intrinsic dimensionality* of the data. In fact, with dimensionality reduction methods it is often possible to approximate  $\mathbf{D} \in \mathbb{R}^{n \times d}$  with a derived data matrix  $\mathbf{D}' \in \mathbb{R}^{n \times k}$ , which has much lower dimensionality, that is,  $k \ll d$ . In this case  $k$  may reflect the “true” intrinsic dimensionality of the data.

**Example 1.5.** The line  $\ell$  in Figure 1.5 is given as  $\ell = \text{span}\left(\begin{pmatrix} -2.15 & 2.75 \end{pmatrix}^T\right)$ , with  $\dim(\ell) = 1$ . After normalization, we obtain the orthonormal basis for  $\ell$  as the unit vector

$$\frac{1}{\sqrt{12.19}} \begin{pmatrix} -2.15 \\ 2.75 \end{pmatrix} = \begin{pmatrix} -0.615 \\ 0.788 \end{pmatrix}$$

## 1.4 DATA: PROBABILISTIC VIEW

---

The probabilistic view of the data assumes that each numeric attribute  $X$  is a *random variable*, defined as a function that assigns a real number to each outcome of an experiment (i.e., some process of observation or measurement). Formally,  $X$  is a function  $X: \mathcal{O} \rightarrow \mathbb{R}$ , where  $\mathcal{O}$ , the domain of  $X$ , is the set of all possible outcomes of the experiment, also called the *sample space*, and  $\mathbb{R}$ , the *range* of  $X$ , is the set of real numbers. If the outcomes are numeric, and represent the observed values of the random variable, then  $X: \mathcal{O} \rightarrow \mathcal{O}$  is simply the identity function:  $X(v) = v$  for all  $v \in \mathcal{O}$ . The distinction between an outcome and a value of a random variable is important, as we may want to treat the observed values differently depending on the context, as seen in [Example 1.6](#).

A random variable  $X$  is called a *discrete random variable* if it takes on only a finite or countably infinite number of values in its range, whereas  $X$  is called a *continuous random variable* if it can take on any value in its range.

**Example 1.6.** Consider the sepal length attribute ( $X_1$ ) for the Iris dataset in [Table 1.1](#). All  $n = 150$  values of this attribute are shown in [Table 1.2](#), which lie in the range [4.3, 7.9], with centimeters as the unit of measurement. Let us assume that these constitute the set of all possible outcomes  $\mathcal{O}$ .

By default, we can consider the attribute  $X_1$  to be a continuous random variable, given as the identity function  $X_1(v) = v$ , because the outcomes (sepal length values) are all numeric.

On the other hand, if we want to distinguish between Iris flowers with short and long sepal lengths, with long being, say, a length of 7 cm or more, we can define a discrete random variable  $A$  as follows:

$$A(v) = \begin{cases} 0 & \text{if } v < 7 \\ 1 & \text{if } v \geq 7 \end{cases}$$

In this case the domain of  $A$  is [4.3, 7.9], and its range is {0, 1}. Thus,  $A$  assumes nonzero probability only at the discrete values 0 and 1.

### Probability Mass Function

If  $X$  is discrete, the *probability mass function* of  $X$  is defined as

$$f(x) = P(X = x) \quad \text{for all } x \in \mathbb{R} \tag{1.13}$$

In other words, the function  $f$  gives the probability  $P(X = x)$  that the random variable  $X$  has the exact value  $x$ . The name “probability mass function” intuitively conveys the fact that the probability is concentrated or massed at only discrete values in the range of  $X$ , and is zero for all other values.  $f$  must also obey the basic rules of probability. That is,  $f$  must be non-negative:

$$f(x) \geq 0$$

Table 1.2. Iris dataset: sepal length (in centimeters).

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 5.9 | 6.9 | 6.6 | 4.6 | 6.0 | 4.7 | 6.5 | 5.8 | 6.7 | 6.7 | 5.1 | 5.1 | 5.7 | 6.1 | 4.9 |
| 5.0 | 5.0 | 5.7 | 5.0 | 7.2 | 5.9 | 6.5 | 5.7 | 5.5 | 4.9 | 5.0 | 5.5 | 4.6 | 7.2 | 6.8 |
| 5.4 | 5.0 | 5.7 | 5.8 | 5.1 | 5.6 | 5.8 | 5.1 | 6.3 | 6.3 | 5.6 | 6.1 | 6.8 | 7.3 | 5.6 |
| 4.8 | 7.1 | 5.7 | 5.3 | 5.7 | 5.7 | 5.6 | 4.4 | 6.3 | 5.4 | 6.3 | 6.9 | 7.7 | 6.1 | 5.6 |
| 6.1 | 6.4 | 5.0 | 5.1 | 5.6 | 5.4 | 5.8 | 4.9 | 4.6 | 5.2 | 7.9 | 7.7 | 6.1 | 5.5 | 4.6 |
| 4.7 | 4.4 | 6.2 | 4.8 | 6.0 | 6.2 | 5.0 | 6.4 | 6.3 | 6.7 | 5.0 | 5.9 | 6.7 | 5.4 | 6.3 |
| 4.8 | 4.4 | 6.4 | 6.2 | 6.0 | 7.4 | 4.9 | 7.0 | 5.5 | 6.3 | 6.8 | 6.1 | 6.5 | 6.7 | 6.7 |
| 4.8 | 4.9 | 6.9 | 4.5 | 4.3 | 5.2 | 5.0 | 6.4 | 5.2 | 5.8 | 5.5 | 7.6 | 6.3 | 6.4 | 6.3 |
| 5.8 | 5.0 | 6.7 | 6.0 | 5.1 | 4.8 | 5.7 | 5.1 | 6.6 | 6.4 | 5.2 | 6.4 | 7.7 | 5.8 | 4.9 |
| 5.4 | 5.1 | 6.0 | 6.5 | 5.5 | 7.2 | 6.9 | 6.2 | 6.5 | 6.0 | 5.4 | 5.5 | 6.7 | 7.7 | 5.1 |

and the sum of all probabilities should add to 1:

$$\sum_x f(x) = 1$$

**Example 1.7 (Bernoulli and Binomial Distribution).** In [Example 1.6](#),  $A$  was defined as a discrete random variable representing long sepal length. From the sepal length data in [Table 1.2](#) we find that only 13 Irises have sepal length of at least 7 cm. We can thus estimate the probability mass function of  $A$  as follows:

$$f(1) = P(A = 1) = \frac{13}{150} = 0.087 = p$$

and

$$f(0) = P(A = 0) = \frac{137}{150} = 0.913 = 1 - p$$

In this case we say that  $A$  has a *Bernoulli distribution* with parameter  $p \in [0, 1]$ , which denotes the probability of a *success*, that is, the probability of picking an Iris with a long sepal length at random from the set of all points. On the other hand,  $1 - p$  is the probability of a *failure*, that is, of not picking an Iris with long sepal length.

Let us consider another discrete random variable  $B$ , denoting the number of Irises with long sepal length in  $m$  independent Bernoulli trials with probability of success  $p$ . In this case,  $B$  takes on the discrete values  $[0, m]$ , and its probability mass function is given by the *binomial distribution*

$$f(k) = P(B = k) = \binom{m}{k} p^k (1 - p)^{m-k}$$

The formula can be understood as follows. There are  $\binom{m}{k}$  ways of picking  $k$  long sepal length Irises out of the  $m$  trials. For each selection of  $k$  long sepal length Irises, the total probability of the  $k$  successes is  $p^k$ , and the total probability of  $m - k$  failures is  $(1 - p)^{m-k}$ . For example, because  $p = 0.087$  from above, the probability of observing exactly  $k = 2$  Irises with long sepal length in  $m = 10$  trials is given as

$$f(2) = P(B = 2) = \binom{10}{2} (0.087)^2 (0.913)^8 = 0.164$$

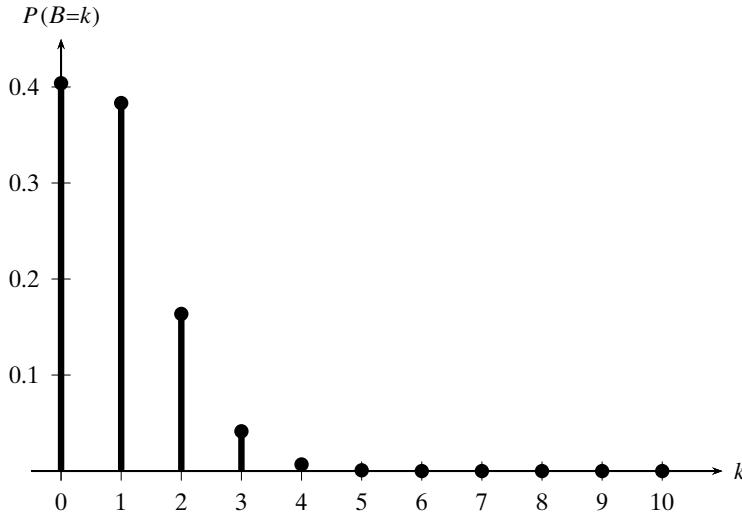


Figure 1.6. Binomial distribution: probability mass function ( $m = 10, p = 0.087$ ).

Figure 1.6 shows the full probability mass function for different values of  $k$  for  $m = 10$ . Because  $p$  is quite small, the probability of  $k$  successes in so few trials falls off rapidly as  $k$  increases, becoming practically zero for values of  $k \geq 6$ .

### Probability Density Function

If  $X$  is continuous, its range is the entire set of real numbers  $\mathbb{R}$ . The probability of any specific value  $x$  is only one out of the infinitely many possible values in the range of  $X$ , which means that  $P(X = x) = 0$  for all  $x \in \mathbb{R}$ . However, this does not mean that the value  $x$  is impossible, because in that case we would conclude that all values are impossible! What it means is that the probability mass is spread so thinly over the range of values that it can be measured only over intervals  $[a, b] \subset \mathbb{R}$ , rather than at specific points. Thus, instead of the probability mass function, we define the *probability density function*, which specifies the probability that the variable  $X$  takes on values in any interval  $[a, b] \subset \mathbb{R}$ :

$$P(X \in [a, b]) = \int_a^b f(x) dx \quad (1.14)$$

As before, the density function  $f$  must satisfy the basic laws of probability:

$$f(x) \geq 0, \quad \text{for all } x \in \mathbb{R}$$

and

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

We can get an intuitive understanding of the density function  $f$  by considering the probability density over a small interval of width  $2\epsilon > 0$ , centered at  $x$ , namely

$[x - \epsilon, x + \epsilon]$ :

$$\begin{aligned} P(X \in [x - \epsilon, x + \epsilon]) &= \int_{x-\epsilon}^{x+\epsilon} f(x) dx \simeq 2\epsilon \cdot f(x) \\ f(x) &\simeq \frac{P(X \in [x - \epsilon, x + \epsilon])}{2\epsilon} \end{aligned} \quad (1.15)$$

$f(x)$  thus gives the probability density at  $x$ , given as the ratio of the probability mass to the width of the interval, that is, the probability mass per unit distance. Thus, it is important to note that  $P(X = x) \neq f(x)$ .

Even though the probability density function  $f(x)$  does not specify the probability  $P(X = x)$ , it can be used to obtain the relative probability of one value  $x_1$  over another  $x_2$  because for a given  $\epsilon > 0$ , by Eq.(1.15), we have

$$\frac{P(X \in [x_1 - \epsilon, x_1 + \epsilon])}{P(X \in [x_2 - \epsilon, x_2 + \epsilon])} \simeq \frac{2\epsilon \cdot f(x_1)}{2\epsilon \cdot f(x_2)} = \frac{f(x_1)}{f(x_2)} \quad (1.16)$$

Thus, if  $f(x_1)$  is larger than  $f(x_2)$ , then values of  $X$  close to  $x_1$  are more probable than values close to  $x_2$ , and vice versa.

**Example 1.8 (Normal Distribution).** Consider again the `sepal_length` values from the Iris dataset, as shown in Table 1.2. Let us assume that these values follow a *Gaussian* or *normal* density function, given as

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

There are two parameters of the normal density distribution, namely,  $\mu$ , which represents the mean value, and  $\sigma^2$ , which represents the variance of the values (these parameters are discussed in Chapter 2). Figure 1.7 shows the characteristic “bell” shape plot of the normal distribution. The parameters,  $\mu = 5.84$  and  $\sigma^2 = 0.681$ , were estimated directly from the data for `sepal_length` in Table 1.2.

Whereas  $f(x = \mu) = f(5.84) = \frac{1}{\sqrt{2\pi \cdot 0.681}} \exp\{0\} = 0.483$ , we emphasize that the probability of observing  $X = \mu$  is zero, that is,  $P(X = \mu) = 0$ . Thus,  $P(X = x)$  is not given by  $f(x)$ , rather,  $P(X = x)$  is given as the area under the curve for an infinitesimally small interval  $[x - \epsilon, x + \epsilon]$  centered at  $x$ , with  $\epsilon > 0$ . Figure 1.7 illustrates this with the shaded region centered at  $\mu = 5.84$ . From Eq.(1.15), we have

$$P(X = \mu) \simeq 2\epsilon \cdot f(\mu) = 2\epsilon \cdot 0.483 = 0.967\epsilon$$

As  $\epsilon \rightarrow 0$ , we get  $P(X = \mu) \rightarrow 0$ . However, based on Eq.(1.16) we can claim that the probability of observing values close to the mean value  $\mu = 5.84$  is 2.69 times the probability of observing values close to  $x = 7$ , as

$$\frac{f(5.84)}{f(7)} = \frac{0.483}{0.18} = 2.69$$

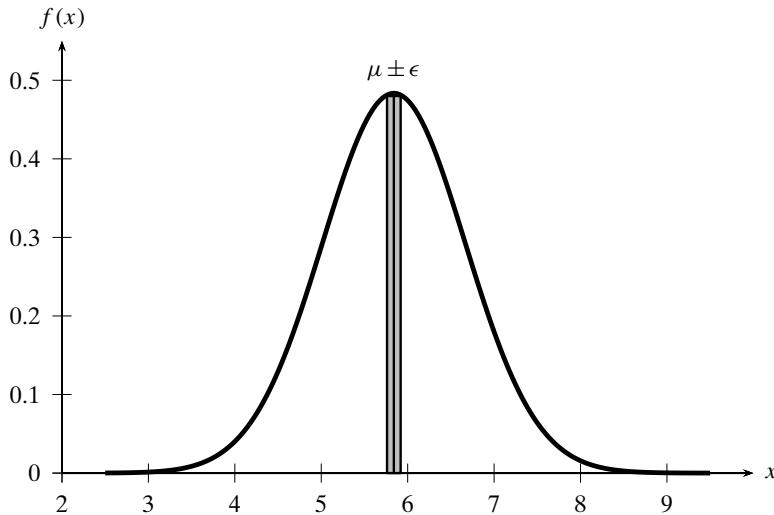


Figure 1.7. Normal distribution: probability density function ( $\mu = 5.84, \sigma^2 = 0.681$ ).

### Cumulative Distribution Function

For any random variable  $X$ , whether discrete or continuous, we can define the *cumulative distribution function (CDF)*  $F : \mathbb{R} \rightarrow [0, 1]$ , which gives the probability of observing a value at most some given value  $x$ :

$$F(x) = P(X \leq x) \quad \text{for all } -\infty < x < \infty \quad (1.17)$$

When  $X$  is discrete,  $F$  is given as

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u)$$

and when  $X$  is continuous,  $F$  is given as

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$$

**Example 1.9 (Cumulative Distribution Function).** Figure 1.8 shows the cumulative distribution function for the binomial distribution in Figure 1.6. It has the characteristic step shape (right continuous, non-decreasing), as expected for a discrete random variable.  $F(x)$  has the same value  $F(k)$  for all  $x \in [k, k + 1)$  with  $0 \leq k < m$ , where  $m$  is the number of trials and  $k$  is the number of successes. The closed (filled) and open circles demarcate the corresponding closed and open interval  $[k, k + 1)$ . For instance,  $F(x) = 0.404 = F(0)$  for all  $x \in [0, 1)$ .

Figure 1.9 shows the cumulative distribution function for the normal density function shown in Figure 1.7. As expected, for a continuous random variable, the CDF is also continuous, and non-decreasing. Because the normal distribution is symmetric about the mean, we have  $F(\mu) = P(X \leq \mu) = 0.5$ .

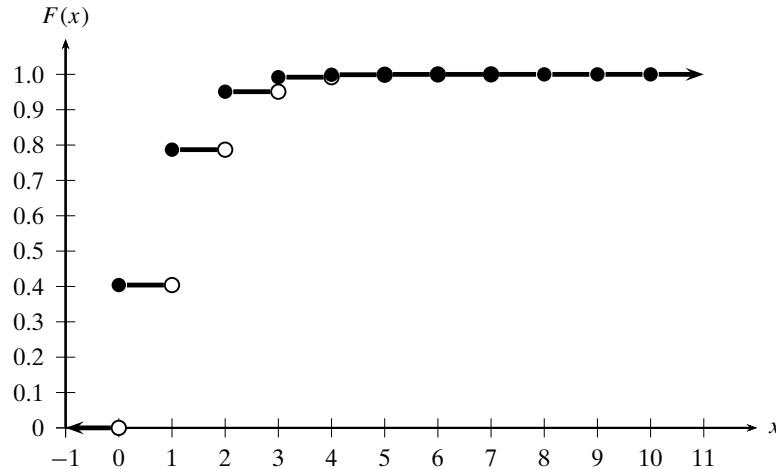


Figure 1.8. Cumulative distribution function for the binomial distribution.

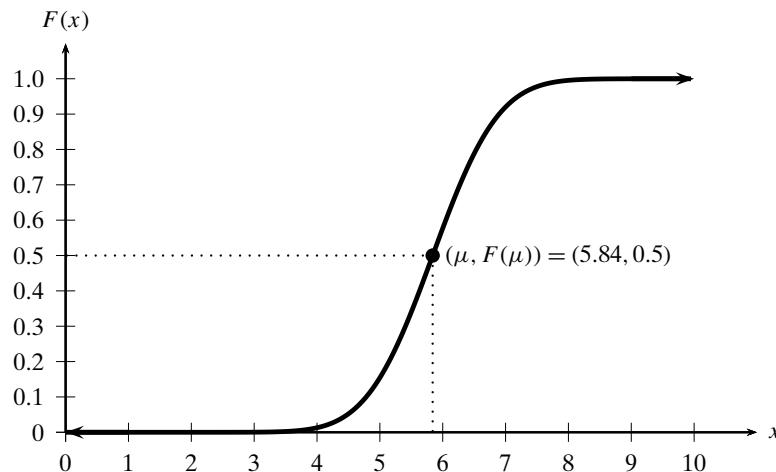


Figure 1.9. Cumulative distribution function for the normal distribution.

### 1.4.1 Bivariate Random Variables

Instead of considering each attribute as a random variable, we can also perform pair-wise analysis by considering a pair of attributes,  $X_1$  and  $X_2$ , as a *bivariate random variable*:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$\mathbf{X} : \mathcal{O} \rightarrow \mathbb{R}^2$  is a function that assigns to each outcome in the sample space, a pair of real numbers, that is, a 2-dimensional vector  $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$ . As in the univariate case, if the outcomes are numeric, then the default is to assume  $\mathbf{X}$  to be the identity function.

### Joint Probability Mass Function

If  $X_1$  and  $X_2$  are both discrete random variables then  $\mathbf{X}$  has a *joint probability mass function* given as follows:

$$f(\mathbf{x}) = f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = P(\mathbf{X} = \mathbf{x})$$

$f$  must satisfy the following two conditions:

$$f(\mathbf{x}) = f(x_1, x_2) \geq 0 \quad \text{for all } -\infty < x_1, x_2 < \infty$$

$$\sum_{\mathbf{x}} f(\mathbf{x}) = \sum_{x_1} \sum_{x_2} f(x_1, x_2) = 1$$

### Joint Probability Density Function

If  $X_1$  and  $X_2$  are both continuous random variables then  $\mathbf{X}$  has a *joint probability density function*  $f$  given as follows:

$$P(\mathbf{X} \in W) = \int_{\mathbf{x} \in W} \int f(\mathbf{x}) d\mathbf{x} = \int_{(x_1, x_2)^T \in W} \int f(x_1, x_2) dx_1 dx_2$$

where  $W \subset \mathbb{R}^2$  is some subset of the 2-dimensional space of reals.  $f$  must also satisfy the following two conditions:

$$f(\mathbf{x}) = f(x_1, x_2) \geq 0 \quad \text{for all } -\infty < x_1, x_2 < \infty$$

$$\int_{\mathbb{R}^2} f(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = 1$$

As in the univariate case, the probability mass  $P(\mathbf{x}) = P((x_1, x_2)^T) = 0$  for any particular point  $\mathbf{x}$ . However, we can use  $f$  to compute the probability density at  $\mathbf{x}$ . Consider the square region  $W = ([x_1 - \epsilon, x_1 + \epsilon], [x_2 - \epsilon, x_2 + \epsilon])$ , that is, a 2-dimensional window of width  $2\epsilon$  centered at  $\mathbf{x} = (x_1, x_2)^T$ . The probability density at  $\mathbf{x}$  can be approximated as

$$\begin{aligned} P(\mathbf{X} \in W) &= P\left(\mathbf{X} \in ([x_1 - \epsilon, x_1 + \epsilon], [x_2 - \epsilon, x_2 + \epsilon])\right) \\ &= \int_{x_1 - \epsilon}^{x_1 + \epsilon} \int_{x_2 - \epsilon}^{x_2 + \epsilon} f(x_1, x_2) dx_1 dx_2 \\ &\simeq 2\epsilon \cdot 2\epsilon \cdot f(x_1, x_2) \end{aligned}$$

which implies that

$$f(x_1, x_2) = \frac{P(\mathbf{X} \in W)}{(2\epsilon)^2}$$

The relative probability of one value  $(a_1, a_2)$  versus another  $(b_1, b_2)$  can therefore be computed via the probability density function:

$$\frac{P(\mathbf{X} \in ([a_1 - \epsilon, a_1 + \epsilon], [a_2 - \epsilon, a_2 + \epsilon])))}{P(\mathbf{X} \in ([b_1 - \epsilon, b_1 + \epsilon], [b_2 - \epsilon, b_2 + \epsilon])))} \simeq \frac{(2\epsilon)^2 \cdot f(a_1, a_2)}{(2\epsilon)^2 \cdot f(b_1, b_2)} = \frac{f(a_1, a_2)}{f(b_1, b_2)}$$

**Example 1.10 (Bivariate Distributions).** Consider the sepal length and sepal width attributes in the Iris dataset, plotted in Figure 1.2. Let  $A$  denote the Bernoulli random variable corresponding to long sepal length (at least 7 cm), as defined in Example 1.7.

Define another Bernoulli random variable  $B$  corresponding to long sepal width, say, at least 3.5 cm. Let  $\mathbf{X} = \begin{pmatrix} A \\ B \end{pmatrix}$  be a discrete bivariate random variable; then the joint probability mass function of  $\mathbf{X}$  can be estimated from the data as follows:

$$f(0, 0) = P(A = 0, B = 0) = \frac{116}{150} = 0.773$$

$$f(0, 1) = P(A = 0, B = 1) = \frac{21}{150} = 0.140$$

$$f(1, 0) = P(A = 1, B = 0) = \frac{10}{150} = 0.067$$

$$f(1, 1) = P(A = 1, B = 1) = \frac{3}{150} = 0.020$$

Figure 1.10 shows a plot of this probability mass function.

Treating attributes  $X_1$  and  $X_2$  in the Iris dataset (see Table 1.1) as continuous random variables, we can define a continuous bivariate random variable  $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ . Assuming that  $\mathbf{X}$  follows a *bivariate normal distribution*, its joint probability density function is given as

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}\right\}$$

Here  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are the parameters of the bivariate normal distribution, representing the 2-dimensional mean vector and covariance matrix, which are discussed in detail

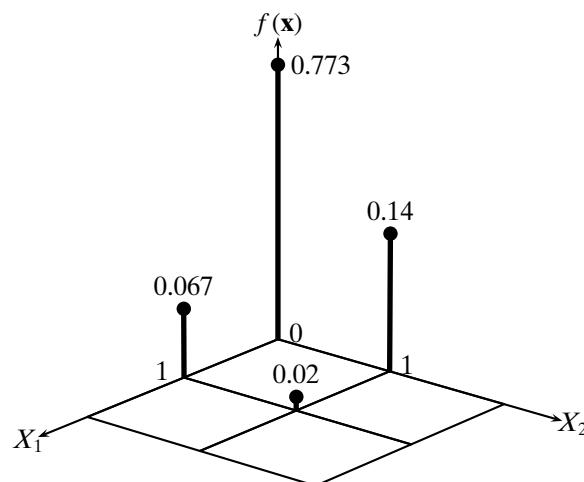


Figure 1.10. Joint probability mass function:  $X_1$  (long sepal length),  $X_2$  (long sepal width).

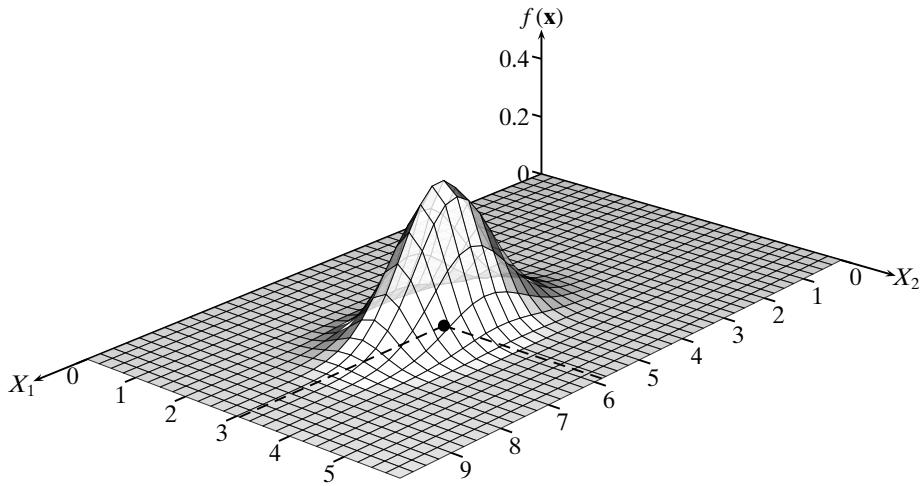


Figure 1.11. Bivariate normal density:  $\mu = (5.843, 3.054)^T$  (solid circle).

in Chapter 2. Further,  $|\Sigma|$  denotes the determinant of  $\Sigma$ . The plot of the bivariate normal density is given in Figure 1.11, with mean

$$\mu = (5.843, 3.054)^T$$

and covariance matrix

$$\Sigma = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

It is important to emphasize that the function  $f(\mathbf{x})$  specifies only the probability density at  $\mathbf{x}$ , and  $f(\mathbf{x}) \neq P(\mathbf{X} = \mathbf{x})$ . As before, we have  $P(\mathbf{X} = \mathbf{x}) = 0$ .

### Joint Cumulative Distribution Function

The *joint cumulative distribution function* for two random variables  $X_1$  and  $X_2$  is defined as the function  $F$ , such that for all values  $x_1, x_2 \in (-\infty, \infty)$ ,

$$F(\mathbf{x}) = F(x_1, x_2) = P(X_1 \leq x_1 \text{ and } X_2 \leq x_2) = P(\mathbf{X} \leq \mathbf{x})$$

### Statistical Independence

Two random variables  $X_1$  and  $X_2$  are said to be (statistically) *independent* if, for every  $W_1 \subset \mathbb{R}$  and  $W_2 \subset \mathbb{R}$ , we have

$$P(X_1 \in W_1 \text{ and } X_2 \in W_2) = P(X_1 \in W_1) \cdot P(X_2 \in W_2)$$

Furthermore, if  $X_1$  and  $X_2$  are independent, then the following two conditions are also satisfied:

$$F(\mathbf{x}) = F(x_1, x_2) = F_1(x_1) \cdot F_2(x_2)$$

$$f(\mathbf{x}) = f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$$

where  $F_i$  is the cumulative distribution function, and  $f_i$  is the probability mass or density function for random variable  $X_i$ .

### 1.4.2 Multivariate Random Variable

A  $d$ -dimensional *multivariate random variable*  $\mathbf{X} = (X_1, X_2, \dots, X_d)^T$ , also called a *vector random variable*, is defined as a function that assigns a vector of real numbers to each outcome in the sample space, that is,  $\mathbf{X} : \mathcal{O} \rightarrow \mathbb{R}^d$ . The range of  $\mathbf{X}$  can be denoted as a vector  $\mathbf{x} = (x_1, x_2, \dots, x_d)^T$ . In case all  $X_j$  are numeric, then  $\mathbf{X}$  is by default assumed to be the identity function. In other words, if all attributes are numeric, we can treat each outcome in the sample space (i.e., each point in the data matrix) as a vector random variable. On the other hand, if the attributes are not all numeric, then  $\mathbf{X}$  maps the outcomes to numeric vectors in its range.

If all  $X_j$  are discrete, then  $\mathbf{X}$  is jointly discrete and its joint probability mass function  $f$  is given as

$$\begin{aligned} f(\mathbf{x}) &= P(\mathbf{X} = \mathbf{x}) \\ f(x_1, x_2, \dots, x_d) &= P(X_1 = x_1, X_2 = x_2, \dots, X_d = x_d) \end{aligned}$$

If all  $X_j$  are continuous, then  $\mathbf{X}$  is jointly continuous and its joint probability density function is given as

$$\begin{aligned} P(\mathbf{X} \in W) &= \int_{\mathbf{x} \in W} \dots \int f(\mathbf{x}) d\mathbf{x} \\ P((X_1, X_2, \dots, X_d)^T \in W) &= \int_{(x_1, x_2, \dots, x_d)^T \in W} \dots \int f(x_1, x_2, \dots, x_d) dx_1 dx_2 \dots dx_d \end{aligned}$$

for any  $d$ -dimensional region  $W \subseteq \mathbb{R}^d$ .

The laws of probability must be obeyed as usual, that is,  $f(\mathbf{x}) \geq 0$  and sum of  $f$  over all  $\mathbf{x}$  in the range of  $\mathbf{X}$  must be 1. The joint cumulative distribution function of  $\mathbf{X} = (X_1, \dots, X_d)^T$  is given as

$$\begin{aligned} F(\mathbf{x}) &= P(\mathbf{X} \leq \mathbf{x}) \\ F(x_1, x_2, \dots, x_d) &= P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_d \leq x_d) \end{aligned}$$

for every point  $\mathbf{x} \in \mathbb{R}^d$ .

We say that  $X_1, X_2, \dots, X_d$  are independent random variables if and only if, for every region  $W_i \subset \mathbb{R}$ , we have

$$\begin{aligned} P(X_1 \in W_1 \text{ and } X_2 \in W_2 \dots \text{ and } X_d \in W_d) \\ = P(X_1 \in W_1) \cdot P(X_2 \in W_2) \dots \cdot P(X_d \in W_d) \quad (1.18) \end{aligned}$$

If  $X_1, X_2, \dots, X_d$  are independent then the following conditions are also satisfied

$$\begin{aligned} F(\mathbf{x}) &= F(x_1, \dots, x_d) = F_1(x_1) \cdot F_2(x_2) \cdot \dots \cdot F_d(x_d) \\ f(\mathbf{x}) &= f(x_1, \dots, x_d) = f_1(x_1) \cdot f_2(x_2) \cdot \dots \cdot f_d(x_d) \quad (1.19) \end{aligned}$$

where  $F_i$  is the cumulative distribution function, and  $f_i$  is the probability mass or density function for random variable  $X_i$ .

### 1.4.3 Random Sample and Statistics

The probability mass or density function of a random variable  $X$  may follow some known form, or as is often the case in data analysis, it may be unknown. When the probability function is not known, it may still be convenient to assume that the values follow some known distribution, based on the characteristics of the data. However, even in this case, the parameters of the distribution may still be unknown. Thus, in general, either the parameters, or the entire distribution, may have to be estimated from the data.

In statistics, the word *population* is used to refer to the set or universe of all entities under study. Usually we are interested in certain characteristics or parameters of the entire population (e.g., the mean age of all computer science students in the United States). However, looking at the entire population may not be feasible or may be too expensive. Instead, we try to make inferences about the population parameters by drawing a random sample from the population, and by computing appropriate *statistics* from the sample that give estimates of the corresponding population parameters of interest.

#### Univariate Sample

Given a random variable  $X$ , a *random sample* of size  $n$  from  $X$  is defined as a set of  $n$  *independent and identically distributed (IID)* random variables  $S_1, S_2, \dots, S_n$ , that is, all of the  $S_i$ 's are statistically independent of each other, and follow the same probability mass or density function as  $X$ .

If we treat attribute  $X$  as a random variable, then each of the observed values of  $X$ , namely,  $x_i$  ( $1 \leq i \leq n$ ), are themselves treated as identity random variables, and the observed data is assumed to be a random sample drawn from  $X$ . That is, all  $x_i$  are considered to be mutually independent and identically distributed as  $X$ . By Eq. (1.19) their joint probability function is given as

$$f(x_1, \dots, x_n) = \prod_{i=1}^n f_X(x_i) \quad (1.20)$$

where  $f_X$  is the probability mass or density function for  $X$ .

#### Multivariate Sample

For multivariate parameter estimation, the  $n$  data points  $\mathbf{x}_i$  (with  $1 \leq i \leq n$ ) constitute a  $d$ -dimensional multivariate random sample drawn from the vector random variable  $\mathbf{X} = (X_1, X_2, \dots, X_d)$ . That is,  $\mathbf{x}_i$  are assumed to be independent and identically distributed, and thus their joint distribution is given as

$$f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \prod_{i=1}^n f_{\mathbf{X}}(\mathbf{x}_i) \quad (1.21)$$

where  $f_{\mathbf{X}}$  is the probability mass or density function for  $\mathbf{X}$ .

Estimating the parameters of a multivariate joint probability distribution is usually difficult and computationally intensive. One simplifying assumption that is typically made is that the  $d$  attributes  $X_1, X_2, \dots, X_d$  are statistically independent. However, we do not assume that they are identically distributed, because that is almost never justified. Under the attribute independence assumption Eq. (1.21) can be rewritten as

$$f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \prod_{i=1}^n f(\mathbf{x}_i) = \prod_{i=1}^n \prod_{j=1}^d f_{X_j}(x_{ij})$$

### Statistic

We can estimate a parameter of the population by defining an appropriate sample *statistic*, which is defined as a function of the sample. More precisely, let  $\{\mathbf{S}_i\}_{i=1}^m$  denote a random sample of size  $m$  drawn from a (multivariate) random variable  $\mathbf{X}$ . A statistic  $\hat{\theta}$  is some function over the random sample, given as

$$\hat{\theta}: (\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_m) \rightarrow \mathbb{R}$$

The statistic  $\hat{\theta}$  is an estimate of the corresponding population parameter  $\theta$ . As such, the statistic  $\hat{\theta}$  is itself a random variable. If we use the value of a statistic to estimate a population parameter, this value is called a *point estimate* of the parameter, and the statistic is called an *estimator* of the parameter. In Chapter 2 we will study different estimators for population parameters that reflect the location (or centrality) and dispersion of values.

**Example 1.11 (Sample Mean).** Consider attribute `sepal length` ( $X_1$ ) in the Iris dataset, whose values are shown in Table 1.2. Assume that the mean value of  $X_1$  is not known. Let us assume that the observed values  $\{x_i\}_{i=1}^n$  constitute a random sample drawn from  $X_1$ .

The *sample mean* is a statistic, defined as the average

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

Plugging in values from Table 1.2, we obtain

$$\hat{\mu} = \frac{1}{150} (5.9 + 6.9 + \dots + 7.7 + 5.1) = \frac{876.5}{150} = 5.84$$

The value  $\hat{\mu} = 5.84$  is a point estimate for the unknown population parameter  $\mu$ , the (true) mean value of variable  $X_1$ .

## 1.5 FURTHER READING

---

For a review of the linear algebra concepts see Strang (2006) and Poole (2010), and for the probabilistic view see Evans and Rosenthal (2011). There are several good books on data mining, and machine and statistical learning; these include Hand, Mannila, and Smyth (2001), Han, Kamber, and Pei (2006), Witten, Frank, and Hall (2011), Tan, Steinbach, and Kumar (2013), and Bishop (2006) and Hastie, Tibshirani, and Friedman (2009).

- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer Science + Business Media.
- Evans, M. and Rosenthal, J. (2011). *Probability and Statistics: The Science of Uncertainty*. 2nd ed. New York: W. H. Freeman.
- Han, J., Kamber, M., and Pei, J. (2006). *Data Mining: Concepts and Techniques*. 2nd ed. The Morgan Kaufmann Series in Data Management Systems. Philadelphia: Elsevier Science.
- Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Adaptive Computation and Machine Learning Series. Cambridge, MA: MIT Press.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd ed. Springer Series in Statistics. New York: Springer Science + Business Media.
- Poole, D. (2010). *Linear Algebra: A Modern Introduction*. 3rd ed. Independence, KY: Cengage Learning.
- Strang, G. (2006). *Linear Algebra and Its Applications*. 4th ed. Independence, KY: Thomson Brooks/Cole, Cengage learning.
- Tan, P., Steinbach, M., and Kumar, V. (2013). *Introduction to Data Mining*. 2nd ed. Upper Saddle River, NJ: Prentice Hall.
- Witten, I., Frank, E., and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. The Morgan Kaufmann Series in Data Management Systems. Philadelphia: Elsevier Science.

## 1.6 EXERCISES

---

**Q1.** Show that the mean of the centered data matrix  $\bar{\mathbf{D}}$  in Eq. (1.9) is  $\mathbf{0}$ .

**Q2.** Prove that for the  $L_p$ -distance in Eq. (1.5), for  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , we have

$$\|\mathbf{x} - \mathbf{y}\|_\infty = \lim_{p \rightarrow \infty} \|\mathbf{x} - \mathbf{y}\|_p = \max_{i=1}^d \{|x_i - y_i|\}$$

In this chapter, we discuss basic statistical methods for exploratory data analysis of numeric attributes. We look at measures of central tendency or location, measures of dispersion, and measures of linear dependence or association between attributes. We emphasize the connection between the probabilistic and the geometric and algebraic views of the data matrix.

## 2.1 UNIVARIATE ANALYSIS

---

Univariate analysis focuses on a single attribute at a time; thus the data matrix  $\mathbf{D}$  can be thought of as an  $n \times 1$  matrix, or simply a column vector, given as

$$\mathbf{D} = \begin{pmatrix} X \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where  $X$  is the numeric attribute of interest, with  $x_i \in \mathbb{R}$ .  $X$  is assumed to be a random variable, with each point  $x_i$  ( $1 \leq i \leq n$ ) itself treated as an identity random variable. We assume that the observed data is a random sample drawn from  $X$ , that is, each variable  $x_i$  is independent and identically distributed as  $X$ . In the vector view, we treat the sample as an  $n$ -dimensional vector, and write  $X \in \mathbb{R}^n$ .

In general, the probability density or mass function  $f(x)$  and the cumulative distribution function  $F(x)$ , for attribute  $X$ , are both unknown. However, we can estimate these distributions directly from the data sample, which also allow us to compute statistics to estimate several important population parameters.

### Empirical Cumulative Distribution Function

The *empirical cumulative distribution function (CDF)* of  $X$  is given as

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x) \quad (2.1)$$

where

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases}$$

is a binary *indicator variable* that indicates whether the given condition is satisfied or not. Intuitively, to obtain the empirical CDF we compute, for each value  $x \in \mathbb{R}$ , how many points in the sample are less than or equal to  $x$ . The empirical CDF puts a probability mass of  $\frac{1}{n}$  at each point  $x_i$ . Note that we use the notation  $\hat{F}$  to denote the fact that the empirical CDF is an estimate for the unknown population CDF  $F$ .

### Inverse Cumulative Distribution Function

Define the *inverse cumulative distribution function* or *quantile function* for a random variable  $X$  as follows:

$$F^{-1}(q) = \min\{x \mid F(x) \geq q\} \quad \text{for } q \in [0, 1] \quad (2.2)$$

That is, the inverse CDF gives the least value of  $X$ , for which  $q$  fraction of the values are lower, and  $1 - q$  fraction of the values are higher. The *empirical inverse cumulative distribution function*  $\hat{F}^{-1}$  can be obtained from Eq. (2.1).

### Empirical Probability Mass Function

The *empirical probability mass function (PMF)* of  $X$  is given as

$$\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^n I(x_i = x) \quad (2.3)$$

where

$$I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}$$

The empirical PMF also puts a probability mass of  $\frac{1}{n}$  at each point  $x_i$ .

#### 2.1.1 Measures of Central Tendency

These measures give an indication about the concentration of the probability mass, the “middle” values, and so on.

##### Mean

The *mean*, also called the *expected value*, of a random variable  $X$  is the arithmetic average of the values of  $X$ . It provides a one-number summary of the *location* or *central tendency* for the distribution of  $X$ .

The mean or expected value of a discrete random variable  $X$  is defined as

$$\mu = E[X] = \sum_x x \cdot f(x) \quad (2.4)$$

where  $f(x)$  is the probability mass function of  $X$ .

The expected value of a continuous random variable  $X$  is defined as

$$\mu = E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx \quad (2.5)$$

where  $f(x)$  is the probability density function of  $X$ .

**Sample Mean** The *sample mean* is a statistic, that is, a function  $\hat{\mu} : \{x_1, x_2, \dots, x_n\} \rightarrow \mathbb{R}$ , defined as the average value of  $x_i$ 's:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.6)$$

It serves as an estimator for the unknown mean value  $\mu$  of  $X$ . It can be derived by plugging in the empirical PMF  $\hat{f}(x)$  in Eq. (2.4):

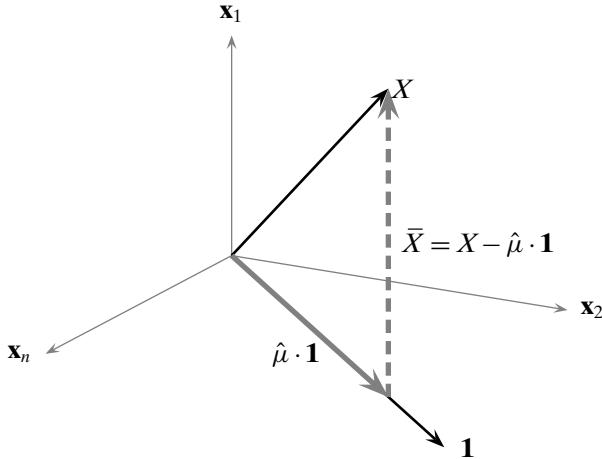
$$\hat{\mu} = \sum_x x \cdot \hat{f}(x) = \sum_x x \left( \frac{1}{n} \sum_{i=1}^n I(x_i = x) \right) = \frac{1}{n} \sum_{i=1}^n x_i$$

**Sample Mean Is Unbiased** An estimator  $\hat{\theta}$  is called an *unbiased estimator* for parameter  $\theta$  if  $E[\hat{\theta}] = \theta$  for every possible value of  $\theta$ . The sample mean  $\hat{\mu}$  is an unbiased estimator for the population mean  $\mu$ , as

$$E[\hat{\mu}] = E \left[ \frac{1}{n} \sum_{i=1}^n x_i \right] = \frac{1}{n} \sum_{i=1}^n E[x_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu \quad (2.7)$$

where we use the fact that the random variables  $x_i$  are IID according to  $X$ , which implies that they have the same mean  $\mu$  as  $X$ , that is,  $E[x_i] = \mu$  for all  $x_i$ . We also used the fact that the expectation function  $E$  is a *linear operator*, that is, for any two random variables  $X$  and  $Y$ , and real numbers  $a$  and  $b$ , we have  $E[aX + bY] = aE[X] + bE[Y]$ .

**Robustness** We say that a statistic is *robust* if it is not affected by extreme values (such as outliers) in the data. The sample mean is unfortunately not robust because a single large value (an outlier) can skew the average. A more robust measure is the *trimmed mean* obtained after discarding a small fraction of extreme values on one or both ends. Furthermore, the mean can be somewhat misleading in that it is typically not a value that occurs in the sample, and it may not even be a value that the random variable can actually assume (for a discrete random variable). For example, the number of cars per capita is an integer-valued random variable, but according to the US Bureau of Transportation Studies, the average number of passenger cars in the United States was 0.45 in 2008 (137.1 million cars, with a population size of 304.4 million). Obviously, one cannot own 0.45 cars; it can be interpreted as saying that on average there are 45 cars per 100 people.



**Figure 2.1.** Geometric interpretation of mean and variance. The vectors are shown in the (conceptual)  $n$ -dimensional space  $\mathbb{R}^n$  spanned by the  $n$  points.

**Geometric Interpretation of Sample Mean** Treating the data sample for attribute  $X$  as a vector in  $n$ -dimensional space, where  $n$  is the sample size, we write  $X = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ . Further, let  $\mathbf{1} = (1, 1, \dots, 1)^T \in \mathbb{R}^n$  be the vector all of whose elements are 1, which is also called the *ones vector*. Consider the projection of  $X$  onto the vector  $\mathbf{1}$ , we have

$$\mathbf{p} = \left( \frac{X^T \mathbf{1}}{\mathbf{1}^T \mathbf{1}} \right) \cdot \mathbf{1} = \left( \frac{\sum_{i=1}^n x_i}{n} \right) \cdot \mathbf{1} = \hat{\mu} \cdot \mathbf{1}$$

Thus, the sample mean is simply the offset or the scalar projection of  $X$  on the vector  $\mathbf{1}$  (see Eq. (1.12)), as shown in Figure 2.1:

$$\hat{\mu} = \text{proj}_{\mathbf{1}}(X) = \left( \frac{X^T \mathbf{1}}{\mathbf{1}^T \mathbf{1}} \right) \quad (2.8)$$

The sample mean can be used to center the attribute  $X$ . Define the *centered attribute vector*,  $\bar{X}$ , as follows:

$$\bar{X} = X - \hat{\mu} \cdot \mathbf{1} = \begin{pmatrix} x_1 - \hat{\mu} \\ x_2 - \hat{\mu} \\ \vdots \\ x_n - \hat{\mu} \end{pmatrix} \quad (2.9)$$

From Figure 2.1 we can observe that  $\hat{\mu} \cdot \mathbf{1}$  is the parallel component and  $\bar{X}$  is the perpendicular component of the projection of  $X$  onto  $\mathbf{1}$ . We can also see that  $\mathbf{1}$  and  $\bar{X}$  are orthogonal to each other, since

$$\mathbf{1}^T \bar{X} = \mathbf{1}^T (X - \hat{\mu} \cdot \mathbf{1}) = \mathbf{1}^T X - \left( \frac{X^T \mathbf{1}}{\mathbf{1}^T \mathbf{1}} \right) \cdot \mathbf{1}^T \mathbf{1} = 0$$

In fact, the subspace containing  $\bar{X}$  is an *orthogonal complement* of the space spanned by  $\mathbf{1}$ .

## Median

The *median* of a random variable is defined as the value  $m$  such that

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

In other words, the median  $m$  is the “middle-most” value; half of the values of  $X$  are less and half of the values of  $X$  are more than  $m$ . In terms of the (inverse) cumulative distribution function, the median is therefore the value  $m$  for which

$$F(m) = 0.5 \text{ or } m = F^{-1}(0.5)$$

The *sample median* can be obtained from the empirical CDF [Eq.(2.1)] or the empirical inverse CDF [Eq.(2.2)] by computing

$$\hat{F}(m) = 0.5 \text{ or } m = \hat{F}^{-1}(0.5)$$

A simpler approach to compute the sample median is to first sort all the values  $x_i$  ( $i \in [1, n]$ ) in increasing order. If  $n$  is odd, the median is the value at position  $\frac{n+1}{2}$ . If  $n$  is even, the values at positions  $\frac{n}{2}$  and  $\frac{n}{2} + 1$  are both medians.

Unlike the mean, median is robust, as it is not affected very much by extreme values. Also, it is a value that occurs in the sample and a value the random variable can actually assume.

## Mode

The *mode* of a random variable  $X$  is the value at which the probability mass function or the probability density function attains its maximum value, depending on whether  $X$  is discrete or continuous, respectively.

The *sample mode* is a value for which the empirical probability mass function [Eq.(2.3)] attains its maximum, given as

$$\text{mode}(X) = \arg \max_x \hat{f}(x)$$

The mode may not be a very useful measure of central tendency for a sample because by chance an unrepresentative element may be the most frequent element. Furthermore, if all values in the sample are distinct, each of them will be the mode.

**Example 2.1 (Sample Mean, Median, and Mode).** Consider the attribute `sepal length` ( $X_1$ ) in the Iris dataset, whose values are shown in Table 1.2. The sample mean is given as follows:

$$\hat{\mu} = \frac{1}{150}(5.9 + 6.9 + \dots + 7.7 + 5.1) = \frac{876.5}{150} = 5.843$$

Figure 2.2 shows all 150 values of `sepal length`, and the sample mean. Figure 2.3(a) shows the empirical CDF and Figure 2.3(b) shows the empirical inverse CDF for `sepal length`.

Because  $n = 150$  is even, the sample median is the value at positions  $\frac{n}{2} = 75$  and  $\frac{n}{2} + 1 = 76$  in sorted order. For `sepal length` both these values are 5.8; thus the sample median is 5.8. From the inverse CDF in Figure 2.3(b), we can see that

$$\hat{F}(5.8) = 0.5 \text{ or } 5.8 = \hat{F}^{-1}(0.5)$$

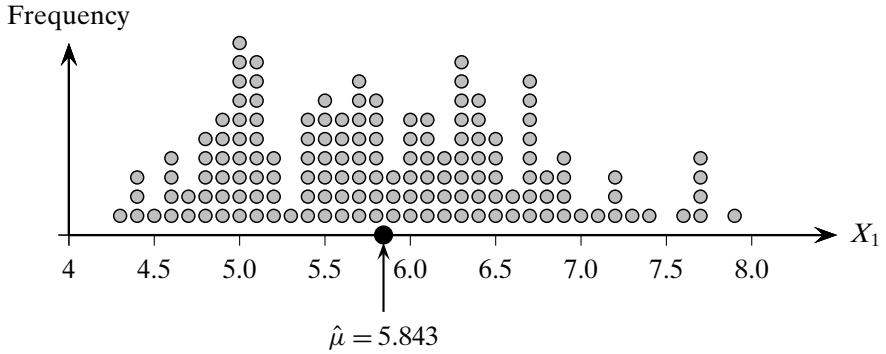


Figure 2.2. Sample mean for sepal length. Multiple occurrences of the same value are shown stacked.

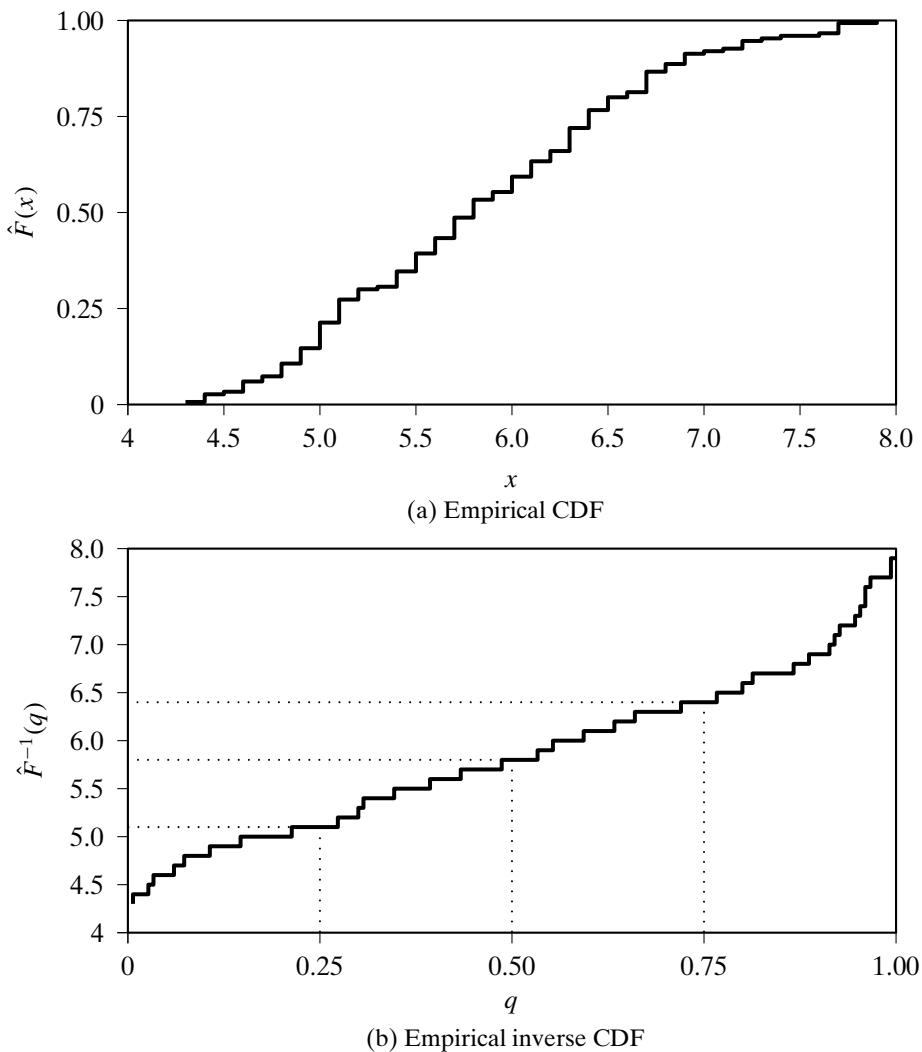


Figure 2.3. Empirical CDF and inverse CDF: sepal length.

The sample mode for sepal length is 5, which can be observed from the frequency of 5 in [Figure 2.2](#). The empirical probability mass at  $x = 5$  is

$$\hat{f}(5) = \frac{10}{150} = 0.067$$

### 2.1.2 Measures of Dispersion

The measures of dispersion give an indication about the spread or variation in the values of a random variable.

#### Range

The *value range* or simply *range* of a random variable  $X$  is the difference between the maximum and minimum values of  $X$ , given as

$$r = \max\{X\} - \min\{X\}$$

The (value) range of  $X$  is a population parameter, not to be confused with the range of the function  $X$ , which is the set of all the values  $X$  can assume. Which range is being used should be clear from the context.

The *sample range* is a statistic, given as

$$\hat{r} = \max_{i=1}^n \{x_i\} - \min_{i=1}^n \{x_i\}$$

By definition, range is sensitive to extreme values, and thus is not robust.

#### Interquartile Range

*Quartiles* are special values of the quantile function [[Eq. \(2.2\)](#)] that divide the data into four equal parts. That is, quartiles correspond to the quantile values of 0.25, 0.5, 0.75, and 1.0. The *first quartile* is the value  $q_1 = F^{-1}(0.25)$ , to the left of which 25% of the points lie; the *second quartile* is the same as the median value  $q_2 = F^{-1}(0.5)$ , to the left of which 50% of the points lie; the third quartile  $q_3 = F^{-1}(0.75)$  is the value to the left of which 75% of the points lie; and the fourth quartile is the maximum value of  $X$ , to the left of which 100% of the points lie.

A more robust measure of the dispersion of  $X$  is the *interquartile range (IQR)*, defined as

$$IQR = q_3 - q_1 = F^{-1}(0.75) - F^{-1}(0.25) \quad (2.10)$$

IQR can also be thought of as a *trimmed range*, where we discard 25% of the low and high values of  $X$ . Or put differently, it is the range for the middle 50% of the values of  $X$ . IQR is robust by definition.

The *sample IQR* can be obtained by plugging in the empirical inverse CDF in [Eq. \(2.10\)](#):

$$\widehat{IQR} = \hat{q}_3 - \hat{q}_1 = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$$

### Variance and Standard Deviation

The *variance* of a random variable  $X$  provides a measure of how much the values of  $X$  deviate from the mean or expected value of  $X$ . More formally, variance is the expected value of the squared deviation from the mean, defined as

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{if } X \text{ is continuous} \end{cases} \quad (2.11)$$

The *standard deviation*,  $\sigma$ , is defined as the positive square root of the variance,  $\sigma^2$ .

We can also write the variance as the difference between the expectation of  $X^2$  and the square of the expectation of  $X$ :

$$\begin{aligned} \sigma^2 = \text{var}(X) &= E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned} \quad (2.12)$$

It is worth noting that variance is in fact the *second moment about the mean*, corresponding to  $r = 2$ , which is a special case of the  *$r$ th moment about the mean* for a random variable  $X$ , defined as  $E[(X - \mu)^r]$ .

**Sample Variance** The *sample variance* is defined as

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad (2.13)$$

It is the average squared deviation of the data values  $x_i$  from the sample mean  $\hat{\mu}$ , and can be derived by plugging in the empirical probability function  $\hat{f}$  from Eq. (2.3) into Eq. (2.11), as

$$\hat{\sigma}^2 = \sum_x (x - \hat{\mu})^2 \hat{f}(x) = \sum_x (x - \hat{\mu})^2 \left( \frac{1}{n} \sum_{i=1}^n I(x_i = x) \right) = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

The *sample standard deviation* is given as the positive square root of the sample variance:

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2}$$

The *standard score*, also called the  *$z$ -score*, of a sample value  $x_i$  is the number of standard deviations the value is away from the mean:

$$z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}} \quad (2.14)$$

Put differently, the  $z$ -score of  $x_i$  measures the deviation of  $x_i$  from the mean value  $\hat{\mu}$ , in units of  $\hat{\sigma}$ .

**Example 2.2.** Consider the data sample for `sepal_length` shown in Figure 2.2. We can see that the sample range is given as

$$\max_i \{x_i\} - \min_i \{x_i\} = 7.9 - 4.3 = 3.6$$

From the inverse CDF for `sepal_length` in Figure 2.3(b), we can find the sample IQR as follows:

$$\hat{q}_1 = \hat{F}^{-1}(0.25) = 5.1$$

$$\hat{q}_3 = \hat{F}^{-1}(0.75) = 6.4$$

$$\widehat{IQR} = \hat{q}_3 - \hat{q}_1 = 6.4 - 5.1 = 1.3$$

The sample variance is given as

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 = 102.168/150 = 0.681$$

The sample standard deviation is then

$$\hat{\sigma} = \sqrt{0.681} = 0.825$$

**Variance of the Sample Mean** Because the sample mean  $\hat{\mu}$  is itself a statistic, we can compute its mean value and variance. The expected value of the sample mean is simply  $\mu$ , as we saw in Eq. (2.7). To derive an expression for the variance of the sample mean, we utilize the fact that the random variables  $x_i$  are all independent, and thus

$$\text{var}\left(\sum_{i=1}^n x_i\right) = \sum_{i=1}^n \text{var}(x_i)$$

Further, because all the  $x_i$ 's are identically distributed as  $X$ , they have the same variance as  $X$ , that is,

$$\text{var}(x_i) = \sigma^2 \text{ for all } i$$

Combining the above two facts, we get

$$\text{var}\left(\sum_{i=1}^n x_i\right) = \sum_{i=1}^n \text{var}(x_i) = \sum_{i=1}^n \sigma^2 = n\sigma^2 \quad (2.15)$$

Further, note that

$$E\left[\sum_{i=1}^n x_i\right] = n\mu \quad (2.16)$$

Using Eqs. (2.12), (2.15), and (2.16), the variance of the sample mean  $\hat{\mu}$  can be computed as

$$\begin{aligned}\text{var}(\hat{\mu}) &= E[(\hat{\mu} - \mu)^2] = E[\hat{\mu}^2] - \mu^2 = E\left[\left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2\right] - \frac{1}{n^2} E\left[\sum_{i=1}^n x_i\right]^2 \\ &= \frac{1}{n^2} \left( E\left[\left(\sum_{i=1}^n x_i\right)^2\right] - E\left[\sum_{i=1}^n x_i\right]^2 \right) = \frac{1}{n^2} \text{var}\left(\sum_{i=1}^n x_i\right) \\ &= \frac{\sigma^2}{n}\end{aligned}\tag{2.17}$$

In other words, the sample mean  $\hat{\mu}$  varies or deviates from the mean  $\mu$  in proportion to the population variance  $\sigma^2$ . However, the deviation can be made smaller by considering larger sample size  $n$ .

**Bias of Sample Variance** The sample variance in Eq. (2.13) is a *biased estimator* for the true population variance,  $\sigma^2$ , that is,  $E[\hat{\sigma}^2] \neq \sigma^2$ . To show this we make use of the identity

$$\sum_{i=1}^n (x_i - \mu)^2 = n(\hat{\mu} - \mu)^2 + \sum_{i=1}^n (x_i - \hat{\mu})^2\tag{2.18}$$

Computing the expectation of  $\hat{\sigma}^2$  by using Eq. (2.18) in the first step, we get

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2\right] = E\left[\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2\right] - E[(\hat{\mu} - \mu)^2]\tag{2.19}$$

Recall that the random variables  $x_i$  are IID according to  $X$ , which means that they have the same mean  $\mu$  and variance  $\sigma^2$  as  $X$ . This means that

$$E[(x_i - \mu)^2] = \sigma^2$$

Further, from Eq. (2.17) the sample mean  $\hat{\mu}$  has variance  $E[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}$ . Plugging these into the Eq. (2.19) we get

$$\begin{aligned}E[\hat{\sigma}^2] &= \frac{1}{n} n\sigma^2 - \frac{\sigma^2}{n} \\ &= \left(\frac{n-1}{n}\right)\sigma^2\end{aligned}$$

The sample variance  $\hat{\sigma}^2$  is a biased estimator of  $\sigma^2$ , as its expected value differs from the population variance by a factor of  $\frac{n-1}{n}$ . However, it is *asymptotically unbiased*, that is, the bias vanishes as  $n \rightarrow \infty$  because

$$\lim_{n \rightarrow \infty} \frac{n-1}{n} = \lim_{n \rightarrow \infty} 1 - \frac{1}{n} = 1$$

Put differently, as the sample size increases, we have

$$E[\hat{\sigma}^2] \rightarrow \sigma^2 \quad \text{as } n \rightarrow \infty$$

The discussion above makes it clear that if we want an unbiased estimate of the sample variance, denoted  $\hat{\sigma}_u^2$ , we must divide by  $n - 1$  instead of  $n$ :

$$\hat{\sigma}_u^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

We can verify that the expected value of  $\hat{\sigma}_u^2$  is given as

$$\begin{aligned} E[\hat{\sigma}_u^2] &= E\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2\right] = \frac{1}{n-1} \cdot E\left[\sum_{i=1}^n (x_i - \mu)^2\right] - \frac{n}{n-1} \cdot E[(\hat{\mu} - \mu)^2] \\ &= \frac{n}{n-1} \sigma^2 - \frac{n}{n-1} \cdot \frac{\sigma^2}{n} \\ &= \frac{n}{n-1} \sigma^2 - \frac{1}{n-1} \sigma^2 = \sigma^2 \end{aligned}$$

**Geometric Interpretation of Sample Variance** Let the data sample for attribute  $X$  denote a vector in  $n$ -dimensional space, where  $n$  is the sample size, and let  $\bar{X}$  denote the centered attribute vector (see Eq. (2.9))

$$\bar{X} = X - \hat{\mu} \cdot \mathbf{1} = \begin{pmatrix} x_1 - \hat{\mu} \\ x_2 - \hat{\mu} \\ \vdots \\ x_n - \hat{\mu} \end{pmatrix}$$

We can then rewrite Eq. (2.13) in terms of the magnitude of  $\bar{X}$ , that is, the dot product of  $\bar{X}$  with itself:

$$\hat{\sigma}^2 = \frac{1}{n} \|\bar{X}\|^2 = \frac{1}{n} \bar{X}^T \bar{X} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad (2.20)$$

The sample variance can thus be interpreted as the squared magnitude of the centered attribute vector, or the dot product of the centered attribute vector with itself, normalized by the sample size.

The geometric interpretation make it clear why dividing by  $n - 1$  make sense to obtain an unbiased estimate for the sample variance. Define the *degrees of freedom* (dof) of a statistical vector as the dimensionality of the subspace that contains the vector. In Figure 2.1 notice that the centered attribute vector  $\bar{X} = X - \hat{\mu} \cdot \mathbf{1}$  lies in a  $n - 1$  dimensional subspace that is an orthogonal complement of the 1 dimensional subspace spanned by the ones vector  $\mathbf{1}$ . Thus, the vector  $\bar{X}$  has only  $n - 1$  degrees of freedom, and the unbiased sample variance is simply the mean or expected squared length of  $\bar{X}$  per dimension

$$\sigma_u^2 = \frac{\|\bar{X}\|^2}{n-1} = \frac{\bar{X}^T \bar{X}}{n-1} = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \hat{\mu})^2$$

## 2.2 BIVARIATE ANALYSIS

---

In bivariate analysis, we consider two attributes at the same time. We are specifically interested in understanding the association or dependence between them, if any. We thus restrict our attention to the two numeric attributes of interest, say  $X_1$  and  $X_2$ , with the data  $\mathbf{D}$  represented as an  $n \times 2$  matrix:

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

Geometrically, we can think of  $\mathbf{D}$  in two ways. It can be viewed as  $n$  points or vectors in 2-dimensional space over the attributes  $X_1$  and  $X_2$ , that is,  $\mathbf{x}_i = (x_{i1}, x_{i2})^T \in \mathbb{R}^2$ . Alternatively, it can be viewed as two points or vectors in an  $n$ -dimensional space comprising the points, that is, each column is a vector in  $\mathbb{R}^n$ , as follows:

$$\begin{aligned} X_1 &= (x_{11}, x_{21}, \dots, x_{n1})^T \\ X_2 &= (x_{12}, x_{22}, \dots, x_{n2})^T \end{aligned}$$

In the probabilistic view, the column vector  $\mathbf{X} = (X_1, X_2)^T$  is considered a bivariate vector random variable, and the points  $\mathbf{x}_i$  ( $1 \leq i \leq n$ ) are treated as a random sample drawn from  $\mathbf{X}$ , that is,  $\mathbf{x}_i$ 's are considered independent and identically distributed as  $\mathbf{X}$ .

### Empirical Joint Probability Mass Function

The *empirical joint probability mass function* for  $\mathbf{X}$  is given as

$$\hat{f}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n I(\mathbf{x}_i = \mathbf{x}) \quad (2.21)$$

$$\hat{f}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = \frac{1}{n} \sum_{i=1}^n I(x_{i1} = x_1, x_{i2} = x_2)$$

where  $\mathbf{x} = (x_1, x_2)^T$  and  $I$  is a indicator variable that takes on the value 1 only when its argument is true:

$$I(\mathbf{x}_i = \mathbf{x}) = \begin{cases} 1 & \text{if } x_{i1} = x_1 \text{ and } x_{i2} = x_2 \\ 0 & \text{otherwise} \end{cases}$$

As in the univariate case, the probability function puts a probability mass of  $\frac{1}{n}$  at each point in the data sample.

### 2.2.1 Measures of Location and Dispersion

#### Mean

The bivariate mean is defined as the expected value of the vector random variable  $\mathbf{X}$ , defined as follows:

$$\boldsymbol{\mu} = E[\mathbf{X}] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad (2.22)$$

In other words, the bivariate mean vector is simply the vector of expected values along each attribute.

The sample mean vector can be obtained from  $\hat{f}_{X_1}$  and  $\hat{f}_{X_2}$ , the empirical probability mass functions of  $X_1$  and  $X_2$ , respectively, using Eq.(2.6). It can also be computed from the joint empirical PMF in Eq.(2.21)

$$\hat{\boldsymbol{\mu}} = \sum_{\mathbf{x}} \mathbf{x} \hat{f}(\mathbf{x}) = \sum_{\mathbf{x}} \mathbf{x} \left( \frac{1}{n} \sum_{i=1}^n I(\mathbf{x}_i = \mathbf{x}) \right) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad (2.23)$$

#### Variance

We can compute the variance along each attribute, namely  $\sigma_1^2$  for  $X_1$  and  $\sigma_2^2$  for  $X_2$  using Eq.(2.11). The *total variance* is given as

$$\sigma_1^2 + \sigma_2^2$$

The sample variances  $\hat{\sigma}_1^2$  and  $\hat{\sigma}_2^2$  can be estimated using Eq.(2.13), and the *sample total variance* [Eq.(1.8)] is simply

$$\text{var}(\mathbf{D}) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2$$

### 2.2.2 Measures of Association

#### Covariance

The *covariance* between two attributes  $X_1$  and  $X_2$  provides a measure of the association or linear dependence between them, and is defined as

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] \quad (2.24)$$

By linearity of expectation, we have

$$\begin{aligned} \sigma_{12} &= E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2 - X_1 \mu_2 - X_2 \mu_1 + \mu_1 \mu_2] \\ &= E[X_1 X_2] - \mu_2 E[X_1] - \mu_1 E[X_2] + \mu_1 \mu_2 = E[X_1 X_2] - \mu_1 \mu_2 \end{aligned}$$

which implies

$$\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2] \quad (2.25)$$

[Eq. \(2.25\)](#) can be seen as a generalization of the univariate variance [[Eq. \(2.12\)](#)] to the bivariate case.

If  $X_1$  and  $X_2$  are independent random variables, then we conclude that their covariance is zero. This is because if  $X_1$  and  $X_2$  are independent, then we have

$$E[X_1 X_2] = E[X_1] \cdot E[X_2]$$

which in turn implies that

$$\sigma_{12} = 0$$

However, the converse is not true. That is, if  $\sigma_{12} = 0$ , one cannot claim that  $X_1$  and  $X_2$  are independent. All we can say is that there is no linear dependence between them, but we cannot rule out that there might be a higher order relationship or dependence between the two attributes.

The *sample covariance* between  $X_1$  and  $X_2$  is given as

$$\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2) \quad (2.26)$$

It can be derived by substituting the empirical joint probability mass function  $\hat{f}(x_1, x_2)$  from [Eq. \(2.21\)](#) into [Eq. \(2.24\)](#), as follows:

$$\begin{aligned} \hat{\sigma}_{12} &= E[(X_1 - \hat{\mu}_1)(X_2 - \hat{\mu}_2)] \\ &= \sum_{\mathbf{x}=(x_1, x_2)^T} (x_1 - \hat{\mu}_1)(x_2 - \hat{\mu}_2) \hat{f}(x_1, x_2) \\ &= \frac{1}{n} \sum_{\mathbf{x}=(x_1, x_2)^T} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1) \cdot (x_{i2} - \hat{\mu}_2) \cdot I(x_{i1} = x_1, x_{i2} = x_2) \\ &= \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2) \end{aligned}$$

Notice that sample covariance is a generalization of the sample variance [[Eq. \(2.13\)](#)] because

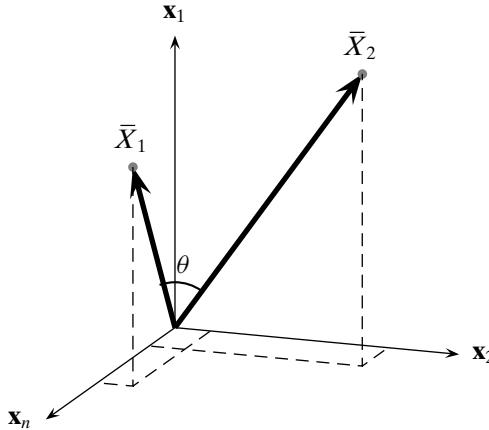
$$\hat{\sigma}_{11} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i1} - \hat{\mu}_1) = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2 = \hat{\sigma}_1^2$$

and similarly,  $\hat{\sigma}_{22} = \hat{\sigma}_2^2$ .

## Correlation

The *correlation* between variables  $X_1$  and  $X_2$  is the *standardized covariance*, obtained by normalizing the covariance with the standard deviation of each variable, given as

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}} \quad (2.27)$$



**Figure 2.4.** Geometric interpretation of covariance and correlation. The two centered attribute vectors are shown in the (conceptual)  $n$ -dimensional space  $\mathbb{R}^n$  spanned by the  $n$  points.

The *sample correlation* for attributes  $X_1$  and  $X_2$  is given as

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2} \sqrt{\sum_{i=1}^n (x_{i2} - \hat{\mu}_2)^2}} \quad (2.28)$$

### Geometric Interpretation of Sample Covariance and Correlation

Let  $\bar{X}_1$  and  $\bar{X}_2$  denote the centered attribute vectors in  $\mathbb{R}^n$ , given as follows:

$$\bar{X}_1 = X_1 - \hat{\mu}_1 \cdot \mathbf{1} = \begin{pmatrix} x_{11} - \hat{\mu}_1 \\ x_{21} - \hat{\mu}_1 \\ \vdots \\ x_{n1} - \hat{\mu}_1 \end{pmatrix} \quad \bar{X}_2 = X_2 - \hat{\mu}_2 \cdot \mathbf{1} = \begin{pmatrix} x_{12} - \hat{\mu}_2 \\ x_{22} - \hat{\mu}_2 \\ \vdots \\ x_{n2} - \hat{\mu}_2 \end{pmatrix}$$

The sample covariance [Eq. (2.26)] can then be written as

$$\hat{\sigma}_{12} = \frac{\bar{X}_1^T \bar{X}_2}{n} \quad (2.29)$$

In other words, the covariance between the two attributes is simply the dot product between the two centered attribute vectors, normalized by the sample size. The above can be seen as a generalization of the univariate sample variance given in Eq. (2.20).

The sample correlation [Eq. (2.28)] can be written as

$$\hat{\rho}_{12} = \frac{\bar{X}_1^T \bar{X}_2}{\sqrt{\bar{X}_1^T \bar{X}_1} \sqrt{\bar{X}_2^T \bar{X}_2}} = \frac{\bar{X}_1^T \bar{X}_2}{\|\bar{X}_1\| \|\bar{X}_2\|} = \left( \frac{\bar{X}_1}{\|\bar{X}_1\|} \right)^T \left( \frac{\bar{X}_2}{\|\bar{X}_2\|} \right) = \cos \theta \quad (2.30)$$

Thus, the correlation coefficient is simply the cosine of the angle [Eq. (1.6)] between the two centered attribute vectors, as illustrated in Figure 2.4.

### Covariance Matrix

The variance–covariance information for the two attributes  $X_1$  and  $X_2$  can be summarized in the square  $2 \times 2$  covariance matrix, given as

$$\begin{aligned}\Sigma &= E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \\ &= E\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} (X_1 - \mu_1 \quad X_2 - \mu_2) \right] \\ &= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\end{aligned}\tag{2.31}$$

Because  $\sigma_{12} = \sigma_{21}$ ,  $\Sigma$  is a *symmetric* matrix. The covariance matrix records the attribute specific variances on the main diagonal, and the covariance information on the off-diagonal elements.

The *total variance* of the two attributes is given as the sum of the diagonal elements of  $\Sigma$ , which is also called the *trace* of  $\Sigma$ , given as

$$tr(\Sigma) = \sigma_1^2 + \sigma_2^2$$

We immediately have  $tr(\Sigma) \geq 0$ .

The *generalized variance* of the two attributes also considers the covariance, in addition to the attribute variances, and is given as the *determinant* of the covariance matrix  $\Sigma$ , denoted as  $|\Sigma|$  or  $\det(\Sigma)$ . The generalized covariance is non-negative, because

$$|\Sigma| = \det(\Sigma) = \sigma_1^2 \sigma_2^2 - \sigma_{12}^2 = \sigma_1^2 \sigma_2^2 - \rho_{12}^2 \sigma_1^2 \sigma_2^2 = (1 - \rho_{12}^2) \sigma_1^2 \sigma_2^2$$

where we used Eq. (2.27), that is,  $\sigma_{12} = \rho_{12} \sigma_1 \sigma_2$ . Note that  $|\rho_{12}| \leq 1$  implies that  $\rho_{12}^2 \leq 1$ , which in turn implies that  $\det(\Sigma) \geq 0$ , that is, the determinant is non-negative.

The *sample covariance matrix* is given as

$$\widehat{\Sigma} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix}\tag{2.32}$$

The sample covariance matrix  $\widehat{\Sigma}$  shares the same properties as  $\Sigma$ , that is, it is symmetric and  $|\widehat{\Sigma}| \geq 0$ , and it can be used to easily obtain the sample generalized variance, given as  $|\widehat{\Sigma}|$ , and the sample total variance, given as

$$\text{var}(\mathbf{D}) = tr(\widehat{\Sigma}) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2\tag{2.33}$$

**Example 2.3 (Sample Mean and Covariance).** Consider the sepal length and sepal width attributes for the Iris dataset, plotted in Figure 2.5. There are  $n = 150$  points in the  $d = 2$  dimensional attribute space. The sample mean vector is given as

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

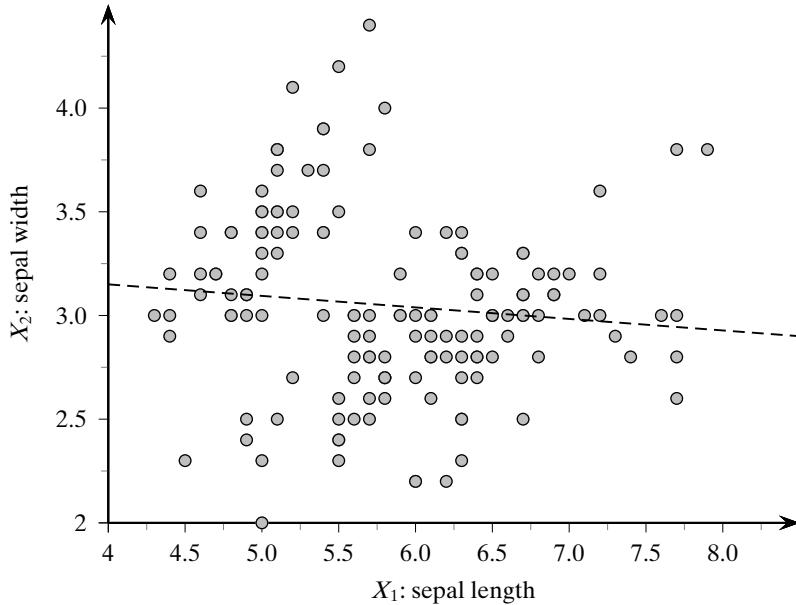


Figure 2.5. Correlation between sepal length and sepal width.

The sample covariance matrix is given as

$$\widehat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

The variance for sepal length is  $\hat{\sigma}_1^2 = 0.681$ , and that for sepal width is  $\hat{\sigma}_2^2 = 0.187$ . The covariance between the two attributes is  $\hat{\sigma}_{12} = -0.039$ , and the correlation between them is

$$\hat{\rho}_{12} = \frac{-0.039}{\sqrt{0.681 \cdot 0.187}} = -0.109$$

Thus, there is a very weak negative correlation between these two attributes, as evidenced by the best linear fit line in Figure 2.5. Alternatively, we can consider the attributes sepal length and sepal width as two points in  $\mathbb{R}^n$ . The correlation is then the cosine of the angle between them; we have

$$\hat{\rho}_{12} = \cos \theta = -0.109, \text{ which implies that } \theta = \cos^{-1}(-0.109) = 96.26^\circ$$

The angle is close to  $90^\circ$ , that is, the two attribute vectors are almost orthogonal, indicating weak correlation. Further, the angle being greater than  $90^\circ$  indicates negative correlation.

The sample total variance is given as

$$tr(\widehat{\Sigma}) = 0.681 + 0.187 = 0.868$$

and the sample generalized variance is given as

$$|\widehat{\Sigma}| = \det(\widehat{\Sigma}) = 0.681 \cdot 0.187 - (-0.039)^2 = 0.126$$

## 2.3 MULTIVARIATE ANALYSIS

---

In multivariate analysis, we consider all the  $d$  numeric attributes  $X_1, X_2, \dots, X_d$ . The full data is an  $n \times d$  matrix, given as

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} = \begin{pmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & & | \\ & & \vdots & \\ & & \mathbf{x}_n^T & \end{pmatrix} = \begin{pmatrix} - & \mathbf{x}_1^T & - \\ - & \mathbf{x}_2^T & - \\ & \vdots & \\ - & \mathbf{x}_n^T & - \end{pmatrix}$$

In the row view, the data can be considered as a set of  $n$  points or vectors in the  $d$ -dimensional attribute space

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})^T \in \mathbb{R}^d$$

In the column view, the data can be considered as a set of  $d$  points or vectors in the  $n$ -dimensional space spanned by the data points

$$X_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T \in \mathbb{R}^n$$

In the probabilistic view, the  $d$  attributes are modeled as a vector random variable,  $\mathbf{X} = (X_1, X_2, \dots, X_d)^T$ , and the points  $\mathbf{x}_i$  are considered to be a random sample drawn from  $\mathbf{X}$ , that is, they are independent and identically distributed as  $\mathbf{X}$ .

### Mean

Generalizing Eq. (2.22), the *multivariate mean vector* is obtained by taking the mean of each attribute, given as

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_d] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}$$

Generalizing Eq. (2.23), the *sample mean* is given as

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

(2.34)

We can also obtain the sample mean by projecting each attribute vector  $X_i$  on the ones vector  $\mathbf{1}$ , which can be written compactly as

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \mathbf{D}^T \mathbf{1}$$

### Covariance Matrix

Generalizing Eq. (2.31) to  $d$  dimensions, the multivariate covariance information is captured by the  $d \times d$  (square) symmetric *covariance matrix* that gives the covariance

for each pair of attributes:

$$\Sigma = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

The diagonal element  $\sigma_i^2$  specifies the attribute variance for  $X_i$ , whereas the off-diagonal elements  $\sigma_{ij} = \sigma_{ji}$  represent the covariance between attribute pairs  $X_i$  and  $X_j$ .

### Covariance Matrix Is Positive Semidefinite

It is worth noting that  $\Sigma$  is a *positive semidefinite* matrix, that is,

$$\mathbf{a}^T \Sigma \mathbf{a} \geq 0 \text{ for any } d\text{-dimensional vector } \mathbf{a}$$

To see this, observe that

$$\begin{aligned} \mathbf{a}^T \Sigma \mathbf{a} &= \mathbf{a}^T E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \mathbf{a} \\ &= E[\mathbf{a}^T (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T \mathbf{a}] \\ &= E[Y^2] \\ &\geq 0 \end{aligned}$$

where  $Y$  is the random variable  $Y = \mathbf{a}^T (\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^d a_i (X_i - \mu_i)$ , and we use the fact that the expectation of a squared random variable is non-negative.

Because  $\Sigma$  is also symmetric, this implies that all the eigenvalues of  $\Sigma$  are real and non-negative. In other words the  $d$  eigenvalues of  $\Sigma$  can be arranged from the largest to the smallest as follows:  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ .

### Total and Generalized Variance

The total variance is given as the trace of the covariance matrix:

$$tr(\Sigma) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2 \quad (2.35)$$

Being a sum of squares, the total variance must be non-negative.

The generalized variance is defined as the determinant of the covariance matrix,  $\det(\Sigma)$ , also denoted as  $|\Sigma|$ ; it gives a single value for the overall multivariate scatter:

$$\det(\Sigma) = |\Sigma| = \prod_{i=1}^d \lambda_i \quad (2.36)$$

Since all the eigenvalues of  $\Sigma$  are non-negative ( $\lambda_i \geq 0$ ), it follows that  $\det(\Sigma) \geq 0$ .

### Sample Covariance Matrix

The *sample covariance matrix* is given as

$$\widehat{\Sigma} = E[(\mathbf{X} - \hat{\mu})(\mathbf{X} - \hat{\mu})^T] = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1d} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \cdots & \hat{\sigma}_d^2 \end{pmatrix} \quad (2.37)$$

Instead of computing the sample covariance matrix element-by-element, we can obtain it via matrix operations. Let  $\bar{\mathbf{D}}$  represent the centered data matrix, given as the matrix of centered attribute vectors  $\bar{X}_i = X_i - \hat{\mu}_i \cdot \mathbf{1}$ , where  $\mathbf{1} \in \mathbb{R}^n$ :

$$\bar{\mathbf{D}} = \mathbf{D} - \mathbf{1} \cdot \hat{\mu}^T = \begin{pmatrix} | & | & | \\ \bar{X}_1 & \bar{X}_2 & \cdots & \bar{X}_d \\ | & | & & | \end{pmatrix}$$

Alternatively, the centered data matrix can also be written in terms of the centered points  $\bar{\mathbf{x}}_i = \mathbf{x}_i - \hat{\mu}$ :

$$\bar{\mathbf{D}} = \mathbf{D} - \mathbf{1} \cdot \hat{\mu}^T = \begin{pmatrix} \mathbf{x}_1^T - \hat{\mu}^T \\ \mathbf{x}_2^T - \hat{\mu}^T \\ \vdots \\ \mathbf{x}_n^T - \hat{\mu}^T \end{pmatrix} = \begin{pmatrix} - & \bar{\mathbf{x}}_1^T & - \\ - & \bar{\mathbf{x}}_2^T & - \\ \vdots & \vdots & \ddots & \vdots \\ - & \bar{\mathbf{x}}_n^T & - \end{pmatrix}$$

In matrix notation, the sample covariance matrix can be written as

$$\widehat{\Sigma} = \frac{1}{n} (\bar{\mathbf{D}}^T \bar{\mathbf{D}}) = \frac{1}{n} \begin{pmatrix} \bar{X}_1^T \bar{X}_1 & \bar{X}_1^T \bar{X}_2 & \cdots & \bar{X}_1^T \bar{X}_d \\ \bar{X}_2^T \bar{X}_1 & \bar{X}_2^T \bar{X}_2 & \cdots & \bar{X}_2^T \bar{X}_d \\ \vdots & \vdots & \ddots & \vdots \\ \bar{X}_d^T \bar{X}_1 & \bar{X}_d^T \bar{X}_2 & \cdots & \bar{X}_d^T \bar{X}_d \end{pmatrix} \quad (2.38)$$

The sample covariance matrix is thus given as the pairwise *inner or dot products* of the centered attribute vectors, normalized by the sample size.

In terms of the centered points  $\bar{\mathbf{x}}_i$ , the sample covariance matrix can also be written as a sum of rank-one matrices obtained as the *outer product* of each centered point:

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \bar{\mathbf{x}}_i \cdot \bar{\mathbf{x}}_i^T \quad (2.39)$$

Also, the sample total variance is given as

$$\text{var}(\mathbf{D}) = \text{tr}(\widehat{\Sigma}) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \cdots + \hat{\sigma}_d^2$$

and the sample generalized variance as  $|\widehat{\Sigma}| = \det(\widehat{\Sigma})$ .

### Sample Scatter Matrix

The *sample scatter matrix* is the  $d \times d$  positive semi-definite matrix defined as

$$\mathbf{S} = \bar{\mathbf{D}}^T \bar{\mathbf{D}} = \sum_{i=1}^n \bar{\mathbf{x}}_i \cdot \bar{\mathbf{x}}_i^T$$

It is simply the un-normalized sample covariance matrix, since  $\mathbf{S} = n \cdot \hat{\Sigma}$ .

**Example 2.4 (Sample Mean and Covariance Matrix).** Let us consider all four numeric attributes for the Iris dataset, namely sepal length, sepal width, petal length, and petal width. The multivariate sample mean vector is given as

$$\hat{\boldsymbol{\mu}} = (5.843 \quad 3.054 \quad 3.759 \quad 1.199)^T$$

and the sample covariance matrix is given as

$$\hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 & 1.265 & 0.513 \\ -0.039 & 0.187 & -0.320 & -0.117 \\ 1.265 & -0.320 & 3.092 & 1.288 \\ 0.513 & -0.117 & 1.288 & 0.579 \end{pmatrix}$$

The sample total variance is

$$\text{var}(\mathbf{D}) = \text{tr}(\hat{\Sigma}) = 0.681 + 0.187 + 3.092 + 0.579 = 4.539$$

and the generalized variance is

$$\det(\hat{\Sigma}) = 1.853 \times 10^{-3}$$

**Example 2.5 (Inner and Outer Product).** To illustrate the inner and outer product-based computation of the sample covariance matrix, consider the 2-dimensional dataset

$$\mathbf{D} = \begin{pmatrix} A_1 & A_2 \\ 1 & 0.8 \\ 5 & 2.4 \\ 9 & 5.5 \end{pmatrix}$$

The mean vector is as follows:

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = \begin{pmatrix} 15/3 \\ 8.7/3 \end{pmatrix} = \begin{pmatrix} 5 \\ 2.9 \end{pmatrix}$$

and the centered data matrix is then given as

$$\bar{\mathbf{D}} = \mathbf{D} - \mathbf{1} \cdot \boldsymbol{\mu}^T = \begin{pmatrix} 1 & 0.8 \\ 5 & 2.4 \\ 9 & 5.5 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} (5 \quad 2.9) = \begin{pmatrix} -4 & -2.1 \\ 0 & -0.5 \\ 4 & 2.6 \end{pmatrix}$$

The inner-product approach [Eq. (2.38)] to compute the sample covariance matrix gives

$$\begin{aligned}\widehat{\Sigma} &= \frac{1}{n} \bar{\mathbf{D}}^T \bar{\mathbf{D}} = \frac{1}{3} \begin{pmatrix} -4 & 0 & 4 \\ -2.1 & -0.5 & 2.6 \end{pmatrix} \cdot \begin{pmatrix} -4 & -2.1 \\ 0 & -0.5 \\ 4 & 2.6 \end{pmatrix} \\ &= \frac{1}{3} \begin{pmatrix} 32 & 18.8 \\ 18.8 & 11.42 \end{pmatrix} = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix}\end{aligned}$$

Alternatively, the outer-product approach [Eq. (2.39)] gives

$$\begin{aligned}\widehat{\Sigma} &= \frac{1}{n} \sum_{i=1}^n \bar{\mathbf{x}}_i \cdot \bar{\mathbf{x}}_i^T \\ &= \frac{1}{3} \left[ \begin{pmatrix} -4 \\ -2.1 \end{pmatrix} \cdot (-4 \quad -2.1) + \begin{pmatrix} 0 \\ -0.5 \end{pmatrix} \cdot (0 \quad -0.5) + \begin{pmatrix} 4 \\ 2.6 \end{pmatrix} \cdot (4 \quad 2.6) \right] \\ &= \frac{1}{3} \left[ \begin{pmatrix} 16.0 & 8.4 \\ 8.4 & 4.41 \end{pmatrix} + \begin{pmatrix} 0.0 & 0.0 \\ 0.0 & 0.25 \end{pmatrix} + \begin{pmatrix} 16.0 & 10.4 \\ 10.4 & 6.76 \end{pmatrix} \right] \\ &= \frac{1}{3} \begin{pmatrix} 32.0 & 18.8 \\ 18.8 & 11.42 \end{pmatrix} = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix}\end{aligned}$$

where the centered points  $\bar{\mathbf{x}}_i$  are the rows of  $\bar{\mathbf{D}}$ . We can see that both the inner and outer product approaches yield the same sample covariance matrix.

## 2.4 DATA NORMALIZATION

---

When analyzing two or more attributes it is often necessary to normalize the values of the attributes, especially in those cases where the values are vastly different in scale.

### Range Normalization

Let  $X$  be an attribute and let  $x_1, x_2, \dots, x_n$  be a random sample drawn from  $X$ . In *range normalization* each value is scaled by the sample range  $\hat{r}$  of  $X$ :

$$x'_i = \frac{x_i - \min_i\{x_i\}}{\hat{r}} = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

After transformation the new attribute takes on values in the range  $[0, 1]$ .

### Standard Score Normalization

In *standard score normalization*, also called  $z$ -normalization, each value is replaced by its  $z$ -score:

$$x'_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

where  $\hat{\mu}$  is the sample mean and  $\hat{\sigma}^2$  is the sample variance of  $X$ . After transformation, the new attribute has mean  $\hat{\mu}' = 0$ , and standard deviation  $\hat{\sigma}' = 1$ .

**Example 2.6.** Consider the example dataset shown in [Table 2.1](#). The attributes Age and Income have very different scales, with the latter having much larger values. Consider the distance between  $\mathbf{x}_1$  and  $\mathbf{x}_2$ :

$$\|\mathbf{x}_1 - \mathbf{x}_2\| = \|(2, 200)^T\| = \sqrt{2^2 + 200^2} = \sqrt{40004} = 200.01$$

As we can observe, the contribution of Age is overshadowed by the value of Income.

The sample range for Age is  $\hat{r} = 40 - 12 = 28$ , with the minimum value 12. After range normalization, the new attribute is given as

$$\text{Age}' = (0, 0.071, 0.214, 0.393, 0.536, 0.571, 0.786, 0.893, 0.964, 1)^T$$

For example, for the point  $\mathbf{x}_2 = (x_{21}, x_{22}) = (14, 500)$ , the value  $x_{21} = 14$  is transformed into

$$x'_{21} = \frac{14 - 12}{28} = \frac{2}{28} = 0.071$$

Likewise, the sample range for Income is  $6000 - 300 = 5700$ , with a minimum value of 300; Income is therefore transformed into

$$\text{Income}' = (0, 0.035, 0.123, 0.298, 0.561, 0.649, 0.702, 1, 0.386, 0.421)^T$$

so that  $x_{22} = 0.035$ . The distance between  $\mathbf{x}_1$  and  $\mathbf{x}_2$  after range normalization is given as

$$\|\mathbf{x}'_1 - \mathbf{x}'_2\| = \|(0, 0)^T - (0.071, 0.035)^T\| = \|(-0.071, -0.035)^T\| = 0.079$$

We can observe that Income no longer skews the distance.

For  $z$ -normalization, we first compute the mean and standard deviation of both attributes:

|                | Age  | Income  |
|----------------|------|---------|
| $\hat{\mu}$    | 27.2 | 2680    |
| $\hat{\sigma}$ | 9.77 | 1726.15 |

Age is transformed into

$$\text{Age}' = (-1.56, -1.35, -0.94, -0.43, -0.02, 0.08, 0.70, 1.0, 1.21, 1.31)^T$$

For instance, the value  $x_{21} = 14$ , for the point  $\mathbf{x}_2 = (x_{21}, x_{22}) = (14, 500)$ , is transformed as

$$x'_{21} = \frac{14 - 27.2}{9.77} = -1.35$$

Likewise, Income is transformed into

$$\text{Income}' = (-1.38, -1.26, -0.97, -0.39, 0.48, 0.77, 0.94, 1.92, -0.10, 0.01)^T$$

so that  $x_{22} = -1.26$ . The distance between  $\mathbf{x}_1$  and  $\mathbf{x}_2$  after  $z$ -normalization is given as

$$\|\mathbf{x}'_1 - \mathbf{x}'_2\| = \|(-1.56, -1.38)^T - (1.35, -1.26)^T\| = \|(-0.18, -0.12)^T\| = 0.216$$

**Table 2.1.** Dataset for normalization

| $\mathbf{x}_i$    | Age ( $X_1$ ) | Income ( $X_2$ ) |
|-------------------|---------------|------------------|
| $\mathbf{x}_1$    | 12            | 300              |
| $\mathbf{x}_2$    | 14            | 500              |
| $\mathbf{x}_3$    | 18            | 1000             |
| $\mathbf{x}_4$    | 23            | 2000             |
| $\mathbf{x}_5$    | 27            | 3500             |
| $\mathbf{x}_6$    | 28            | 4000             |
| $\mathbf{x}_7$    | 34            | 4300             |
| $\mathbf{x}_8$    | 37            | 6000             |
| $\mathbf{x}_9$    | 39            | 2500             |
| $\mathbf{x}_{10}$ | 40            | 2700             |

## 2.5 NORMAL DISTRIBUTION

---

The normal distribution is one of the most important probability density functions, especially because many physically observed variables follow an approximately normal distribution. Furthermore, the sampling distribution of the mean of any arbitrary probability distribution follows a normal distribution. The normal distribution also plays an important role as the parametric distribution of choice in clustering, density estimation, and classification.

### 2.5.1 Univariate Normal Distribution

A random variable  $X$  has a normal distribution, with the parameters mean  $\mu$  and variance  $\sigma^2$ , if the probability density function of  $X$  is given as follows:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad (2.40)$$

The term  $(x - \mu)^2$  measures the distance of a value  $x$  from the mean  $\mu$  of the distribution, and thus the probability density decreases exponentially as a function of the distance from the mean. The maximum value of the density occurs at the mean value  $x = \mu$ , given as  $f(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}$ , which is inversely proportional to the standard deviation  $\sigma$  of the distribution.

**Example 2.7.** Figure 2.6 plots the standard normal distribution, which has the parameters  $\mu = 0$  and  $\sigma^2 = 1$ . The normal distribution has a characteristic *bell* shape, and it is symmetric about the mean. The figure also shows the effect of different values of standard deviation on the shape of the distribution. A smaller value (e.g.,  $\sigma = 0.5$ ) results in a more “peaked” distribution that decays faster, whereas a larger value (e.g.,  $\sigma = 2$ ) results in a flatter distribution that decays slower. Because the normal distribution is symmetric, the mean  $\mu$  is also the median, as well as the mode, of the distribution.

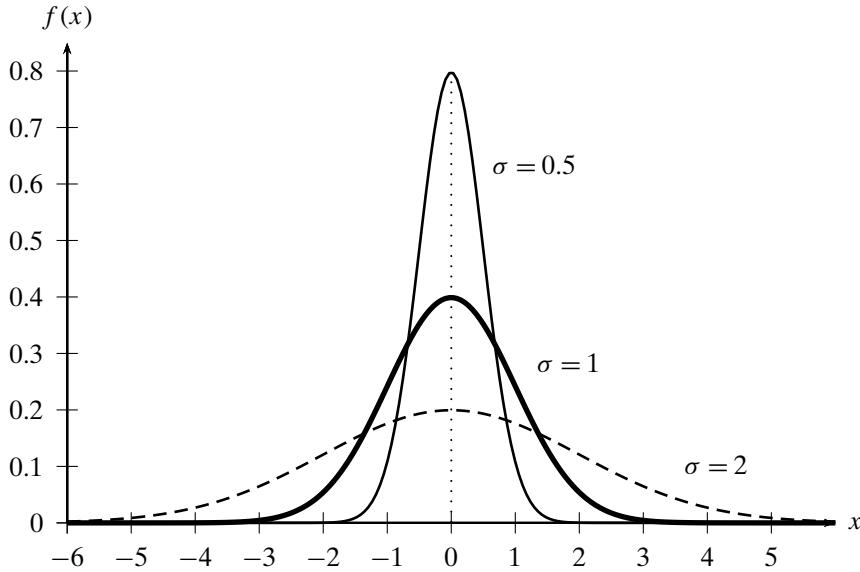


Figure 2.6. Normal distribution:  $\mu = 0$ , and different variances.

### Probability Mass

Given an interval  $[a, b]$  the probability mass of the normal distribution within that interval is given as

$$P(a \leq x \leq b) = \int_a^b f(x | \mu, \sigma^2) dx$$

In particular, we are often interested in the probability mass concentrated within  $k$  standard deviations from the mean, that is, for the interval  $[\mu - k\sigma, \mu + k\sigma]$ , which can be computed as

$$P(\mu - k\sigma \leq x \leq \mu + k\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-k\sigma}^{\mu+k\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

Via a change of variable  $z = \frac{x-\mu}{\sigma}$ , we get an equivalent formulation in terms of the standard normal distribution:

$$P(-k \leq z \leq k) = \frac{1}{\sqrt{2\pi}} \int_{-k}^k e^{-\frac{1}{2}z^2} dz = \frac{2}{\sqrt{2\pi}} \int_0^k e^{-\frac{1}{2}z^2} dz$$

The last step follows from the fact that  $e^{-\frac{1}{2}z^2}$  is symmetric, and thus the integral over the range  $[-k, k]$  is equivalent to 2 times the integral over the range  $[0, k]$ . Finally, via another change of variable  $t = \frac{z}{\sqrt{2}}$ , we get

$$P(-k \leq z \leq k) = 2 \cdot P(0 \leq t \leq k/\sqrt{2}) = \frac{2}{\sqrt{\pi}} \int_0^{k/\sqrt{2}} e^{-t^2} dt = \operatorname{erf}(k/\sqrt{2}) \quad (2.41)$$

where  $\text{erf}$  is the *Gauss error function*, defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Using Eq.(2.41) we can compute the probability mass within  $k$  standard deviations of the mean. In particular, for  $k = 1$ , we have

$$P(\mu - \sigma \leq x \leq \mu + \sigma) = \text{erf}(1/\sqrt{2}) = 0.6827$$

which means that 68.27% of all points lie within 1 standard deviation from the mean. For  $k = 2$ , we have  $\text{erf}(2/\sqrt{2}) = 0.9545$ , and for  $k = 3$  we have  $\text{erf}(3/\sqrt{2}) = 0.9973$ . Thus, almost the entire probability mass (i.e., 99.73%) of a normal distribution is within  $\pm 3\sigma$  from the mean  $\mu$ .

### 2.5.2 Multivariate Normal Distribution

Given the  $d$ -dimensional vector random variable  $\mathbf{X} = (X_1, X_2, \dots, X_d)^T$ , we say that  $\mathbf{X}$  has a multivariate normal distribution, with the parameters mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ , if its joint multivariate probability density function is given as follows:

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}|}} \exp \left\{ -\frac{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2} \right\} \quad (2.42)$$

where  $|\boldsymbol{\Sigma}|$  is the determinant of the covariance matrix. As in the univariate case, the term

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (2.43)$$

measures the distance, called the *Mahalanobis distance*, of the point  $\mathbf{x}$  from the mean  $\boldsymbol{\mu}$  of the distribution, taking into account all of the variance–covariance information between the attributes. The Mahalanobis distance is a generalization of Euclidean distance because if we set  $\boldsymbol{\Sigma} = \mathbf{I}$ , where  $\mathbf{I}$  is the  $d \times d$  identity matrix (with diagonal elements as 1's and off-diagonal elements as 0's), we get

$$(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{I}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|^2$$

The Euclidean distance thus ignores the covariance information between the attributes, whereas the Mahalanobis distance explicitly takes it into consideration.

The *standard multivariate normal distribution* has parameters  $\boldsymbol{\mu} = \mathbf{0}$  and  $\boldsymbol{\Sigma} = \mathbf{I}$ . Figure 2.7(a) plots the probability density of the standard bivariate ( $d = 2$ ) normal distribution, with parameters

$$\boldsymbol{\mu} = \mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\boldsymbol{\Sigma} = \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

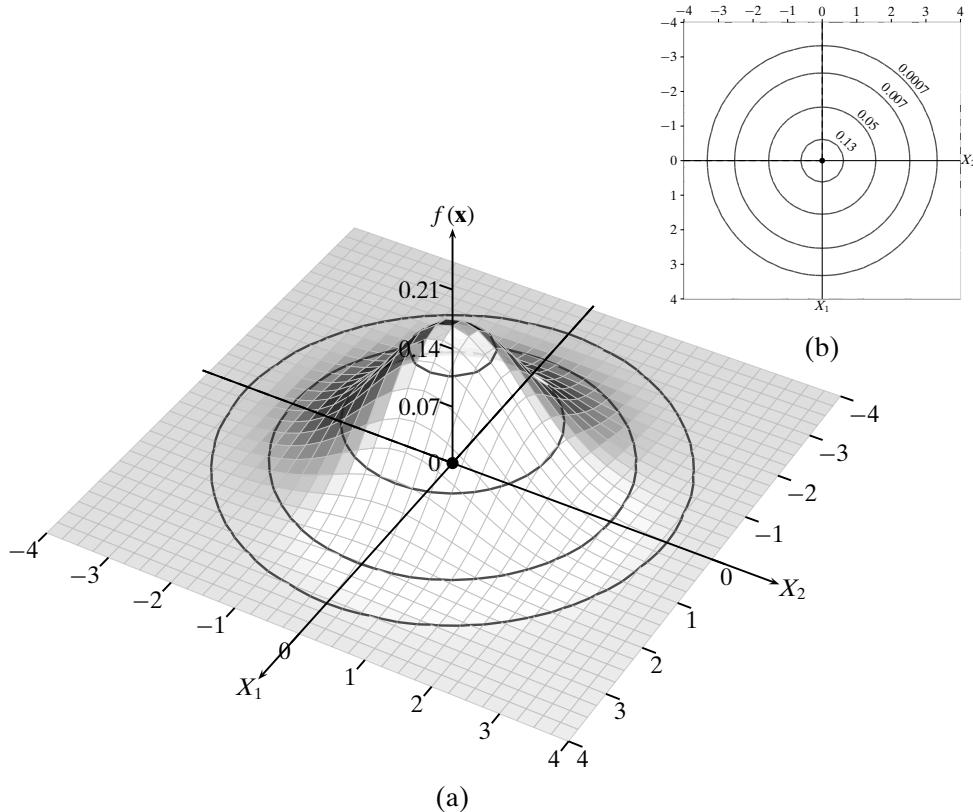


Figure 2.7. (a) Standard bivariate normal density and (b) its contour plot. Parameters:  $\mu = (0, 0)^T$ ,  $\Sigma = \mathbf{I}$ .

This corresponds to the case where the two attributes are independent, and both follow the standard normal distribution. The symmetric nature of the standard normal distribution can be clearly seen in the contour plot shown in Figure 2.7(b). Each level curve represents the set of points  $\mathbf{x}$  with a fixed density value  $f(\mathbf{x})$ .

### Geometry of the Multivariate Normal

Let us consider the geometry of the multivariate normal distribution for an arbitrary mean  $\mu$  and covariance matrix  $\Sigma$ . Compared to the standard normal distribution, we can expect the density contours to be shifted, scaled, and rotated. The shift or translation comes from the fact that the mean  $\mu$  is not necessarily the origin  $\mathbf{0}$ . The scaling or skewing is a result of the attribute variances, and the rotation is a result of the covariances.

The shape or geometry of the normal distribution becomes clear by considering the eigen-decomposition of the covariance matrix. Recall that  $\Sigma$  is a  $d \times d$  symmetric positive semidefinite matrix. The eigenvector equation for  $\Sigma$  is given as

$$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

Here  $\lambda_i$  is an eigenvalue of  $\Sigma$  and the vector  $\mathbf{u}_i \in \mathbb{R}^d$  is the eigenvector corresponding to  $\lambda_i$ . Because  $\Sigma$  is symmetric and positive semidefinite it has  $d$  real and non-negative

eigenvalues, which can be arranged in order from the largest to the smallest as follows:  $\lambda_1 \geq \lambda_2 \geq \dots \lambda_d \geq 0$ . The diagonal matrix  $\Lambda$  is used to record these eigenvalues:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

Further, the eigenvectors are unit vectors (normal) and are mutually orthogonal, that is, they are orthonormal:

$$\begin{aligned} \mathbf{u}_i^T \mathbf{u}_i &= 1 \quad \text{for all } i \\ \mathbf{u}_i^T \mathbf{u}_j &= 0 \quad \text{for all } i \neq j \end{aligned}$$

The eigenvectors can be put together into an orthogonal matrix  $\mathbf{U}$ , defined as a matrix with normal and mutually orthogonal columns:

$$\mathbf{U} = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{pmatrix}$$

The eigen-decomposition of  $\Sigma$  can then be expressed compactly as follows:

$$\Sigma = \mathbf{U} \Lambda \mathbf{U}^T$$

This equation can be interpreted geometrically as a change in basis vectors. From the original  $d$  dimensions corresponding to the  $d$  attributes  $X_j$ , we derive  $d$  new dimensions  $\mathbf{u}_i$ .  $\Sigma$  is the covariance matrix in the original space, whereas  $\Lambda$  is the covariance matrix in the new coordinate space. Because  $\Lambda$  is a diagonal matrix, we can immediately conclude that after the transformation, each new dimension  $\mathbf{u}_i$  has variance  $\lambda_i$ , and further that all covariances are zero. In other words, in the new space, the normal distribution is axis aligned (has no rotation component), but is skewed in each axis proportional to the eigenvalue  $\lambda_i$ , which represents the variance along that dimension (further details are given in [Section 7.2.4](#)).

### Total and Generalized Variance

The determinant of the covariance matrix is given as  $\det(\Sigma) = \prod_{i=1}^d \lambda_i$ . Thus, the generalized variance of  $\Sigma$  is the product of its eigenvalues.

Given the fact that the trace of a square matrix is invariant to similarity transformation, such as a change of basis, we conclude that the total variance  $\text{var}(\mathbf{D})$  for a dataset  $\mathbf{D}$  is invariant, that is,

$$\text{var}(\mathbf{D}) = \text{tr}(\Sigma) = \sum_{i=1}^d \sigma_i^2 = \sum_{i=1}^d \lambda_i = \text{tr}(\Lambda)$$

In other words  $\sigma_1^2 + \dots + \sigma_d^2 = \lambda_1 + \dots + \lambda_d$ .

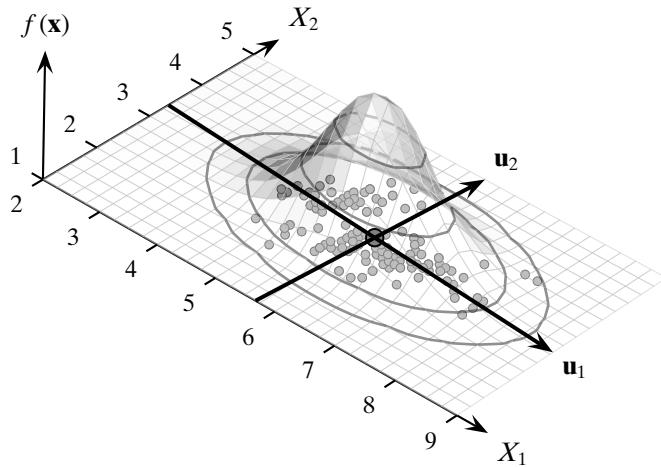


Figure 2.8. Iris: sepal length and sepal width, bivariate normal density and contours.

**Example 2.8 (Bivariate Normal Density).** Treating attributes sepal length ( $X_1$ ) and sepal width ( $X_2$ ) in the Iris dataset (see Table 1.1) as continuous random variables, we can define a continuous bivariate random variable  $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ . Assuming that  $\mathbf{X}$  follows a bivariate normal distribution, we can estimate its parameters from the sample. The sample mean is given as

$$\hat{\mu} = (5.843, 3.054)^T$$

and the sample covariance matrix is given as

$$\hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

The plot of the bivariate normal density for the two attributes is shown in Figure 2.8. The figure also shows the contour lines and the data points.

Consider the point  $\mathbf{x}_2 = (6.9, 3.1)^T$ . We have

$$\mathbf{x}_2 - \hat{\mu} = \begin{pmatrix} 6.9 \\ 3.1 \end{pmatrix} - \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix} = \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix}$$

The Mahalanobis distance between  $\mathbf{x}_2$  and  $\hat{\mu}$  is

$$\begin{aligned} (\mathbf{x}_i - \hat{\mu})^T \hat{\Sigma}^{-1} (\mathbf{x}_i - \hat{\mu}) &= (1.057 \quad 0.046) \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}^{-1} \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix} \\ &= (1.057 \quad 0.046) \begin{pmatrix} 1.486 & 0.31 \\ 0.31 & 5.42 \end{pmatrix} \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix} \\ &= 1.701 \end{aligned}$$

whereas the squared Euclidean distance between them is

$$\|(\mathbf{x}_2 - \hat{\boldsymbol{\mu}})\|^2 = (1.057 \quad 0.046) \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix} = 1.119$$

The eigenvalues and the corresponding eigenvectors of  $\hat{\Sigma}$  are as follows:

$$\begin{aligned}\lambda_1 &= 0.684 & \mathbf{u}_1 &= (-0.997, 0.078)^T \\ \lambda_2 &= 0.184 & \mathbf{u}_2 &= (-0.078, -0.997)^T\end{aligned}$$

These two eigenvectors define the new axes in which the covariance matrix is given as

$$\Lambda = \begin{pmatrix} 0.684 & 0 \\ 0 & 0.184 \end{pmatrix}$$

The angle between the original axes  $\mathbf{e}_1 = (1, 0)^T$  and  $\mathbf{u}_1$  specifies the rotation angle for the multivariate normal:

$$\begin{aligned}\cos \theta &= \mathbf{e}_1^T \mathbf{u}_1 = -0.997 \\ \theta &= \cos^{-1}(-0.997) = 175.5^\circ\end{aligned}$$

[Figure 2.8](#) illustrates the new coordinate axes and the new variances. We can see that in the original axes, the contours are only slightly rotated by angle  $175.5^\circ$  (or  $-4.5^\circ$ ).

## 2.6 FURTHER READING

---

There are several good textbooks that cover the topics discussed in this chapter in more depth; see Evans and Rosenthal (2011) and Wasserman (2004) and Rencher and Christensen (2012).

Evans, M. and Rosenthal, J. (2011). *Probability and Statistics: The Science of Uncertainty*. 2nd ed. New York: W. H. Freeman.

Rencher, A. C. and Christensen, W. F. (2012). *Methods of Multivariate Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer Science + Business Media.

## 2.7 EXERCISES

---

**Q1.** True or False:

- (a) Mean is robust against outliers.
- (b) Median is robust against outliers.
- (c) Standard deviation is robust against outliers.

- Q2.** Let  $X$  and  $Y$  be two random variables, denoting age and weight, respectively. Consider a random sample of size  $n = 20$  from these two variables

$$X = (69, 74, 68, 70, 72, 67, 66, 70, 76, 68, 72, 79, 74, 67, 66, 71, 74, 75, 75, 76)$$

$$Y = (153, 175, 155, 135, 172, 150, 115, 137, 200, 130, 140, 265, 185, 112, 140, \\ 150, 165, 185, 210, 220)$$

- (a) Find the mean, median, and mode for  $X$ .
  - (b) What is the variance for  $Y$ ?
  - (c) Plot the normal distribution for  $X$ .
  - (d) What is the probability of observing an age of 80 or higher?
  - (e) Find the 2-dimensional mean  $\hat{\mu}$  and the covariance matrix  $\hat{\Sigma}$  for these two variables.
  - (f) What is the correlation between age and weight?
  - (g) Draw a scatterplot to show the relationship between age and weight.
- Q3.** Show that the identity in [Eq. \(2.18\)](#) holds, that is,

$$\sum_{i=1}^n (x_i - \mu)^2 = n(\hat{\mu} - \mu)^2 + \sum_{i=1}^n (x_i - \hat{\mu})^2$$

- Q4.** Prove that if  $x_i$  are independent random variables, then

$$\text{var}\left(\sum_{i=1}^n x_i\right) = \sum_{i=1}^n \text{var}(x_i)$$

This fact was used in [Eq. \(2.15\)](#).

- Q5.** Define a measure of deviation called *mean absolute deviation* for a random variable  $X$  as follows:

$$\frac{1}{n} \sum_{i=1}^n |x_i - \mu|$$

Is this measure robust? Why or why not?

- Q6.** Prove that the expected value of a vector random variable  $\mathbf{X} = (X_1, X_2)^T$  is simply the vector of the expected value of the individual random variables  $X_1$  and  $X_2$  as given in [Eq. \(2.22\)](#).
- Q7.** Show that the correlation [[Eq. \(2.27\)](#)] between any two random variables  $X_1$  and  $X_2$  lies in the range  $[-1, 1]$ .
- Q8.** Given the dataset in [Table 2.2](#), compute the covariance matrix and the generalized variance.

**Table 2.2.** Dataset for Q8

|                  | $X_1$ | $X_2$ | $X_3$ |
|------------------|-------|-------|-------|
| $\mathbf{x}_1^T$ | 17    | 17    | 12    |
| $\mathbf{x}_2^T$ | 11    | 9     | 13    |
| $\mathbf{x}_3^T$ | 11    | 8     | 19    |

- Q9.** Consider [Table 2.3](#). Assume that both the attributes  $X$  and  $Y$  are numeric, and the table represents the entire population. If we know that the correlation between  $X$  and  $Y$  is zero, what can you infer about the values of  $Y$ ?

**Table 2.3.** Dataset for Q9

| $X$ | $Y$ |
|-----|-----|
| 1   | $a$ |
| 0   | $b$ |
| 1   | $c$ |
| 0   | $a$ |
| 0   | $c$ |

- Q10.** Show that the outer-product in [Eq. \(2.39\)](#) for the sample covariance matrix is equivalent to [Eq. \(2.37\)](#).
- Q11.** Assume that we are given two univariate normal distributions,  $N_A$  and  $N_B$ , and let their mean and standard deviation be as follows:  $\mu_A = 4$ ,  $\sigma_A = 1$  and  $\mu_B = 8$ ,  $\sigma_B = 2$ .
- (a) For each of the following values  $x_i \in \{5, 6, 7\}$  find out which is the more likely normal distribution to have produced it.
  - (b) Derive an expression for the point for which the probability of having been produced by both the normals is the same.
- Q12.** Under what conditions will the covariance matrix  $\Sigma$  be identical to the correlation matrix, whose  $(i, j)$  entry gives the correlation between attributes  $X_i$  and  $X_j$ ? What can you conclude about the two variables?
- Q13.** Show that the total variance of a dataset  $\mathbf{D}$  is given as the trace of the sample covariance matrix  $\hat{\Sigma}$ . That is, show that  $\text{var}(\mathbf{D}) = \text{tr}(\hat{\Sigma})$ .

In this chapter we present methods to analyze categorical attributes. Because categorical attributes have only symbolic values, many of the arithmetic operations cannot be performed directly on the symbolic values. However, we can compute the frequencies of these values and use them to analyze the attributes.

### 3.1 UNIVARIATE ANALYSIS

---

We assume that the data consists of values for a single categorical attribute,  $X$ . Let the domain of  $X$  consist of  $m$  symbolic values  $\text{dom}(X) = \{a_1, a_2, \dots, a_m\}$ . The data  $\mathbf{D}$  is thus an  $n \times 1$  symbolic data matrix given as

$$\mathbf{D} = \begin{pmatrix} X \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where each point  $x_i \in \text{dom}(X)$ .

#### 3.1.1 Bernoulli Variable

Let us first consider the case when the categorical attribute  $X$  has domain  $\{a_1, a_2\}$ , with  $m = 2$ . We can model  $X$  as a Bernoulli random variable, which takes on two distinct values, 1 and 0, according to the mapping

$$X(v) = \begin{cases} 1 & \text{if } v = a_1 \\ 0 & \text{if } v = a_2 \end{cases}$$

The probability mass function (PMF) of  $X$  is given as

$$P(X=x) = f(x) = \begin{cases} p_1 & \text{if } x = 1 \\ p_0 & \text{if } x = 0 \end{cases}$$

where  $p_1$  and  $p_0$  are the parameters of the distribution, which must satisfy the condition

$$p_1 + p_0 = 1$$

Because there is only one free parameter, it is customary to denote  $p_1 = p$ , from which it follows that  $p_0 = 1 - p$ . The PMF of Bernoulli random variable  $X$  can then be written compactly as

$$P(X=x) = f(x) = p^x(1-p)^{1-x} \quad (3.1)$$

We can see that  $P(X=1) = p^1(1-p)^0 = p$  and  $P(X=0) = p^0(1-p)^1 = 1-p$ , as desired.

### Mean and Variance

The expected value of  $X$  is given as

$$\mu = E[X] = 1 \cdot p + 0 \cdot (1-p) = p \quad (3.2)$$

and the variance of  $X$  is given as

$$\sigma^2 = \text{var}(X) = E[X^2] - (E[X])^2 = (1^2 \cdot p + 0^2 \cdot (1-p)) - p^2 = p - p^2$$

which implies

$$\sigma^2 = p(1-p) \quad (3.3)$$

### Sample Mean and Variance

To estimate the parameters of the Bernoulli variable  $X$ , we assume that each symbolic point has been mapped to its binary value. Thus, the set  $\{x_1, x_2, \dots, x_n\}$  is assumed to be a random sample drawn from  $X$  (i.e., each  $x_i$  is IID with  $X$ ).

The sample mean is given as

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{n_1}{n} = \hat{p} \quad (3.4)$$

where  $n_1$  is the number of points with  $x_i = 1$  in the random sample (equal to the number of occurrences of symbol  $a_1$ ).

Let  $n_0 = n - n_1$  denote the number of points with  $x_i = 0$  in the random sample. The sample variance is given as

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \\ &= \frac{n_1}{n} (1 - \hat{p})^2 + \frac{n - n_1}{n} (0 - \hat{p})^2 = \hat{p}(1 - \hat{p})^2 + (1 - \hat{p})\hat{p}^2 \\ &= \hat{p}(1 - \hat{p})(1 - \hat{p} + \hat{p}) = \hat{p}(1 - \hat{p}) \end{aligned}$$

The sample variance could also have been obtained directly from Eq.(3.3), by substituting  $\hat{p}$  for  $p$ .

**Example 3.1.** Consider the sepal length attribute ( $X_1$ ) for the Iris dataset in [Table 1.1](#). Let us define an Iris flower as Long if its sepal length is in the range  $[7, \infty]$ , and Short if its sepal length is in the range  $[-\infty, 7)$ . Then  $X_1$  can be treated as a categorical attribute with domain {Long, Short}. From the observed sample of size  $n = 150$ , we find 13 long Irises. The sample mean of  $X_1$  is

$$\hat{\mu} = \hat{p} = 13/150 = 0.087$$

and its variance is

$$\hat{\sigma}^2 = \hat{p}(1 - \hat{p}) = 0.087(1 - 0.087) = 0.087 \cdot 0.913 = 0.079$$

### Binomial Distribution: Number of Occurrences

Given the Bernoulli variable  $X$ , let  $\{x_1, x_2, \dots, x_n\}$  denote a random sample of size  $n$  drawn from  $X$ . Let  $N$  be the random variable denoting the number of occurrences of the symbol  $a_1$  (value  $X = 1$ ) in the sample.  $N$  has a binomial distribution, given as

$$f(N = n_1 | n, p) = \binom{n}{n_1} p^{n_1} (1-p)^{n-n_1} \quad (3.5)$$

In fact,  $N$  is the sum of the  $n$  independent Bernoulli random variables  $x_i$  IID with  $X$ , that is,  $N = \sum_{i=1}^n x_i$ . By linearity of expectation, the mean or expected number of occurrences of symbol  $a_1$  is given as

$$\mu_N = E[N] = E \left[ \sum_{i=1}^n x_i \right] = \sum_{i=1}^n E[x_i] = \sum_{i=1}^n p = np \quad (3.6)$$

Because  $x_i$  are all independent, the variance of  $N$  is given as

$$\sigma_N^2 = \text{var}(N) = \sum_{i=1}^n \text{var}(x_i) = \sum_{i=1}^n p(1-p) = np(1-p) \quad (3.7)$$

**Example 3.2.** Continuing with [Example 3.1](#), we can use the estimated parameter  $\hat{p} = 0.087$  to compute the expected number of occurrences  $N$  of Long sepal length Irises via the binomial distribution:

$$E[N] = n\hat{p} = 150 \cdot 0.087 = 13$$

In this case, because  $p$  is estimated from the sample via  $\hat{p}$ , it is not surprising that the expected number of occurrences of long Irises coincides with the actual occurrences.

However, what is more interesting is that we can compute the variance in the number of occurrences:

$$\text{var}(N) = n\hat{p}(1 - \hat{p}) = 150 \cdot 0.079 = 11.9$$

As the sample size increases, the binomial distribution given in Eq. 3.5 tends to a normal distribution with  $\mu = 13$  and  $\sigma = \sqrt{11.9} = 3.45$  for our example. Thus, with confidence greater than 95% we can claim that the number of occurrences of  $a_1$  will lie in the range  $\mu \pm 2\sigma = [9.55, 16.45]$ , which follows from the fact that for a normal distribution 95.45% of the probability mass lies within two standard deviations from the mean (see Section 2.5.1).

### 3.1.2 Multivariate Bernoulli Variable

We now consider the general case when  $X$  is a categorical attribute with domain  $\{a_1, a_2, \dots, a_m\}$ . We can model  $X$  as an  $m$ -dimensional Bernoulli random variable  $\mathbf{X} = (A_1, A_2, \dots, A_m)^T$ , where each  $A_i$  is a Bernoulli variable with parameter  $p_i$  denoting the probability of observing symbol  $a_i$ . However, because  $X$  can assume only one of the symbolic values at any one time, if  $X = a_i$ , then  $A_i = 1$ , and  $A_j = 0$  for all  $j \neq i$ . The range of the random variable  $\mathbf{X}$  is thus the set  $\{0, 1\}^m$ , with the further restriction that if  $X = a_i$ , then  $\mathbf{X} = \mathbf{e}_i$ , where  $\mathbf{e}_i$  is the  $i$ th standard basis vector  $\mathbf{e}_i \in \mathbb{R}^m$ , given as

$$\mathbf{e}_i = (\overbrace{0, \dots, 0}^{i-1}, 1, \overbrace{0, \dots, 0}^{m-i})^T$$

In  $\mathbf{e}_i$ , only the  $i$ th element is 1 ( $e_{ii} = 1$ ), whereas all other elements are zero ( $e_{ij} = 0, \forall j \neq i$ ).

This is precisely the definition of a *multivariate Bernoulli variable*, which is a generalization of a Bernoulli variable from two outcomes to  $m$  outcomes. We thus model the categorical attribute  $X$  as a multivariate Bernoulli variable  $\mathbf{X}$  defined as

$$\mathbf{X}(v) = \mathbf{e}_i \text{ if } v = a_i$$

This is also referred to as a *one-hot encoding* of the variable  $X$ . The range of  $\mathbf{X}$  consists of  $m$  distinct vector values  $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$ , with the PMF of  $\mathbf{X}$  given as

$$P(\mathbf{X} = \mathbf{e}_i) = f(\mathbf{e}_i) = p_i$$

where  $p_i$  is the probability of observing value  $a_i$ . These parameters must satisfy the condition

$$\sum_{i=1}^m p_i = 1$$