DATA MINING AND ANALYSIS

The fundamental algorithms in data mining and analysis form the basis for the emerging field of data science, which includes automated methods to analyze patterns and models for all kinds of data, with applications ranging from scientific discovery to business intelligence and analytics. This textbook for senior undergraduate and graduate data mining courses provides a broad yet in-depth overview of data mining, integrating related concepts from machine learning and statistics. The main parts of the book include exploratory data analysis, pattern mining, clustering, and classification. The book lays the basic foundations of these tasks and also covers cutting-edge topics such as kernel methods, high-dimensional data analysis, and complex graphs and networks. With its comprehensive coverage, algorithmic perspective, and wealth of examples, this book offers solid guidance in data mining for students, researchers, and practitioners alike.

Key Features:

- Covers both core methods and cutting-edge research
- Algorithmic approach with open-source implementations
- Minimal prerequisites, as all key mathematical concepts are presented, as is the intuition behind the formulas
- Short, self-contained chapters with class-tested examples and exercises that allow for flexibility in designing a course and for easy reference
- Supplementary online resource containing lecture slides, videos, project ideas, and more

Mohammed J. Zaki is a Professor of Computer Science at Rensselaer Polytechnic Institute, Troy, New York.

Wagner Meira Jr. is a Professor of Computer Science at Universidade Federal de Minas Gerais, Brazil.

DATA MINING AND ANALYSIS

Fundamental Concepts and Algorithms

MOHAMMED J. ZAKI
Rensselaer Polytechnic Institute, Troy, New York

WAGNER MEIRA JR.
Universidade Federal de Minas Gerais, Brazil

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

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
Information on this title: www.cambridge.org/9780521766333

First published 2014

Printed in the United States of America

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

Library of Congress Cataloging in Publication Data
Zaki, Mohammed J., 1971–
Data mining and analysis: fundamental concepts and algorithms / Mohammed J. Zaki, Rensselaer Polytechnic Institute, Troy, New York, Wagner Meira Jr., Universidade Federal de Minas Gerais, Brazil.
pages cm
Includes bibliographical references and index.
ISBN 978-0-521-76633-3 (hardback)
1. Data mining. I. Meira, Wagner, 1967– II. Title.
QA76.9.D343Z36 2014
006.3′12–dc23 2013037544

ISBN 978-0-521-76633-3 Hardback

Contents

Preface

This book is an outgrowth of data mining courses at Rensselaer Polytechnic Institute (RPI) and Universidade Federal de Minas Gerais (UFMG); the RPI course has been offered every Fall since 1998, whereas the UFMG course has been offered since 2002. Although there are several good books on data mining and related topics, we felt that many of them are either too high-level or too advanced. Our goal was to write an introductory text that focuses on the fundamental algorithms in data mining and analysis. It lays the mathematical foundations for the core data mining methods, with key concepts explained when first encountered; the book also tries to build the intuition behind the formulas to aid understanding.

The main parts of the book include exploratory data analysis, frequent pattern mining, clustering, and classification. The book lays the basic foundations of these tasks, and it also covers cutting-edge topics such as kernel methods, high-dimensional data analysis, and complex graphs and networks. It integrates concepts from related disciplines such as machine learning and statistics and is also ideal for a course on data analysis. Most of the prerequisite material is covered in the text, especially on linear algebra, and probability and statistics.

The book includes many examples to illustrate the main technical concepts. 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. We suggest that readers use their favorite data analysis and mining software to work through our examples and to implement the algorithms we describe in text; we recommend the R software or the Python language with its NumPy package. The datasets used and other supplementary material such as project ideas and slides are available online at the book's companion site and its mirrors at RPI and UFMG:

- `http://dataminingbook.info`
- `http://www.cs.rpi.edu/~zaki/dataminingbook`
- `http://www.dcc.ufmg.br/dataminingbook`

Having understood the basic principles and algorithms in data mining and data analysis, readers will be well equipped to develop their own methods or use more advanced techniques.

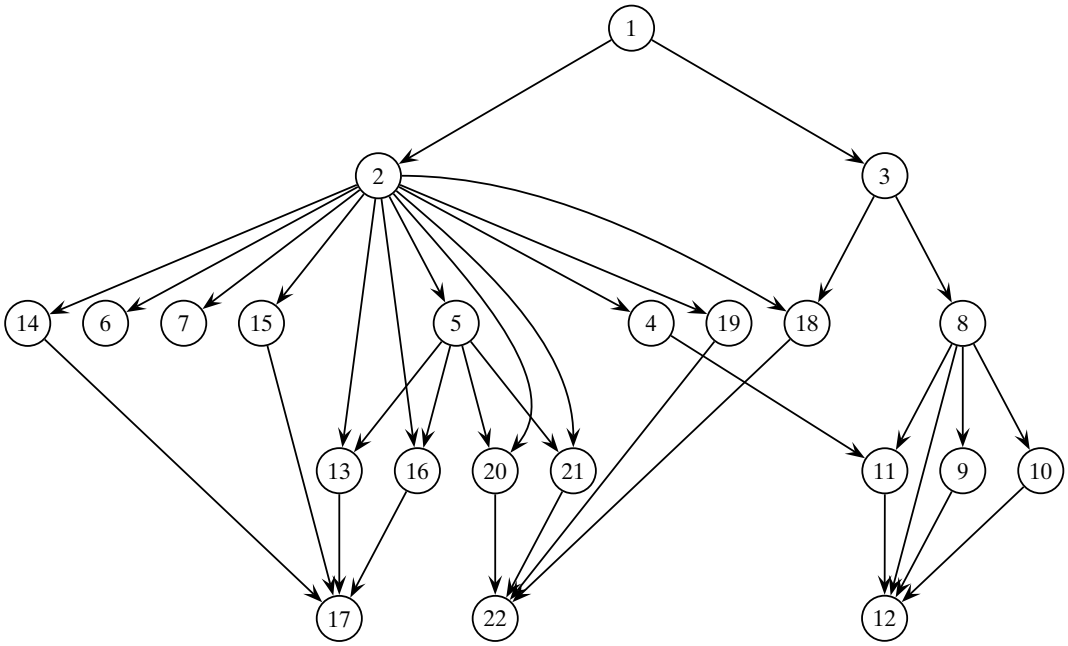


Figure 0.1. Chapter dependencies

Suggested Roadmaps

The chapter dependency graph is shown in Figure 0.1. We suggest some typical roadmaps for courses and readings based on this book. For an undergraduate-level course, we suggest the following chapters: 1–3, 8, 10, 12–15, 17–19, and 21–22. For an undergraduate course without exploratory data analysis, we recommend Chapters 1, 8–15, 17–19, and 21–22. For a graduate course, one possibility is to quickly go over the material in Part I or to assume it as background reading and to directly cover Chapters 9–22; the other parts of the book, namely frequent pattern mining (Part II), clustering (Part III), and classification (Part IV), can be covered in any order. For a course on data analysis the chapters covered must include 1–7, 13–14, 15 (Section 2), and 20. Finally, for a course with an emphasis on graphs and kernels we suggest Chapters 4, 5, 7 (Sections 1–3), 11–12, 13 (Sections 1–2), 16–17, and 20–22.

Acknowledgments

Initial drafts of this book have been used in several data mining courses. We received many valuable comments and corrections from both the faculty and students. Our thanks go to

- Muhammad Abulaish, Jamia Millia Islamia, India
- Mohammad Al Hasan, Indiana University Purdue University at Indianapolis
- Marcio Luiz Bunte de Carvalho, Universidade Federal de Minas Gerais, Brazil
- Loïc Cerf, Universidade Federal de Minas Gerais, Brazil
- Ayhan Demiriz, Sakarya University, Turkey
- Murat Dundar, Indiana University Purdue University at Indianapolis
- Jun Luke Huan, University of Kansas
- Ruoming Jin, Kent State University
- Latifur Khan, University of Texas, Dallas

- Pauli Miettinen, Max-Planck-Institut für Informatik, Germany
- Suat Ozdemir, Gazi University, Turkey
- Naren Ramakrishnan, Virginia Polytechnic and State University
- Leonardo Chaves Dutra da Rocha, Universidade Federal de São João del-Rei, Brazil
- Saeed Salem, North Dakota State University
- Ankur Teredesai, University of Washington, Tacoma
- Hannu Toivonen, University of Helsinki, Finland
- Adriano Alonso Veloso, Universidade Federal de Minas Gerais, Brazil
- Jason T.L. Wang, New Jersey Institute of Technology
- Jianyong Wang, Tsinghua University, China
- Jiong Yang, Case Western Reserve University
- Jieping Ye, Arizona State University

We would like to thank all the students enrolled in our data mining courses at RPI and UFMG, as well as the anonymous reviewers who provided technical comments on various chapters. We appreciate the collegial and supportive environment within the computer science departments at RPI and UFMG and at the Qatar Computing Research Institute. In addition, we thank NSF, CNPq, CAPES, FAPEMIG, Inweb – the National Institute of Science and Technology for the Web, and Brazil's Science without Borders program for their support. We thank Lauren Cowles, our editor at Cambridge University Press, for her guidance and patience in realizing this book.

Finally, on a more personal front, MJZ dedicates the book to his wife, Amina, for her love, patience and support over all these years, and to his children, Abrar and Afsah, and his parents. WMJ gratefully dedicates the book to his wife Patricia; to his children, Gabriel and Marina; and to his parents, Wagner and Marlene, for their love, encouragement, and inspiration.

CHAPTER 1 Data Mining and Analysis

Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data. 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. We then discuss the main data mining tasks, which span exploratory data analysis, frequent pattern mining, clustering, and classification, laying out the roadmap for the book.

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

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

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

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

and X_j denotes the jth column, which is an n-tuple given as

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

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

Table 1.1. Extract from the Iris dataset

	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class 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

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*.

Example 1.1. Table 1.1 shows an extract of the Iris dataset; the complete data forms a 150×5 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

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

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 $domain(\texttt{Age}) = \mathbb{N}$, where $\mathbb{N}$ denotes the set of natural numbers (non-negative integers), is numeric, and so is `petal length` in Table 1.1, with $domain(\texttt{petal length}) = \mathbb{R}^+$ (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 °C or °F is interval-scaled. If it is 20 °C on one day and 10 °C on the following day, it is meaningful to talk about a temperature drop of 10 °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

$$domain(\texttt{Sex}) = \{\texttt{M}, \texttt{F}\}$$
$$domain(\texttt{Education}) = \{\texttt{HighSchool}, \texttt{BS}, \texttt{MS}, \texttt{PhD}\}$$

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 $domain(\texttt{class}) = \{\texttt{iris-setosa}, \texttt{iris-versicolor}, \texttt{iris-virginica}\}$.
- *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:

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

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

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

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 jth *standard basis vector* $\mathbf{e}_j$ is the d-dimensional unit vector whose jth component is 1 and the rest of the components are 0

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

Any other vector in $\mathbb{R}^d$ can be written as *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 jth 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 $\mathbf{x}_1 = (5.9, 3.0, 4.2, 1.5, \texttt{Iris-versicolor})$ on the first two attributes is shown in Figure 1.1a. 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 1.1b 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 $\mathbb{R}^3$:

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

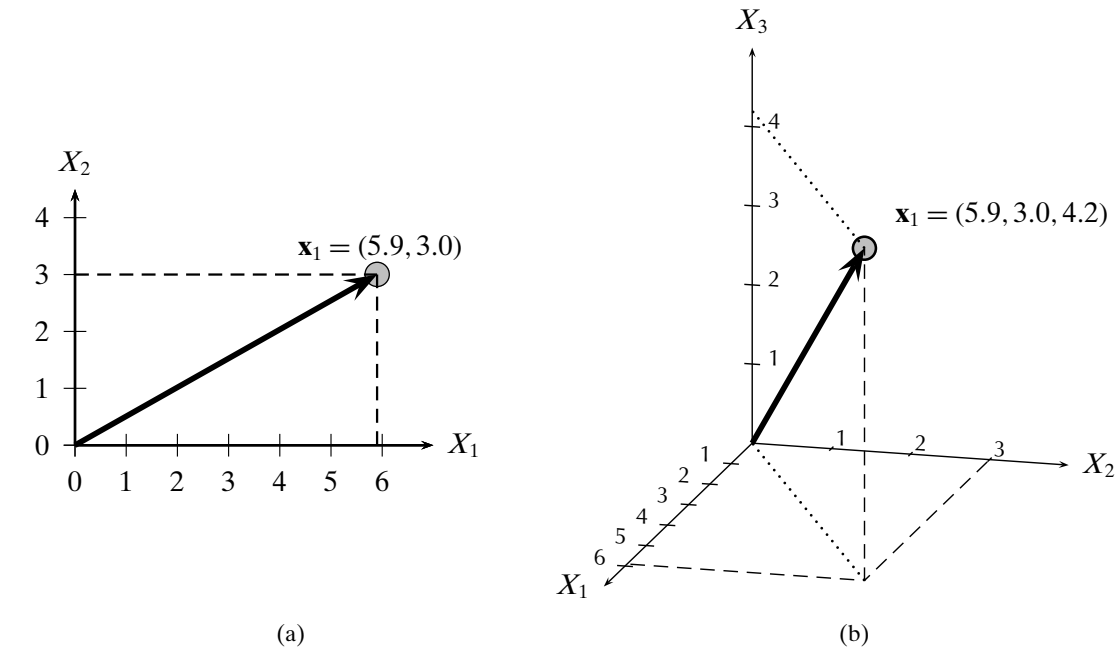


Figure 1.1. Row $\mathbf{x}_1$ as a point and vector in (a) $\mathbb{R}^2$ and (b) $\mathbb{R}^3$.

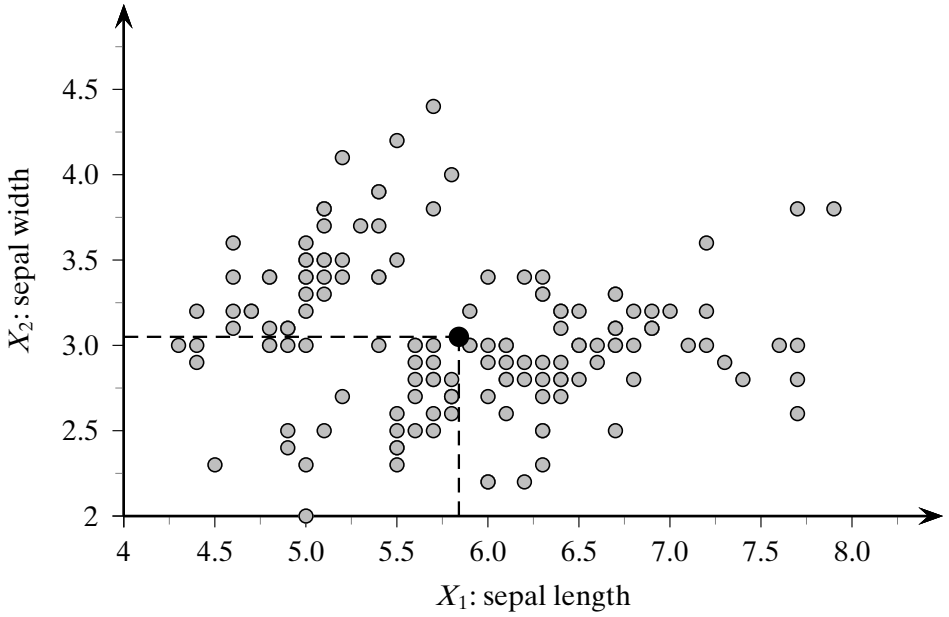


Figure 1.2. Scatterplot: `sepal length` versus `sepal width`. The solid circle shows the mean point.

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

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

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

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

Dot Product

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

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

Length

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

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

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

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

for any $p \neq 0$. Thus, the Euclidean norm corresponds to the case when $p = 2$.

Distance

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

$$\delta(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\| = \sqrt{(\mathbf{a} - \mathbf{b})^T(\mathbf{a} - \mathbf{b})} = \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2} \tag{1.1}$$

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

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

$$\delta_p(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_p \tag{1.2}$$

If p is unspecified, as in Eq. (1.1), 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

$$\cos\theta = \frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|} = \left(\frac{\mathbf{a}}{\|\mathbf{a}\|}\right)^T \left(\frac{\mathbf{b}}{\|\mathbf{b}\|}\right) \tag{1.3}$$

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

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

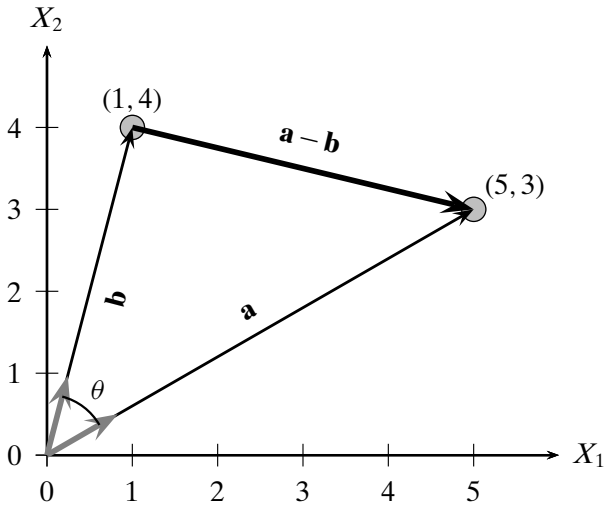


Figure 1.3. Distance and angle. Unit vectors are shown in gray.

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°, to -1, corresponding to an angle of 180° (or π 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° 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

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

Using Eq. (1.1), the Euclidean distance between them is given as

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

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 **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. (1.3) the cosine of the angle between **a** and **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:

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

Let us consider the L_p-norm for **a** with $p = 3$; we get

$$\|\mathbf{a}\|_3 = \left(5^3 + 3^3\right)^{1/3} = (153)^{1/3} = 5.34$$

The distance between **a** and **b** using Eq. (1.2) for the L_p-norm with $p = 3$ is given as

$$\|\mathbf{a} - \mathbf{b}\|_3 = \left\|(4,-1)^T\right\|_3 = \left(4^3 + (-1)^3\right)^{1/3} = (63)^{1/3} = 3.98$$

1.3.2 Mean and Total Variance

Mean

The *mean* of the data matrix **D** is the vector obtained as the average of all the points:

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

Total Variance

The *total variance* of the data matrix **D** is the average squared distance of each point from the mean:

$$var(\mathbf{D}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{x}_i, \boldsymbol{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \tag{1.4}$$

Simplifying Eq. (1.4) we obtain

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

$$= \frac{1}{n}\left(\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - 2n\boldsymbol{\mu}^T\boldsymbol{\mu} + n\|\boldsymbol{\mu}\|^2\right)$$
$$= \frac{1}{n}\left(\sum_{i=1}^{n} \|\mathbf{x}_i\|^2\right) - \|\boldsymbol{\mu}\|^2$$

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:

$$\mathbf{Z} = \mathbf{D} - \mathbf{1}\cdot\boldsymbol{\mu}^T = \begin{pmatrix}\mathbf{x}_1^T\\ \mathbf{x}_2^T\\ \vdots\\ \mathbf{x}_n^T\end{pmatrix} - \begin{pmatrix}\boldsymbol{\mu}^T\\ \boldsymbol{\mu}^T\\ \vdots\\ \boldsymbol{\mu}^T\end{pmatrix} = \begin{pmatrix}\mathbf{x}_1^T - \boldsymbol{\mu}^T\\ \mathbf{x}_2^T - \boldsymbol{\mu}^T\\ \vdots\\ \mathbf{x}_n^T - \boldsymbol{\mu}^T\end{pmatrix} = \begin{pmatrix}\mathbf{z}_1^T\\ \mathbf{z}_2^T\\ \vdots\\ \mathbf{z}_n^T\end{pmatrix} \tag{1.5}$$

where $\mathbf{z}_i = \mathbf{x}_i - \boldsymbol{\mu}$ 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 $\mathbf{Z}$ 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

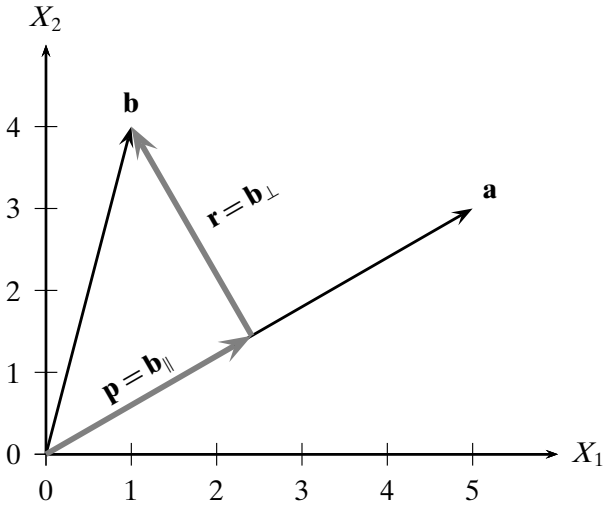


Figure 1.4. Orthogonal projection.

of another vector $\mathbf{a}$, illustrated in Figure 1.4, is given as

$$\mathbf{b} = \mathbf{b}_{\parallel} + \mathbf{b}_{\perp} = \mathbf{p} + \mathbf{r} \tag{1.6}$$

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 vector between the points $\mathbf{b}$ and $\mathbf{p}$.

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

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

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

$$\mathbf{p} = \mathbf{b}_{\parallel} = c\mathbf{a} = \left(\frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}}\right)\mathbf{a} \tag{1.7}$$

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

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

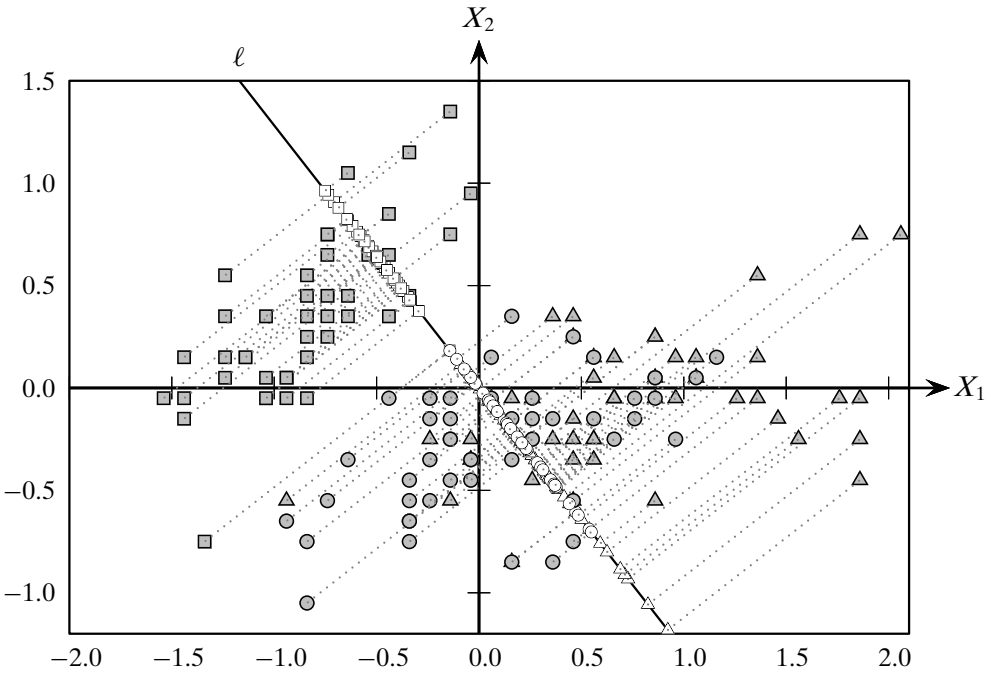


Figure 1.5. Projecting the centered data onto the line ℓ.

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 $var(\mathbf{D}) = 0.868$ (centering does not change this value).

Figure 1.5 shows the projection of each point onto the line ℓ, 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 ℓ 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

$$\mathbf{D} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{pmatrix}^T = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \end{pmatrix}$$

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 $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ 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 $span(\mathbf{v}_1, \ldots, \mathbf{v}_k)$, which is itself a vector space being a *subspace* of $\mathbb{R}^m$. If $span(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \mathbb{R}^m$, then we say that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ 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 $col(\mathbf{D})$, is the set of all linear combinations of the d attributes $X_j \in \mathbb{R}^n$, that is,

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

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

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

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

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

Linear Independence

We say that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ 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 $c_1, c_2, \ldots, c_k$, 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, $\mathbf{v}_1, \cdots, \mathbf{v}_k$ 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 $\mathbf{v}_1, \ldots, \mathbf{v}_k$, that are linearly independent and they span S, that is, $span(\mathbf{v}_1, \ldots, \mathbf{v}_k) = S$. 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

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

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 $rank(\mathbf{D}) \leq \min(n, d)$, 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 $rank(\mathbf{D}) < d$, then the data points reside in a lower dimensional subspace of $\mathbb{R}^d$, and in this case $rank(\mathbf{D})$ 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 ℓ in Figure 1.5 is given as $\ell = span\left(\begin{pmatrix} -2.15 & 2.75 \end{pmatrix}^T\right)$, with $dim(\ell) = 1$. After normalization, we obtain the orthonormal basis for ℓ as the unit vector

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

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

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 $X: \mathcal{O} \to \mathbb{R}$, 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 $X: \mathcal{O} \to \mathcal{O}$ is simply the identity function: $X(v) = v$ for all $v \in \mathcal{O}$. The distinction between the outcomes and the value of the 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

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

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:

$$f(x) \geq 0$$

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 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 a trials falls off rapidly as k increases, becoming practically zero for values of $k \geq 6$.

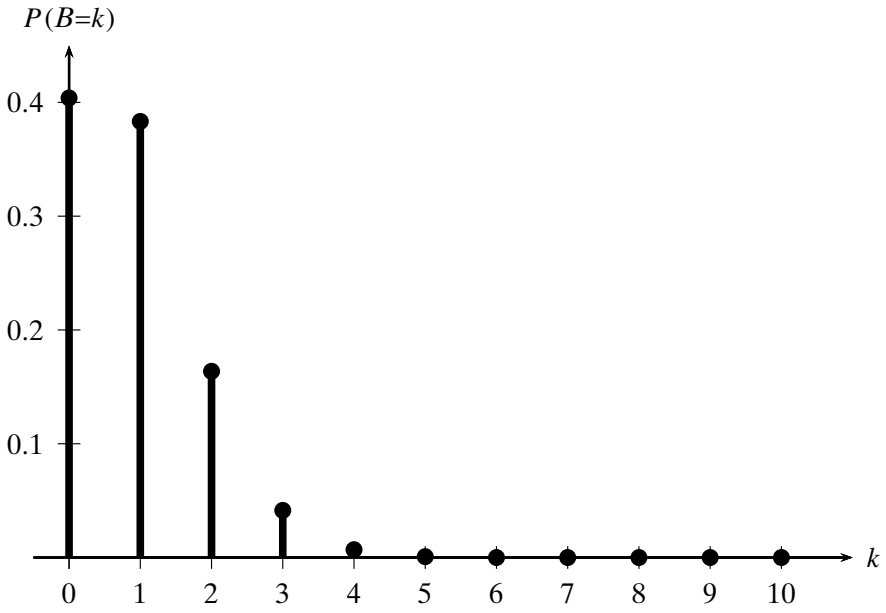


Figure 1.6. Binomial distribution: probability mass function ($m = 10, p = 0.087$).

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 $[a, b] \subset \mathbb{R}$:

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

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

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

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

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

$$P\big(X \in [x-\epsilon, x+\epsilon]\big) = \int_{x-\epsilon}^{x+\epsilon} f(x)\, dx \simeq 2\epsilon \cdot f(x)$$

$$f(x) \simeq \frac{P\big(X \in [x-\epsilon, x+\epsilon]\big)}{2\epsilon} \tag{1.8}$$

$f(x)$ 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. (1.8), we have

$$\frac{P(X \in [x_1-\epsilon, x_1+\epsilon])}{P(X \in [x_2-\epsilon, x_2+\epsilon])} \simeq \frac{2\epsilon \cdot f(x_1)}{2\epsilon \cdot f(x_2)} = \frac{f(x_1)}{f(x_2)} \tag{1.9}$$

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, μ, which represents the mean value, and σ^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. (1.8), we have

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

As $\epsilon \to 0$, we get $P(X=\mu) \to 0$. However, based on Eq. (1.9) 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

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

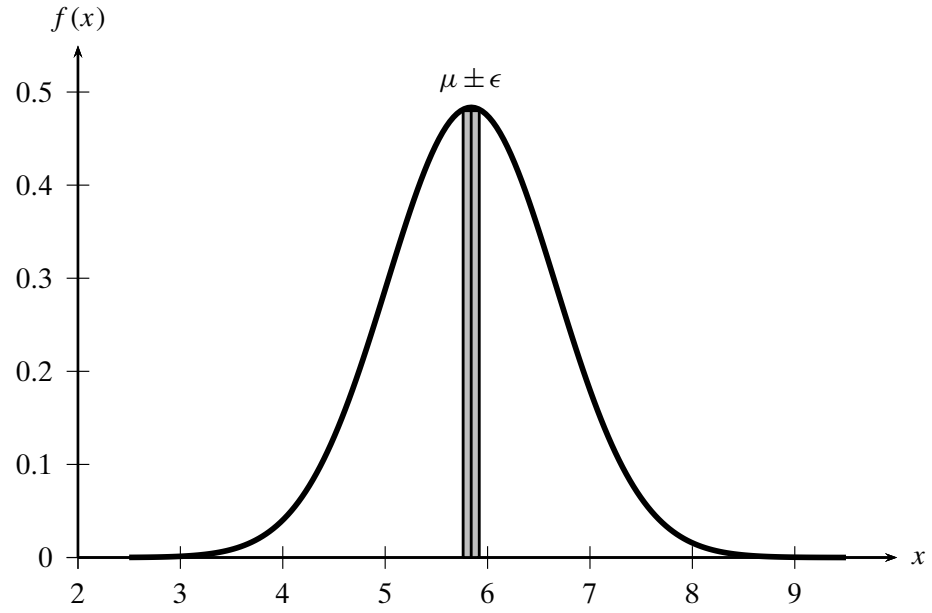


Figure 1.7. Normal distribution: probability density function ($\mu = 5.84$, $\sigma^2 = 0.681$).

Cumulative Distribution Function

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

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

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 $F(\mu) = P(X \leq \mu) = 0.5$.

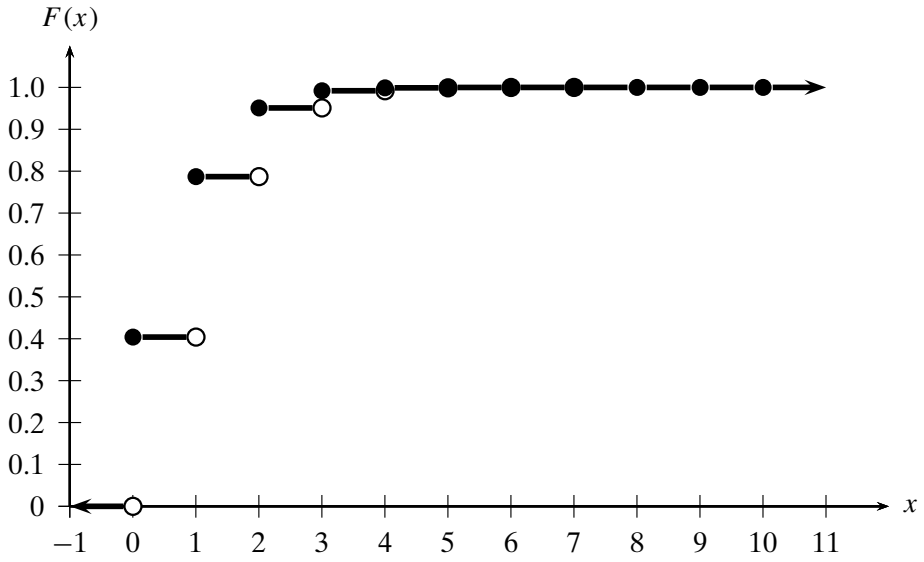


Figure 1.8. Cumulative distribution function for the binomial distribution.

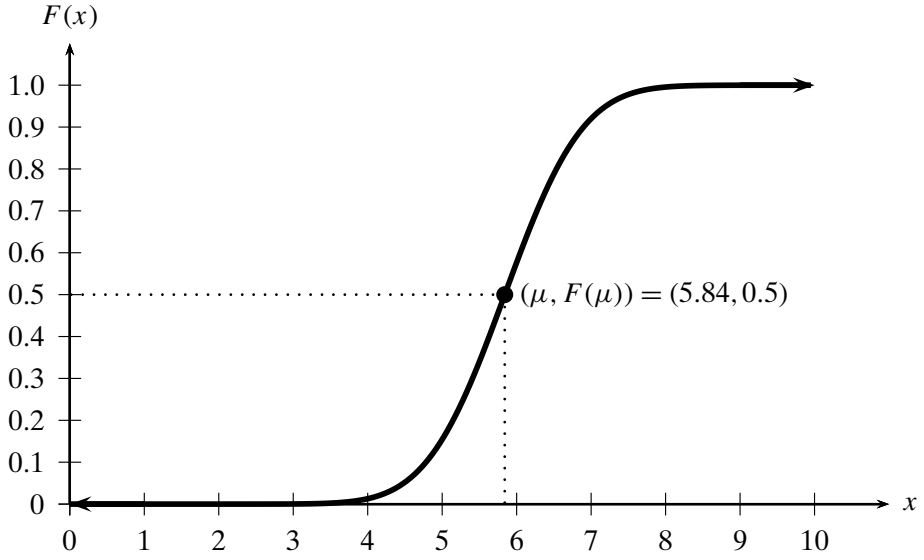


Figure 1.9. Cumulative distribution function for the normal distribution.

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:

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

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

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:

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

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

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

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

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:

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

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

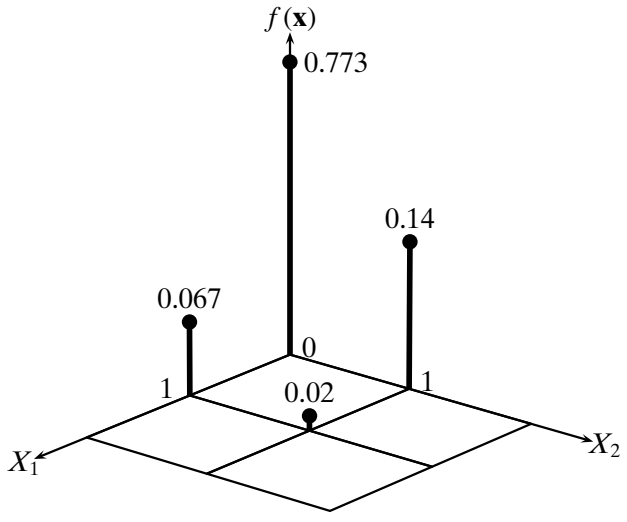


Figure 1.10. Joint probability mass function: X_1 (long sepal length), X_2 (long sepal width).

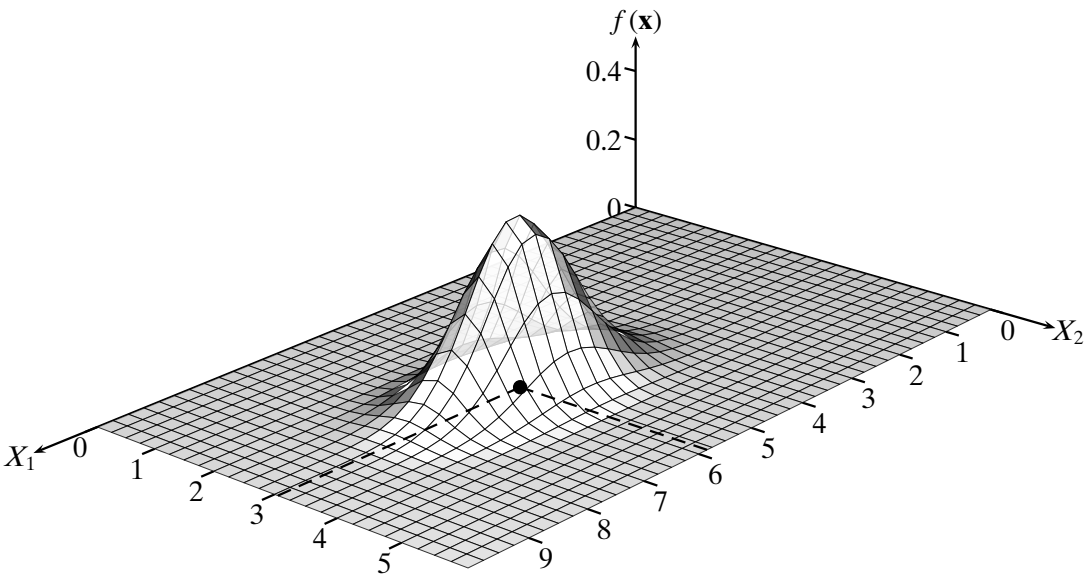


Figure 1.11. Bivariate normal density: $\boldsymbol{\mu} = (5.843, 3.054)^T$ (solid circle).

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

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

and covariance matrix

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

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* $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$, 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, $\mathbf{X} : \mathcal{O} \to \mathbb{R}^d$. The range of $\mathbf{X}$ can be denoted as a vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$. 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

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

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

$$P(\mathbf{X} \in W) = \int_{\mathbf{x} \in W} \cdots \int f(\mathbf{x})\, d\mathbf{x}$$
$$P\left((X_1, X_2, \ldots, X_d)^T \in W\right) = \int_{(x_1, x_2, \ldots, x_d)^T \in W} \cdots \int f(x_1, x_2, \ldots, x_d)\, dx_1\, dx_2 \ldots dx_d$$

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 $\mathbf{X} = (X_1, \ldots, X_d)^T$ is given as

$$F(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x})$$
$$F(x_1, x_2, \ldots, x_d) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_d)$$

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

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

$$P(X_1 \in W_1 \text{ and } X_2 \in W_2 \cdots \text{ and } X_d \in W_d) = P(X_1 \in W_1) \cdot P(X_2 \in W_2) \cdot \cdots \cdot P(X_d \in W_d) \quad (1.10)$$

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

$$F(\mathbf{x}) = F(x_1, \ldots, x_d) = F_1(x_1) \cdot F_2(x_2) \cdot \ldots \cdot F_d(x_d)$$
$$f(\mathbf{x}) = f(x_1, \ldots, x_d) = f_1(x_1) \cdot f_2(x_2) \cdot \ldots \cdot f_d(x_d) \quad (1.11)$$

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 $S_1, S_2, \ldots, S_n$, 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. (1.11) their joint probability function is given as

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

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

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 $\mathbf{X} = (X_1, X_2, \ldots, X_d)$. That is, $\mathbf{x}_i$ are assumed to be independent and identically distributed, and thus their joint distribution is given as

$$f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f_{\mathbf{X}}(\mathbf{x}_i) \tag{1.12}$$

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 $X_1, X_2, \ldots, X_d$ 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. (1.12) can be rewritten as

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

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 the random sample of size m drawn from a (multivariate) random variable $\mathbf{X}$. A statistic $\hat{\theta}$ is a function $\hat{\theta} : (\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_m) \to \mathbb{R}$. The statistic is an estimate of the corresponding population parameter θ. 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

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

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

1.5 DATA MINING

Data mining comprises the core algorithms that enable one to gain fundamental insights and knowledge from massive data. It is an interdisciplinary field merging concepts from allied areas such as database systems, statistics, machine learning, and pattern recognition. In fact, data mining is part of a larger knowledge discovery process, which includes pre-processing tasks such as data extraction, data cleaning, data fusion, data reduction and feature construction, as well as post-processing steps

such as pattern and model interpretation, hypothesis confirmation and generation, and so on. This knowledge discovery and data mining process tends to be highly iterative and interactive.

The algebraic, geometric, and probabilistic viewpoints of data play a key role in data mining. Given a dataset of n points in a d-dimensional space, the fundamental analysis and mining tasks covered in this book include exploratory data analysis, frequent pattern discovery, data clustering, and classification models, which are described next.

1.5.1 Exploratory Data Analysis

Exploratory data analysis aims to explore the numeric and categorical attributes of the data individually or jointly to extract key characteristics of the data sample via statistics that give information about the centrality, dispersion, and so on. Moving away from the IID assumption among the data points, it is also important to consider the statistics that deal with the data as a graph, where the nodes denote the points and weighted edges denote the connections between points. This enables one to extract important topological attributes that give insights into the structure and models of networks and graphs. Kernel methods provide a fundamental connection between the independent pointwise view of data, and the viewpoint that deals with pairwise similarities between points. Many of the exploratory data analysis and mining tasks can be cast as kernel problems via the *kernel trick*, that is, by showing that the operations involve only dot-products between pairs of points. However, kernel methods also enable us to perform nonlinear analysis by using familiar linear algebraic and statistical methods in high-dimensional spaces comprising "nonlinear" dimensions. They further allow us to mine complex data as long as we have a way to measure the pairwise similarity between two abstract objects. Given that data mining deals with massive datasets with thousands of attributes and millions of points, another goal of exploratory analysis is to reduce the amount of data to be mined. For instance, feature selection and dimensionality reduction methods are used to select the most important dimensions, discretization methods can be used to reduce the number of values of an attribute, data sampling methods can be used to reduce the data size, and so on.

Part I of this book 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. Next, 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 Erdös–Rényi random graph model, the Watts–Strogatz model, and the Barabási–Albert model. Kernel methods are then introduced in Chapter 5, which provide 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 surface 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 the 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.

1.5.2 Frequent Pattern Mining

Frequent pattern mining refers to the task of extracting informative and useful patterns in massive and complex datasets. Patterns comprise sets of co-occurring attribute values, called *itemsets*, or more complex patterns, such as sequences, which consider explicit precedence relationships (either positional or temporal), and graphs, which consider arbitrary relationships between points. The key goal is to discover hidden trends and behaviors in the data to understand better the interactions among the points and attributes.

Part II begins by presenting efficient algorithms for frequent itemset mining in Chapter 8. The key methods include the level-wise Apriori algorithm, the "vertical" intersection based Eclat algorithm, and the frequent pattern tree and projection based FPGrowth method. Typically the mining process results in too many frequent patterns that can be hard to interpret. In Chapter 9 we consider approaches to summarize the mined patterns; these include maximal (GenMax algorithm), closed (Charm algorithm), and non-derivable itemsets. We describe effective methods for frequent sequence mining in Chapter 10, which include the level-wise GSP method, the vertical SPADE algorithm, and the projection-based PrefixSpan approach. We also describe how consecutive subsequences, also called substrings, can be mined much more efficiently via Ukkonen's linear time and space suffix tree method. Moving

beyond sequences to arbitrary graphs, we describe the popular and efficient gSpan algorithm for frequent subgraph mining in Chapter 11. Graph mining involves two key steps, namely graph isomorphism checks to eliminate duplicate patterns during pattern enumeration and subgraph isomorphism checks during frequency computation. These operations can be performed in polynomial time for sets and sequences, but for graphs it is known that subgraph isomorphism is NP-hard, and thus there is no polynomial time method possible unless P = NP. The gSpan method proposes a new canonical code and a systematic approach to subgraph extension, which allow it to efficiently detect duplicates and to perform several subgraph isomorphism checks much more efficiently than performing them individually. Given that pattern mining methods generate many output results it is very important to assess the mined patterns. We discuss strategies for assessing both the frequent patterns and rules that can be mined from them in Chapter 12, emphasizing methods for significance testing.

1.5.3 Clustering

Clustering is the task of partitioning the points into *natural groups* called clusters, such that points within a group are very similar, whereas points across clusters are as dissimilar as possible. Depending on the data and desired cluster characteristics, there are different types of clustering paradigms such as representative-based, hierarchical, density-based, graph-based, and spectral clustering.

Part III starts with representative-based clustering methods (Chapter 13), which include the K-means and Expectation-Maximization (EM) algorithms. K-means is a greedy algorithm that minimizes the squared error of points from their respective cluster means, and it performs hard clustering, that is, each point is assigned to only one cluster. We also show how kernel K-means can be used for nonlinear clusters. EM generalizes K-means by modeling the data as a mixture of normal distributions, and it finds the cluster parameters (the mean and covariance matrix) by maximizing the likelihood of the data. It is a soft clustering approach, that is, instead of making a hard assignment, it returns the probability that a point belongs to each cluster. In Chapter 14 we consider various agglomerative hierarchical clustering methods, which start from each point in its own cluster, and successively merge (or agglomerate) pairs of clusters until the desired number of clusters have been found. We consider various cluster proximity measures that distinguish the different hierarchical methods. There are some datasets where the points from different clusters may in fact be closer in distance than points from the same cluster; this usually happens when the clusters are nonconvex in shape. Density-based clustering methods described in Chapter 15 use the density or connectedness properties to find such nonconvex clusters. The two main methods are DBSCAN and its generalization DENCLUE, which is based on kernel density estimation. We consider graph clustering methods in Chapter 16, which are typically based on spectral analysis of graph data. Graph clustering can be considered as an optimization problem over a k-way cut in a graph; different objectives can be cast as spectral decomposition of different graph matrices, such as the (normalized) adjacency matrix, Laplacian matrix, and so on, derived from the original graph data or from the kernel matrix. Finally, given the proliferation of different types of clustering methods,

it is important to assess the mined clusters as to how good they are in capturing the natural groups in data. In Chapter 17, we describe various clustering validation and evaluation strategies, spanning external and internal measures to compare a clustering with the ground-truth if it is available, or to compare two clusterings. We also highlight methods for clustering stability, that is, the sensitivity of the clustering to data perturbation, and clustering tendency, that is, the clusterability of the data. We also consider methods to choose the parameter k, which is the user-specified value for the number of clusters to discover.

1.5.4 Classification

The classification task is to predict the label or class for a given unlabeled point. Formally, a classifier is a model or function M that predicts the class label $\hat{y}$ for a given input example $\mathbf{x}$, that is, $\hat{y} = M(\mathbf{x})$, where $\hat{y} \in \{c_1, c_2, \ldots, c_k\}$ and each c_i is a class label (a categorical attribute value). To build the model we require a set of points with their correct class labels, which is called a *training set*. After learning the model M, we can automatically predict the class for any new point. Many different types of classification models have been proposed such as decision trees, probabilistic classifiers, support vector machines, and so on.

Part IV starts with the powerful Bayes classifier, which is an example of the probabilistic classification approach (Chapter 18). It uses the Bayes theorem to predict the class as the one that maximizes the posterior probability $P(c_i|\mathbf{x})$. The main task is to estimate the joint probability density function $f(\mathbf{x})$ for each class, which is modeled via a multivariate normal distribution. One limitation of the Bayes approach is the number of parameters to be estimated which scales as $O(d^2)$. The naive Bayes classifier makes the simplifying assumption that all attributes are independent, which requires the estimation of only $O(d)$ parameters. It is, however, surprisingly effective for many datasets. In Chapter 19 we consider the popular decision tree classifier, one of whose strengths is that it yields models that are easier to understand compared to other methods. A decision tree recursively partitions the data space into "pure" regions that contain data points from only one class, with relatively few exceptions. Next, in Chapter 20, we consider the task of finding an optimal direction that separates the points from two classes via linear discriminant analysis. It can be considered as a dimensionality reduction method that also takes the class labels into account, unlike PCA, which does not consider the class attribute. We also describe the generalization of linear to kernel discriminant analysis, which allows us to find nonlinear directions via the kernel trick. In Chapter 21 we describe the support vector machine (SVM) approach in detail, which is one of the most effective classifiers for many different problem domains. The goal of SVMs is to find the optimal hyperplane that maximizes the *margin* between the classes. Via the kernel trick, SVMs can be used to find nonlinear boundaries, which nevertheless correspond to some linear hyperplane in some high-dimensional "nonlinear" space. One of the important tasks in classification is to assess how good the models are. We conclude this part with Chapter 22, which presents the various methodologies for assessing classification models. We define various classification performance measures including ROC analysis. We then describe the bootstrap and cross-validation approaches for classifier evaluation. Finally, we

discuss the bias–variance tradeoff in classification, and how ensemble classifiers can help improve the variance or the bias of a classifier.

1.6 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); 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*. Adaptative 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. NewYork: 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: Practical Machine Learning Tools and Techniques*, 3rd ed. The Morgan Kaufmann Series in Data Management Systems. Philadelphia: Elsevier Science.

1.7 EXERCISES

Q1. Show that the mean of the centered data matrix $\mathbf{Z}$ in (1.5) is $\mathbf{0}$.

Q2. Prove that for the L_p-distance in Eq. (1.2), we have

$$\delta_\infty(\mathbf{x}, \mathbf{y}) = \lim_{p \to \infty} \delta_p(\mathbf{x}, \mathbf{y}) = \max_{i=1}^{d} \{|x_i - y_i|\}$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

PART ONE

DATA ANALYSIS FOUNDATIONS

CHAPTER 2 Numeric Attributes

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

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

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

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

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:

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

That is, the inverse CDF gives the least value of X, for which q fraction of the values are higher, and $1 - q$ fraction of the values are lower. 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

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

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 given 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

$$\mu = E[X] = \sum_{x} x f(x) \tag{2.4}$$

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

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

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

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

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

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.5}$$

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

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

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

$$E[\hat{\mu}] = E\left[\frac{1}{n} \sum_{i=1}^{n} x_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[x_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu \tag{2.6}$$

where we use the fact that the random variables x_i are IID according to X, which implies that they have the same mean μ 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.

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:

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

Figure 2.1 shows all 150 values of `sepal length`, and the sample mean. Figure 2.2a shows the empirical CDF and Figure 2.2b 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 2.2b, we can see that

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

The sample mode for `sepal length` is 5, which can be observed from the frequency of 5 in Figure 2.1. The empirical probability mass at $x = 5$ is

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

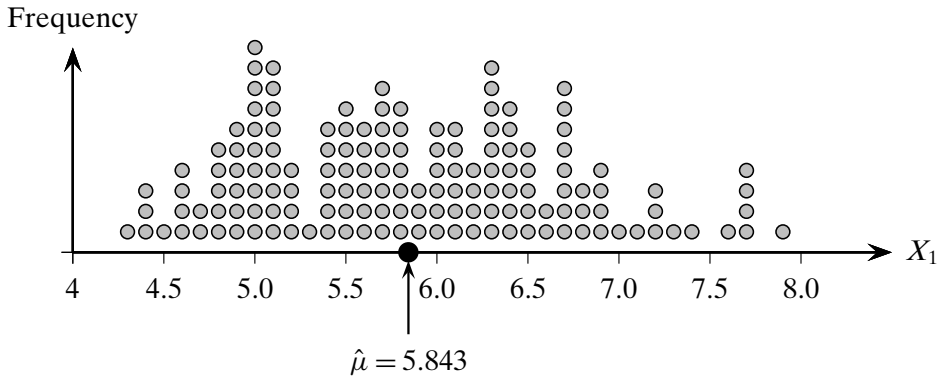
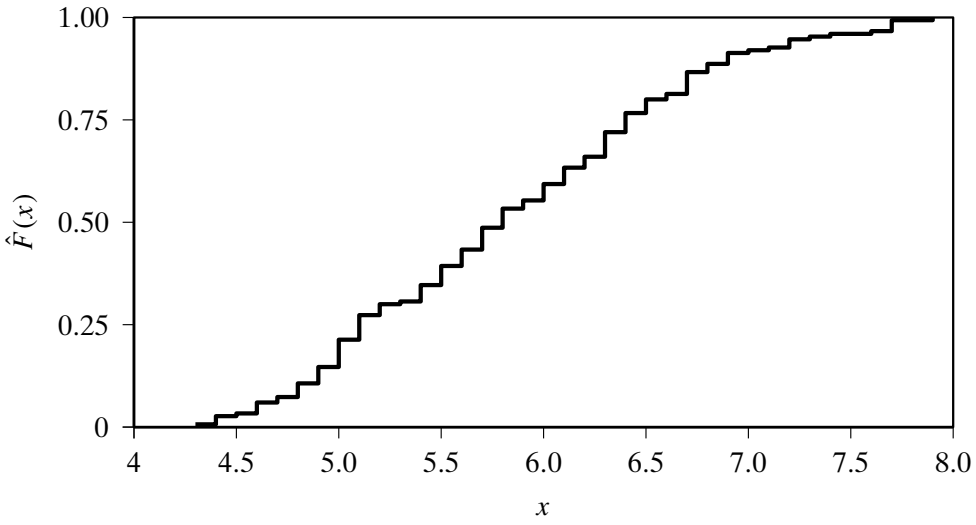
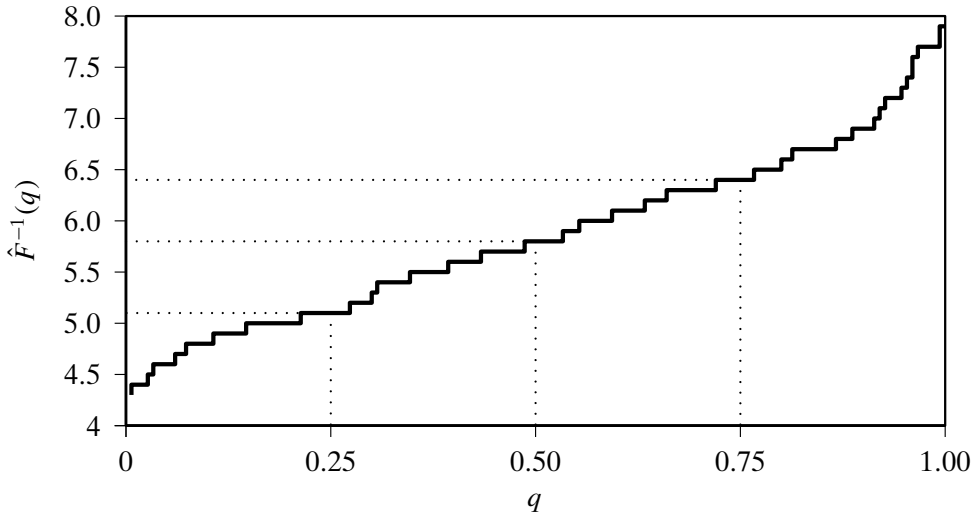


Figure 2.1. Sample mean for `sepal length`. Multiple occurrences of the same value are shown stacked.



(a) Empirical CDF



(b) Empirical inverse CDF

Figure 2.2. Empirical CDF and inverse CDF: `sepal length`.

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

$$IQR = q_3 - q_1 = F^{-1}(0.75) - F^{-1}(0.25) \tag{2.7}$$

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. (2.7):

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

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

$$\sigma^2 = var(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 \, f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 \, f(x)\, dx & \text{if } X \text{ is continuous} \end{cases} \tag{2.8}$$

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

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

$$\begin{aligned} \sigma^2 = var(X) &= E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned} \tag{2.9}$$

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 *rth moment about the mean* for a random variable X, defined as $E[(\mathbf{x} - \mu)^r]$.

Sample Variance The *sample variance* is defined as

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{2.10}$$

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. (2.8), 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:

$$z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

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

Geometric Interpretation of Sample Variance We can treat the data sample for attribute X as a vector in n-dimensional space, where n is the sample size. That is, we write $X = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$. Further, let

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

denote the mean subtracted attribute vector, where $\mathbf{1} \in \mathbb{R}^n$ is the n-dimensional vector all of whose elements have value 1. We can rewrite Eq. (2.10) in terms of the magnitude of Z, that is, the dot product of Z with itself:

$$\hat{\sigma}^2 = \frac{1}{n} \|Z\|^2 = \frac{1}{n} Z^T Z = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{2.11}$$

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.

Example 2.2. Consider the data sample for `sepal length` shown in Figure 2.1. 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 2.2b, 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 can be computed from the centered data vector via Eq. (2.11):

$$\hat{\sigma}^2 = \frac{1}{n}(X - \mathbf{1} \cdot \hat{\mu})^T (X - \mathbf{1} \cdot \hat{\mu}) = 102.168/150 = 0.681$$

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 μ, as we saw in Eq. (2.6). 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

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

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

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

Combining the above two facts, we get

$$var\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} var(x_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2 \tag{2.12}$$

Further, note that

$$E\left[\sum_{i=1}^{n} x_i\right] = n\mu \tag{2.13}$$

Using Eqs. (2.9), (2.12), and (2.13), the variance of the sample mean $\hat{\mu}$ can be computed as

$$\begin{aligned} var(\hat{\mu}) = E[(\hat{\mu} - \mu)^2] = E[\hat{\mu}^2] - \mu^2 &= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \frac{1}{n^2} E\left[\sum_{i=1}^{n} x_i\right]^2 \\ &= \frac{1}{n^2}\left(E\left[\left(\sum_{i=1}^{n} x_i\right)^2\right] - E\left[\sum_{i=1}^{n} x_i\right]^2\right) = \frac{1}{n^2}\, var\left(\sum_{i=1}^{n} x_i\right) \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{2.14}$$

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

Sample Variance Is Biased, but Is Asymptotically Unbiased The sample variance in Eq. (2.10) is a *biased estimator* for the true population variance, σ^2, that is, $E[\hat{\sigma}^2] \neq \sigma^2$. To show this we make use of the identity

$$\sum_{i=1}^{n} (x_i - \mu)^2 = n(\hat{\mu} - \mu)^2 + \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{2.15}$$

Computing the expectation of $\hat{\sigma}^2$ by using Eq. (2.15) in the first step, we get

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2\right] - E[(\hat{\mu} - \mu)^2] \tag{2.16}$$

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

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

Further, from Eq. (2.14) the sample mean $\hat{\mu}$ has variance $E[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}$. Plugging these into the Eq. (2.16) 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 σ^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 $n \to \infty$ because

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

Put differently, as the sample size increases, we have

$$E[\hat{\sigma}^2] \to \sigma^2 \qquad \text{as } n \to \infty$$

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:

$$X_1 = (x_{11}, x_{21}, \ldots, x_{n1})^T$$
$$X_2 = (x_{12}, x_{22}, \ldots, x_{n2})^T$$

In the probabilistic view, the column vector $\mathbf{X} = (X_1, X_2)^T$ is considered a bivariate vector random variable, and the points $\mathbf{x}_i$ $(1 \le i \le n)$ 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

$$\hat{f}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x}) \tag{2.17}$$

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

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:

$$\boldsymbol{\mu} = E[\mathbf{X}] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{2.18}$$

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. (2.5). It can also be computed from the joint empirical PMF in Eq. (2.17)

$$\hat{\boldsymbol{\mu}} = \sum_{\mathbf{x}} \mathbf{x} \hat{f}(\mathbf{x}) = \sum_{\mathbf{x}} \mathbf{x} \left(\frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x}) \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \tag{2.19}$$

Variance

We can compute the variance along each attribute, namely σ_1^2 for X_1 and σ_2^2 for X_2 using Eq. (2.8). The *total variance* [Eq. (1.4)] is given as

$$var(\mathbf{D}) = \sigma_1^2 + \sigma_2^2$$

The sample variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ can be estimated using Eq. (2.10), and the *sample total variance* is simply $\hat{\sigma}_1^2 + \hat{\sigma}_2^2$.

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

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] \tag{2.20}$$

By linearity of expectation, we have

$$\begin{aligned}\sigma_{12} &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E[X_1X_2 - X_1\mu_2 - X_2\mu_1 + \mu_1\mu_2] \\ &= E[X_1X_2] - \mu_2 E[X_1] - \mu_1 E[X_2] + \mu_1\mu_2 \\ &= E[X_1X_2] - \mu_1\mu_2 \\ &= E[X_1X_2] - E[X_1]E[X_2] \end{aligned} \tag{2.21}$$

Eq. (2.21) can be seen as a generalization of the univariate variance [Eq. (2.9)] 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

$$E[X_1X_2] = E[X_1] \cdot E[X_2]$$

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

$$\hat{\sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2) \tag{2.22}$$

It can be derived by substituting the empirical joint probability mass function $\hat{f}(x_1, x_2)$ from Eq. (2.17) into Eq. (2.20), as follows:

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

Notice that sample covariance is a generalization of the sample variance [Eq. (2.10)] because

$$\hat{\sigma}_{11} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_1)(x_i - \mu_1) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_1)^2 = \hat{\sigma}_1^2$$

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

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}} \tag{2.23}$$

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

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n}(x_{i2} - \hat{\mu}_2)^2}} \tag{2.24}$$

Geometric Interpretation of Sample Covariance and Correlation

Let Z_1 and Z_2 denote the centered attribute vectors in $\mathbb{R}^n$, given as follows:

$$Z_1 = X_1 - \mathbf{1} \cdot \hat{\mu}_1 = \begin{pmatrix} x_{11} - \hat{\mu}_1 \\ x_{21} - \hat{\mu}_1 \\ \vdots \\ x_{n1} - \hat{\mu}_1 \end{pmatrix} \qquad Z_2 = X_2 - \mathbf{1} \cdot \hat{\mu}_2 = \begin{pmatrix} x_{12} - \hat{\mu}_2 \\ x_{22} - \hat{\mu}_2 \\ \vdots \\ x_{n2} - \hat{\mu}_2 \end{pmatrix}$$

The sample covariance [Eq. (2.22)] can then be written as

$$\hat{\sigma}_{12} = \frac{Z_1^T Z_2}{n}$$

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. (2.11).

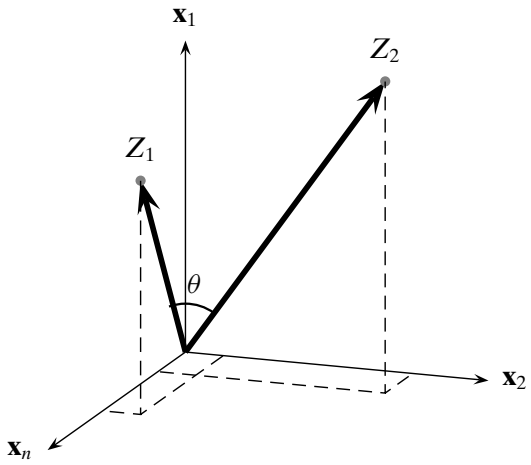


Figure 2.3. Geometric interpretation of covariance and correlation. The two centered attribute vectors are shown in the (conceptual) n-dimensional space $\mathbb{R}^n$ spanned by the n points.

The sample correlation [Eq. (2.24)] can be written as

$$\hat{\rho}_{12} = \frac{Z_1^T Z_2}{\sqrt{Z_1^T Z_1}\sqrt{Z_2^T Z_2}} = \frac{Z_1^T Z_2}{\|Z_1\| \, \|Z_2\|} = \left(\frac{Z_1}{\|Z_1\|}\right)^T \left(\frac{Z_2}{\|Z_2\|}\right) = \cos\theta \tag{2.25}$$

Thus, the correlation coefficient is simply the cosine of the angle [Eq. (1.3)] between the two centered attribute vectors, as illustrated in Figure 2.3.

Covariance Matrix

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

$$\begin{aligned}
\boldsymbol{\Sigma} &= E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \\
&= E\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} \begin{pmatrix} X_1 - \mu_1 & X_2 - \mu_2 \end{pmatrix}\right] \\
&= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}
\end{aligned} \tag{2.26}$$

Because $\sigma_{12} = \sigma_{21}$, $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$, which is also called the *trace* of $\boldsymbol{\Sigma}$, given as

$$var(\mathbf{D}) = tr(\boldsymbol{\Sigma}) = \sigma_1^2 + \sigma_2^2$$

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

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 $\boldsymbol{\Sigma}$, denoted as $|\boldsymbol{\Sigma}|$ or $\det(\boldsymbol{\Sigma})$. The generalized covariance is non-negative, because

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

where we used Eq. (2.23), 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 $\det(\boldsymbol{\Sigma}) \geq 0$, that is, the determinant is non-negative.

The *sample covariance matrix* is given as

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

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

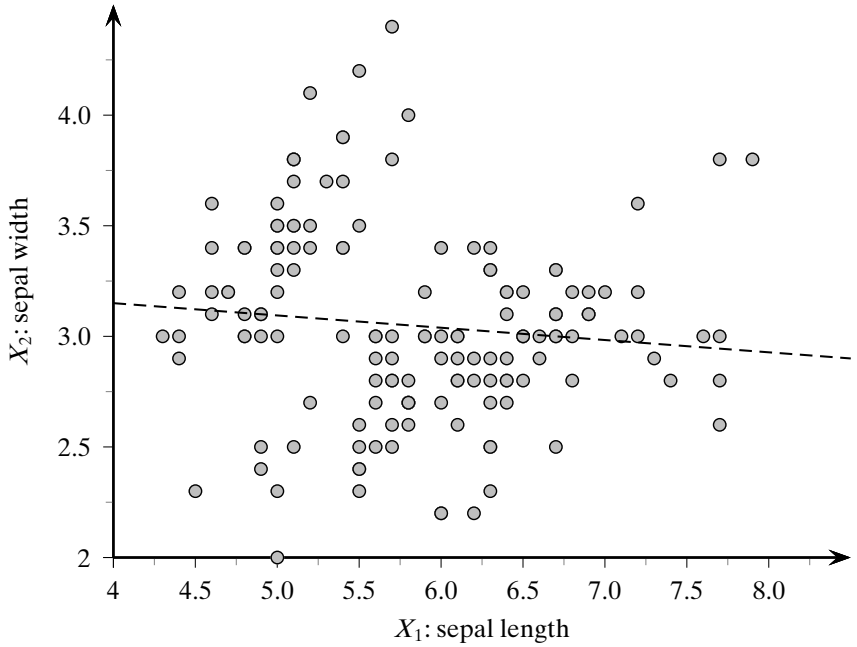


Figure 2.4. Correlation between `sepal length` and `sepal width`.

Example 2.3 (Sample Mean and Covariance). Consider the `sepal length` and `sepal width` attributes for the Iris dataset, plotted in Figure 2.4. There are $n = 150$ points in the $d = 2$ dimensional attribute space. The sample mean vector is given as

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

The sample covariance matrix is given as

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

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 2.4. 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°, that is, the two attribute vectors are almost orthogonal, indicating weak correlation. Further, the angle being greater than 90° indicates negative correlation.

The sample total variance is given as

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

and the sample generalized variance is given as

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

2.3 MULTIVARIATE ANALYSIS

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

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ \hline x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

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

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

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

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

In the probabilistic view, the d attributes are modeled as a vector random variable, $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$, 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. (2.18), 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. (2.19), the *sample mean* is given as

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

Covariance Matrix

Generalizing Eq. (2.26) 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:

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

The diagonal element σ_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 $\mathbf{\Sigma}$ is a *positive semidefinite* matrix, that is,

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

To see this, observe that

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

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

$$\det(\mathbf{\Sigma}) = \prod_{i=1}^{d} \lambda_i \geq 0 \tag{2.27}$$

Total and Generalized Variance

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

$$var(\mathbf{D}) = tr(\mathbf{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2 \tag{2.28}$$

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

The generalized variance is defined as the determinant of the covariance matrix, $\det(\mathbf{\Sigma})$, also denoted as $|\mathbf{\Sigma}|$. It gives a single value for the overall multivariate scatter. From Eq. (2.27) we have $\det(\mathbf{\Sigma}) \geq 0$.

Sample Covariance Matrix

The *sample covariance matrix* is given as

$$\widehat{\boldsymbol{\Sigma}} = E[(\mathbf{X} - \hat{\boldsymbol{\mu}})(\mathbf{X} - \hat{\boldsymbol{\mu}})^T] = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1d} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \cdots & \hat{\sigma}_d^2 \end{pmatrix} \tag{2.29}$$

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

$$\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \hat{\boldsymbol{\mu}}^T = \begin{pmatrix} | & | & & | \\ Z_1 & Z_2 & \cdots & Z_d \\ | & | & & | \end{pmatrix}$$

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

$$\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \hat{\boldsymbol{\mu}}^T = \begin{pmatrix} \mathbf{x}_1^T - \hat{\boldsymbol{\mu}}^T \\ \mathbf{x}_2^T - \hat{\boldsymbol{\mu}}^T \\ \vdots \\ \mathbf{x}_n^T - \hat{\boldsymbol{\mu}}^T \end{pmatrix} = \begin{pmatrix} — & \mathbf{z}_1^T & — \\ — & \mathbf{z}_2^T & — \\ & \vdots & \\ — & \mathbf{z}_n^T & — \end{pmatrix}$$

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

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\left(\mathbf{Z}^T\mathbf{Z}\right) = \frac{1}{n}\begin{pmatrix} Z_1^TZ_1 & Z_1^TZ_2 & \cdots & Z_1^TZ_d \\ Z_2^TZ_1 & Z_2^TZ_2 & \cdots & Z_2^TZ_d \\ \vdots & \vdots & \ddots & \vdots \\ Z_d^TZ_1 & Z_d^TZ_2 & \cdots & Z_d^TZ_d \end{pmatrix} \tag{2.30}$$

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

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \cdot \mathbf{z}_i^T \tag{2.31}$$

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

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 5.843 & 3.054 & 3.759 & 1.199 \end{pmatrix}^T$$

and the sample covariance matrix is given as

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

The sample total variance is

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

and the generalized variance is

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

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

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

The mean vector is as follows:

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

and the centered data matrix is then given as

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

The inner-product approach [Eq. (2.30)] to compute the sample covariance matrix gives

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z} = \frac{1}{3}\begin{pmatrix} -4 & 0 & 4 \\ -2.1 & -0.5 & 2.6 \end{pmatrix} \cdot \begin{pmatrix} -4 & -2.1 \\ 0 & -0.5 \\ 4 & 2.6 \end{pmatrix}$$

$$= \frac{1}{3}\begin{pmatrix} 32 & 18.8 \\ 18.8 & 11.42 \end{pmatrix} = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix}$$

Alternatively, the outer-product approach [Eq. (2.31)] gives

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \cdot \mathbf{z}_i^T$$

$$= \frac{1}{3}\left[\begin{pmatrix} -4 \\ -2.1 \end{pmatrix} \cdot \begin{pmatrix} -4 & -2.1 \end{pmatrix} + \begin{pmatrix} 0 \\ -0.5 \end{pmatrix} \cdot \begin{pmatrix} 0 & -0.5 \end{pmatrix} + \begin{pmatrix} 4 \\ 2.6 \end{pmatrix} \cdot \begin{pmatrix} 4 & 2.6 \end{pmatrix} \right]$$

$$= \frac{1}{3}\left[\begin{pmatrix}16.0 & 8.4\\ 8.4 & 4.41\end{pmatrix} + \begin{pmatrix}0.0 & 0.0\\ 0.0 & 0.25\end{pmatrix} + \begin{pmatrix}16.0 & 10.4\\ 10.4 & 6.76\end{pmatrix}\right]$$

$$= \frac{1}{3}\begin{pmatrix}32.0 & 18.8\\ 18.8 & 11.42\end{pmatrix} = \begin{pmatrix}10.67 & 6.27\\ 6.27 & 3.81\end{pmatrix}$$

where the centered points $\mathbf{z}_i$ are the rows of $\mathbf{Z}$. 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 $x_1, x_2, \ldots, x_n$ 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 $\mathbf{x}_2$:

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

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

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

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

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

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

Likewise, the sample range for Income is $2700 - 300 = 2400$, with a minimum value of 300; Income is therefore transformed into

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

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

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

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

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

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

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

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

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

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 μ and variance σ^2, if the probability density function of X is given as follows:

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

The term $(x-\mu)^2$ measures the distance of a value x from the mean μ 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 σ of the distribution.

Example 2.7. Figure 2.5 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 μ is also the median, as well as the mode, of the distribution.

Probability Mass

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

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

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

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

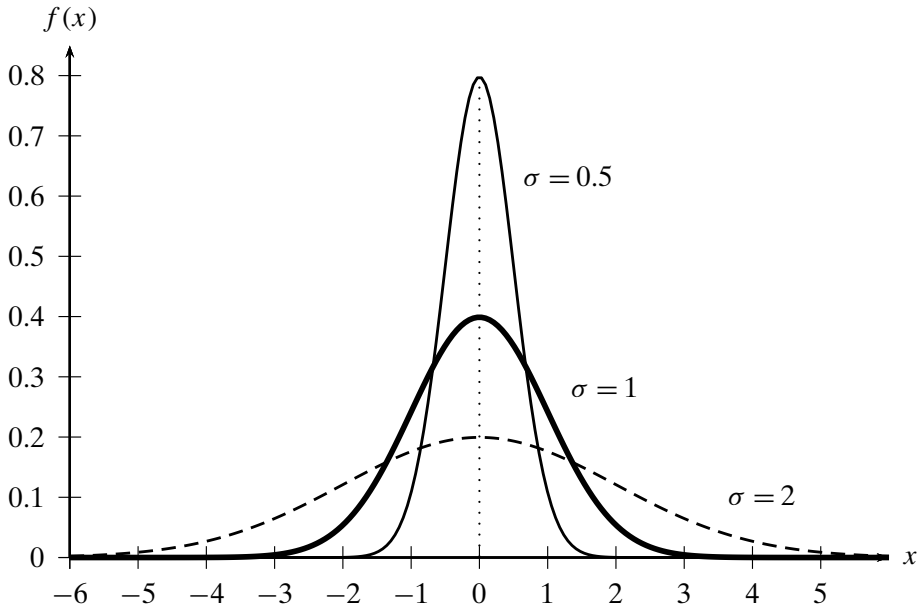


Figure 2.5. Normal distribution: $\mu = 0$, and different variances.

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

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

$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{k} \mathrm{e}^{-\frac{1}{2}z^2} dz$$

The last step follows from the fact that $\mathrm{e}^{-\frac{1}{2}z^2}$ 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

$$P(-k \le z \le k) = P\left(0 \le t \le k/\sqrt{2}\right) = \frac{2}{\sqrt{\pi}} \int_{0}^{k/\sqrt{2}} \mathrm{e}^{-t^2} dt = \mathrm{erf}\left(k/\sqrt{2}\right) \tag{2.32}$$

where erf is the *Gauss error function*, defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} \mathrm{e}^{-t^2} dt$$

Using Eq. (2.32) we can compute the probability mass within k standard deviations of the mean. In particular, for $k = 1$, we have

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

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 μ.

2.5.2 Multivariate Normal Distribution

Given the d-dimensional vector random variable $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$, 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:

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}\right\} \tag{2.33}$$

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

$$(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \tag{2.34}$$

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

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

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 $\mu = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$. Figure 2.6a 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

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

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 2.6b. 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 $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. 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 $\boldsymbol{\mu}$ is not necessarily the origin $\mathbf{0}$. The

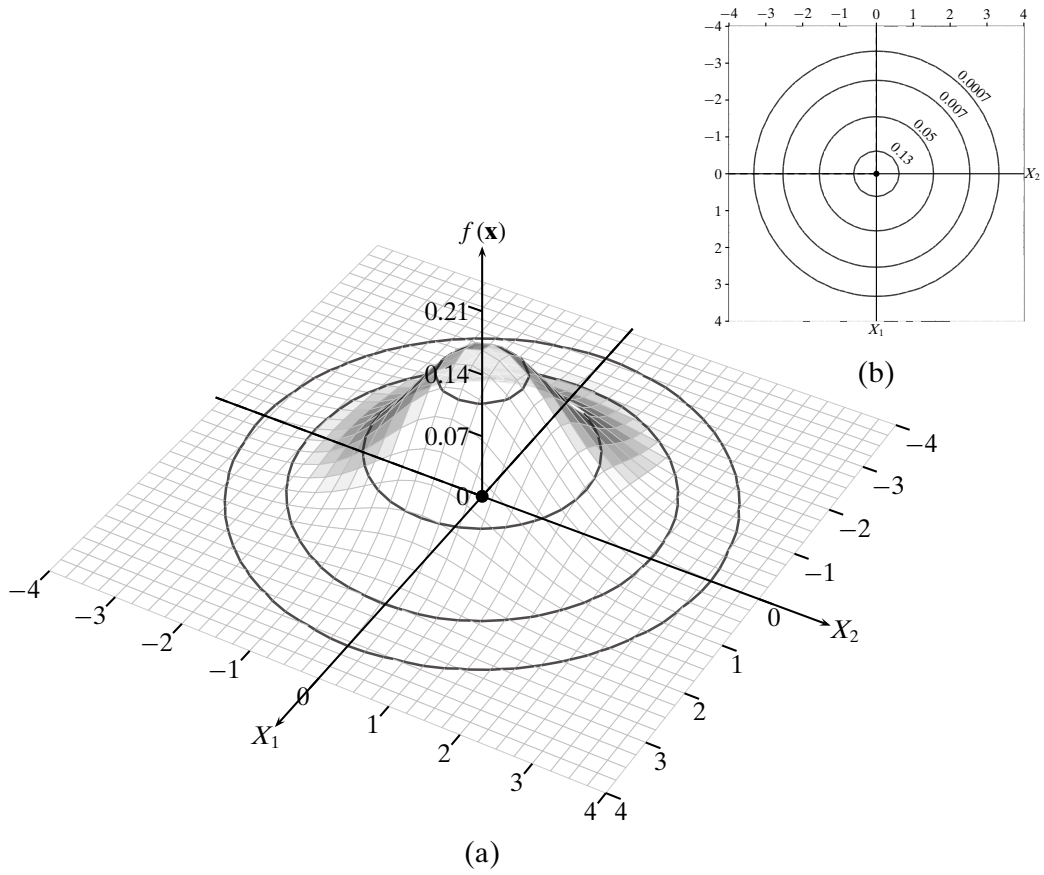


Figure 2.6. (a) Standard bivariate normal density and (b) its contour plot. Parameters: $\boldsymbol{\mu} = (0, 0)^T$, $\boldsymbol{\Sigma} = \mathbf{I}$.

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 $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric positive semidefinite matrix. The eigenvector equation for $\boldsymbol{\Sigma}$ is given as

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

Here λ_i is an eigenvalue of $\boldsymbol{\Sigma}$ and the vector $\mathbf{u}_i \in \mathbb{R}^d$ is the eigenvector corresponding to λ_i. Because $\boldsymbol{\Sigma}$ 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: $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$. The diagonal matrix $\boldsymbol{\Lambda}$ is used to record these eigenvalues:

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

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

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

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 $\mathbf{\Sigma}$ can then be expressed compactly as follows:

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

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 $\mathbf{u}_i$. $\mathbf{\Sigma}$ is the covariance matrix in the original space, whereas $\mathbf{\Lambda}$ is the covariance matrix in the new coordinate space. Because $\mathbf{\Lambda}$ is a diagonal matrix, we can immediately conclude that after the transformation, each new dimension $\mathbf{u}_i$ has variance λ_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 λ_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 is given as $\det(\mathbf{\Sigma}) = \prod_{i=1}^{d} \lambda_i$. Thus, the generalized variance of $\mathbf{\Sigma}$ is the product of its eigenvectors.

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 $var(\mathbf{D})$ for a dataset $\mathbf{D}$ is invariant, that is,

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

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

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

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

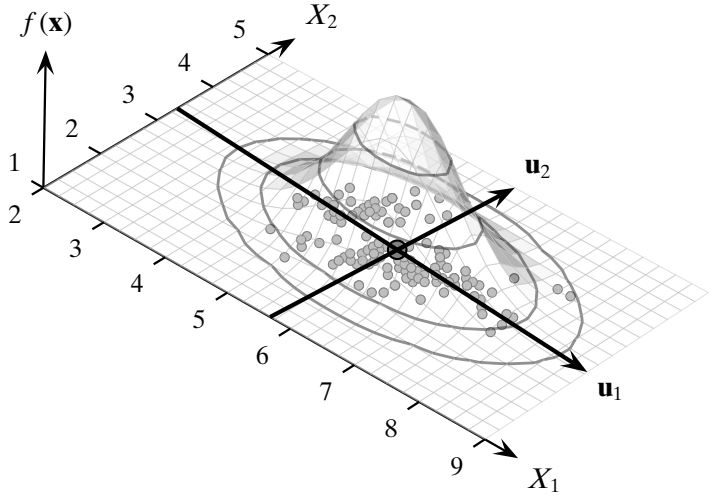


Figure 2.7. Iris: `sepal length` and `sepal width`, bivariate normal density and contours.

and the sample covariance matrix is given as

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

The plot of the bivariate normal density for the two attributes is shown in Figure 2.7. The figure also shows the contour lines and the data points.

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

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

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

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

whereas the squared Euclidean distance between them is

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

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

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

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

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

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

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

Figure 2.7 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° (or -4.5°).

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); 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 $\hat{\boldsymbol{\mu}}$ and the covariance matrix $\widehat{\boldsymbol{\Sigma}}$ 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. (2.15) 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

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

This fact was used in Eq. (2.12).

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. (2.18).

Q7. Show that the correlation [Eq. (2.23)] 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
$\mathbf{x}_1$	17	17	12
$\mathbf{x}_2$	11	9	13
$\mathbf{x}_3$	11	8	19

Q9. Show that the outer-product in Eq. (2.31) for the sample covariance matrix is equivalent to Eq. (2.29).

Q10. 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.

Q11. 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 Q11

X	Y
1	a
0	b
1	c
0	a
0	c

Q12. Under what conditions will the covariance matrix $\mathbf{\Sigma}$ 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?

CHAPTER 3 Categorical Attributes

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 $dom(X) = \{a_1, a_2, \ldots, a_m\}$. The data $\mathbf{D}$ is thus an $n \times 1$ symbolic data matrix given as

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

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

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

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

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

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

and the variance of X is given as

$$\begin{aligned} \sigma^2 = var(X) &= E[X^2] - (E[X])^2 \\ &= (1^2 \cdot p + 0^2 \cdot (1-p)) - p^2 = p - p^2 = p(1-p) \end{aligned} \tag{3.1}$$

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 $\{x_1, x_2, \ldots, x_n\}$ 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

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{n_1}{n} = \hat{p} \tag{3.2}$$

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

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

The sample variance could also have been obtained directly from Eq. (3.1), 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 $\{x_1, x_2, \ldots, x_n\}$ 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

$$f(N = n_1 \,|\, n, p) = \binom{n}{n_1} p^{n_1} (1 - p)^{n - n_1} \tag{3.3}$$

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

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

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

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

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:

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

As the sample size increases, the binomial distribution given in Eq. 3.3 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 $\{a_1, a_2, \ldots, a_m\}$. We can model X as an m-dimensional Bernoulli random variable $\mathbf{X} = (A_1, A_2, \ldots, A_m)^T$, 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 ith standard basis vector $\mathbf{e}_i \in \mathbb{R}^m$ given as

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

In $\mathbf{e}_i$, only the ith 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

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

The range of $\mathbf{X}$ consists of m distinct vector values $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m\}$, 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

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

The PMF can be written compactly as follows:

$$P(\mathbf{X} = \mathbf{e}_i) = f(\mathbf{e}_i) = \prod_{j=1}^{m} p_j^{e_{ij}} \tag{3.4}$$

Because $e_{ii} = 1$, and $e_{ij} = 0$ for $j \neq i$, we can see that, as expected, we have

$$f(\mathbf{e}_i) = \prod_{j=1}^{m} p_j^{e_{ij}} = p_1^{e_{i0}} \times \cdots p_i^{e_{ii}} \cdots \times p_m^{e_{im}} = p_1^0 \times \cdots p_i^1 \cdots \times p_m^0 = p_i$$

Table 3.1. Discretized `sepal length` attribute

Bins	Domain	Counts
[4.3, 5.2]	Very Short (a_1)	$n_1 = 45$
(5.2, 6.1]	Short (a_2)	$n_2 = 50$
(6.1, 7.0]	Long (a_3)	$n_3 = 43$
(7.0, 7.9]	Very Long (a_4)	$n_4 = 12$

Example 3.3. Let us consider the `sepal length` attribute (X_1) for the Iris dataset shown in Table 1.2. We divide the sepal length into four equal-width intervals, and give each interval a name as shown in Table 3.1. We consider X_1 as a categorical attribute with domain

$$\{a_1 = \texttt{VeryShort}, a_2 = \texttt{Short}, a_3 = \texttt{Long}, a_4 = \texttt{VeryLong}\}$$

We model the categorical attribute X_1 as a multivariate Bernoulli variable $\mathbf{X}$, defined as

$$\mathbf{X}(v) = \begin{cases} \mathbf{e}_1 = (1,0,0,0) & \text{if } v = a_1 \\ \mathbf{e}_2 = (0,1,0,0) & \text{if } v = a_2 \\ \mathbf{e}_3 = (0,0,1,0) & \text{if } v = a_3 \\ \mathbf{e}_4 = (0,0,0,1) & \text{if } v = a_4 \end{cases}$$

For example, the symbolic point $x_1 = \texttt{Short} = a_2$ is represented as the vector $(0,1,0,0)^T = \mathbf{e}_2$.

Mean

The mean or expected value of $\mathbf{X}$ can be obtained as

$$\boldsymbol{\mu} = E[\mathbf{X}] = \sum_{i=1}^{m} \mathbf{e}_i f(\mathbf{e}_i) = \sum_{i=1}^{m} \mathbf{e}_i p_i = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} p_1 + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} p_m = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix} = \mathbf{p} \tag{3.5}$$

Sample Mean

Assume that each symbolic point $x_i \in \mathbf{D}$ is mapped to the variable $\mathbf{x}_i = \mathbf{X}(x_i)$. The mapped dataset $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is then assumed to be a random sample IID with $\mathbf{X}$. We can compute the sample mean by placing a probability mass of $\frac{1}{n}$ at each point

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \sum_{i=1}^{m} \frac{n_i}{n} \mathbf{e}_i = \begin{pmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_m/n \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_m \end{pmatrix} = \hat{\mathbf{p}} \tag{3.6}$$

where n_i is the number of occurrences of the vector value $\mathbf{e}_i$ in the sample, which is equivalent to the number of occurrences of the symbol a_i. Furthermore, we have

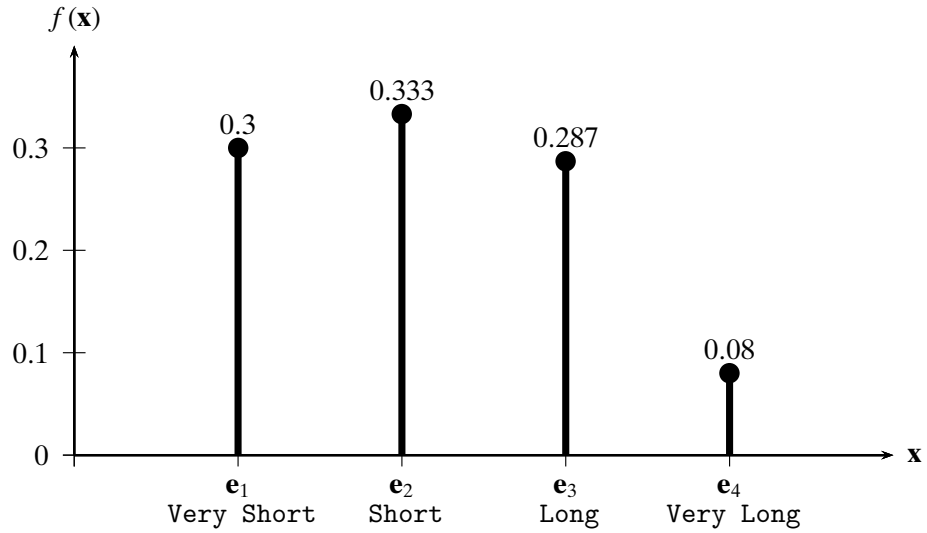


Figure 3.1. Probability mass function: `sepal length`.

$\sum_{i=1}^{m} n_i = n$, which follows from the fact that $\mathbf{X}$ can take on only m distinct values $\mathbf{e}_i$, and the counts for each value must add up to the sample size n.

Example 3.4 (Sample Mean). Consider the observed counts n_i for each of the values a_i ($\mathbf{e}_i$) of the discretized `sepal length` attribute, shown in Table 3.1. Because the total sample size is $n = 150$, from these we can obtain the estimates $\hat{p}_i$ as follows:

$$\hat{p}_1 = 45/150 = 0.3$$
$$\hat{p}_2 = 50/150 = 0.333$$
$$\hat{p}_3 = 43/150 = 0.287$$
$$\hat{p}_4 = 12/150 = 0.08$$

The PMF for $\mathbf{X}$ is plotted in Figure 3.1, and the sample mean for $\mathbf{X}$ is given as

$$\hat{\boldsymbol{\mu}} = \hat{\mathbf{p}} = \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix}$$

Covariance Matrix

Recall that an m-dimensional multivariate Bernoulli variable is simply a vector of m Bernoulli variables. For instance, $\mathbf{X} = (A_1, A_2, \ldots, A_m)^T$, where A_i is the Bernoulli variable corresponding to symbol a_i. The variance–covariance information between the constituent Bernoulli variables yields a covariance matrix for $\mathbf{X}$.

Let us first consider the variance along each Bernoulli variable A_i. By Eq. (3.1), we immediately have

$$\sigma_i^2 = var(A_i) = p_i(1 - p_i)$$

Next consider the covariance between A_i and A_j. Utilizing the identity in Eq. (2.21), we have

$$\sigma_{ij} = E[A_i A_j] - E[A_i] \cdot E[A_j] = 0 - p_i p_j = -p_i p_j$$

which follows from the fact that $E[A_i A_j] = 0$, as A_i and A_j cannot both be 1 at the same time, and thus their product $A_i A_j = 0$. This same fact leads to the negative relationship between A_i and A_j. What is interesting is that the degree of negative association is proportional to the product of the mean values for A_i and A_j.

From the preceding expressions for variance and covariance, the $m \times m$ covariance matrix for $\mathbf{X}$ is given as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \ldots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_m \\ -p_1 p_2 & p_2(1-p_2) & \cdots & -p_2 p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_1 p_m & -p_2 p_m & \cdots & p_m(1-p_m) \end{pmatrix}$$

Notice how each row in $\boldsymbol{\Sigma}$ sums to zero. For example, for row i, we have

$$-p_i p_1 - p_i p_2 - \cdots + p_i(1 - p_i) - \cdots - p_i p_m = p_i - p_i \sum_{j=1}^{m} p_j = p_i - p_i = 0 \tag{3.7}$$

Because $\boldsymbol{\Sigma}$ is symmetric, it follows that each column also sums to zero.

Define $\mathbf{P}$ as the $m \times m$ diagonal matrix:

$$\mathbf{P} = \text{diag}(\mathbf{p}) = \text{diag}(p_1, p_2, \ldots, p_m) = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_m \end{pmatrix}$$

We can compactly write the covariance matrix of $\mathbf{X}$ as

$$\boldsymbol{\Sigma} = \mathbf{P} - \mathbf{p} \cdot \mathbf{p}^T \tag{3.8}$$

Sample Covariance Matrix

The sample covariance matrix can be obtained from Eq. (3.8) in a straightforward manner:

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{P}} - \hat{\mathbf{p}} \cdot \hat{\mathbf{p}}^T \tag{3.9}$$

where $\widehat{\mathbf{P}} = \text{diag}(\hat{\mathbf{p}})$, and $\hat{\mathbf{p}} = \hat{\boldsymbol{\mu}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)^T$ denotes the empirical probability mass function for $\mathbf{X}$.

Example 3.5. Returning to the discretized `sepal length` attribute in Example 3.4, we have $\hat{\boldsymbol{\mu}} = \hat{\mathbf{p}} = (0.3, 0.333, 0.287, 0.08)^T$. The sample covariance matrix is given as

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}} &= \widehat{\mathbf{P}} - \hat{\mathbf{p}} \cdot \hat{\mathbf{p}}^T \\
&= \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.333 & 0 & 0 \\ 0 & 0 & 0.287 & 0 \\ 0 & 0 & 0 & 0.08 \end{pmatrix} - \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} \begin{pmatrix} 0.3 & 0.333 & 0.287 & 0.08 \end{pmatrix} \\
&= \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.333 & 0 & 0 \\ 0 & 0 & 0.287 & 0 \\ 0 & 0 & 0 & 0.08 \end{pmatrix} - \begin{pmatrix} 0.09 & 0.1 & 0.086 & 0.024 \\ 0.1 & 0.111 & 0.096 & 0.027 \\ 0.086 & 0.096 & 0.082 & 0.023 \\ 0.024 & 0.027 & 0.023 & 0.006 \end{pmatrix} \\
&= \begin{pmatrix} 0.21 & -0.1 & -0.086 & -0.024 \\ -0.1 & 0.222 & -0.096 & -0.027 \\ -0.086 & -0.096 & 0.204 & -0.023 \\ -0.024 & -0.027 & -0.023 & 0.074 \end{pmatrix}
\end{aligned}$$

One can verify that each row (and column) in $\widehat{\boldsymbol{\Sigma}}$ sums to zero.

It is worth emphasizing that whereas the modeling of categorical attribute X as a multivariate Bernoulli variable, $\mathbf{X} = (A_1, A_2, \ldots, A_m)^T$, makes the structure of the mean and covariance matrix explicit, the same results would be obtained if we simply treat the mapped values $\mathbf{X}(x_i)$ as a new $n \times m$ binary data matrix, and apply the standard definitions of the mean and covariance matrix from multivariate numeric attribute analysis (see Section 2.3). In essence, the mapping from symbols a_i to binary vectors $\mathbf{e}_i$ is the key idea in categorical attribute analysis.

Example 3.6. Consider the sample $\mathbf{D}$ of size $n = 5$ for the `sepal length` attribute X_1 in the Iris dataset, shown in Table 3.2a. As in Example 3.1, we assume that X_1 has only two categorical values {`Long`, `Short`}. We model X_1 as the multivariate Bernoulli variable $\mathbf{X}_1$ defined as

$$\mathbf{X}_1(v) = \begin{cases} \mathbf{e}_1 = (1,0)^T & \text{if } v = \texttt{Long}(a_1) \\ \mathbf{e}_2 = (0,1)^T & \text{if } v = \texttt{Short}(a_2) \end{cases}$$

The sample mean [Eq. (3.6)] is

$$\hat{\boldsymbol{\mu}} = \hat{\mathbf{p}} = (2/5, 3/5)^T = (0.4, 0.6)^T$$

and the sample covariance matrix [Eq. (3.9)] is

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{P}} - \hat{\mathbf{p}}\hat{\mathbf{p}}^T &= \begin{pmatrix} 0.4 & 0 \\ 0 & 0.6 \end{pmatrix} - \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} \begin{pmatrix} 0.4 & 0.6 \end{pmatrix} \\
&= \begin{pmatrix} 0.4 & 0 \\ 0 & 0.6 \end{pmatrix} - \begin{pmatrix} 0.16 & 0.24 \\ 0.24 & 0.36 \end{pmatrix} = \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}
\end{aligned}$$

Table 3.2. (a) Categorical dataset. (b) Mapped binary dataset. (c) Centered dataset.

(a)

	X
x_1	Short
x_2	Short
x_3	Long
x_4	Short
x_5	Long

(b)

	A_1	A_2
$\mathbf{x}_1$	0	1
$\mathbf{x}_2$	0	1
$\mathbf{x}_3$	1	0
$\mathbf{x}_4$	0	1
$\mathbf{x}_5$	1	0

(c)

	Z_1	Z_2
$\mathbf{z}_1$	−0.4	0.4
$\mathbf{z}_2$	−0.4	0.4
$\mathbf{z}_3$	0.6	−0.6
$\mathbf{z}_4$	−0.4	0.4
$\mathbf{z}_5$	0.6	−0.6

To show that the same result would be obtained via standard numeric analysis, we map the categorical attribute X to the two Bernoulli attributes A_1 and A_2 corresponding to symbols Long and Short, respectively. The mapped dataset is shown in Table 3.2b. The sample mean is simply

$$\hat{\boldsymbol{\mu}} = \frac{1}{5}\sum_{i=1}^{5}\mathbf{x}_i = \frac{1}{5}(2,3)^T = (0.4, 0.6)^T$$

Next, we center the dataset by subtracting the mean value from each attribute. After centering, the mapped dataset is as shown in Table 3.2c, with attribute Z_i as the centered attribute A_i. We can compute the covariance matrix using the inner-product form [Eq. (2.30)] on the centered column vectors. We have

$$\sigma_1^2 = \frac{1}{5}Z_1^T Z_1 = 1.2/5 = 0.24$$

$$\sigma_2^2 = \frac{1}{5}Z_2^T Z_2 = 1.2/5 = 0.24$$

$$\sigma_{12} = \frac{1}{5}Z_1^T Z_2 = -1.2/5 = -0.24$$

Thus, the sample covariance matrix is given as

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}$$

which matches the result obtained by using the multivariate Bernoulli modeling approach.

Multinomial Distribution: Number of Occurrences

Given a multivariate Bernoulli variable $\mathbf{X}$ and a random sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ drawn from $\mathbf{X}$. Let N_i be the random variable corresponding to the number of occurrences of symbol a_i in the sample, and let $\mathbf{N} = (N_1, N_2, \ldots, N_m)^T$ denote the vector random variable corresponding to the joint distribution of the number of occurrences over all the symbols. Then $\mathbf{N}$ has a multinomial distribution, given as

$$f\big(\mathbf{N} = (n_1, n_2, \ldots, n_m) \mid \mathbf{p}\big) = \binom{n}{n_1 n_2 \ldots n_m}\prod_{i=1}^{m} p_i^{n_i}$$

We can see that this is a direct generalization of the binomial distribution in Eq. (3.3). The term

$$\binom{n}{n_1 n_2 \ldots n_m} = \frac{n!}{n_1! n_2! \ldots n_m!}$$

denotes the number of ways of choosing n_i occurrences of each symbol a_i from a sample of size n, with $\sum_{i=1}^{m} n_i = n$.

The mean and covariance matrix of $\mathbf{N}$ are given as n times the mean and covariance matrix of $\mathbf{X}$. That is, the mean of $\mathbf{N}$ is given as

$$\boldsymbol{\mu}_{\mathbf{N}} = E[\mathbf{N}] = nE[\mathbf{X}] = n \cdot \boldsymbol{\mu} = n \cdot \mathbf{p} = \begin{pmatrix} np_1 \\ \vdots \\ np_m \end{pmatrix}$$

and its covariance matrix is given as

$$\boldsymbol{\Sigma}_{\mathbf{N}} = n \cdot (\mathbf{P} - \mathbf{p}\mathbf{p}^T) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_m \\ -np_1p_2 & np_2(1-p_2) & \cdots & -np_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -np_1p_m & -np_2p_m & \cdots & np_m(1-p_m) \end{pmatrix}$$

Likewise the sample mean and covariance matrix for $\mathbf{N}$ are given as

$$\hat{\boldsymbol{\mu}}_{\mathbf{N}} = n\hat{\mathbf{p}} \qquad\qquad \widehat{\boldsymbol{\Sigma}}_{\mathbf{N}} = n\left(\widehat{\mathbf{P}} - \hat{\mathbf{p}}\hat{\mathbf{p}}^T\right)$$

3.2 BIVARIATE ANALYSIS

Assume that the data comprises two categorical attributes, X_1 and X_2, with

$$dom(X_1) = \{a_{11}, a_{12}, \ldots, a_{1m_1}\}$$
$$dom(X_2) = \{a_{21}, a_{22}, \ldots, a_{2m_2}\}$$

We are given n categorical points of the form $\mathbf{x}_i = (x_{i1}, x_{i2})^T$ with $x_{i1} \in dom(X_1)$ and $x_{i2} \in dom(X_2)$. The dataset is thus an $n \times 2$ symbolic data matrix:

$$\mathbf{D} = \left(\begin{array}{cc} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{array}\right)$$

We can model X_1 and X_2 as multivariate Bernoulli variables $\mathbf{X}_1$ and $\mathbf{X}_2$ with dimensions m_1 and m_2, respectively. The probability mass functions for $\mathbf{X}_1$ and $\mathbf{X}_2$ are

given according to Eq. (3.4):

$$P(\mathbf{X}_1 = \mathbf{e}_{1i}) = f_1(\mathbf{e}_{1i}) = p_i^1 = \prod_{k=1}^{m_1} (p_i^1)^{e_{ik}^1}$$

$$P(\mathbf{X}_2 = \mathbf{e}_{2j}) = f_2(\mathbf{e}_{2j}) = p_j^2 = \prod_{k=1}^{m_2} (p_j^2)^{e_{jk}^2}$$

where $\mathbf{e}_{1i}$ is the ith standard basis vector in $\mathbb{R}^{m_1}$ (for attribute X_1) whose kth component is e_{ik}^1, and $\mathbf{e}_{2j}$ is the jth standard basis vector in $\mathbb{R}^{m_2}$ (for attribute X_2) whose kth component is e_{jk}^2. Further, the parameter p_i^1 denotes the probability of observing symbol a_{1i}, and p_j^2 denotes the probability of observing symbol a_{2j}. Together they must satisfy the conditions: $\sum_{i=1}^{m_1} p_i^1 = 1$ and $\sum_{j=1}^{m_2} p_j^2 = 1$.

The joint distribution of $\mathbf{X}_1$ and $\mathbf{X}_2$ is modeled as the $d' = m_1 + m_2$ dimensional vector variable $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, specified by the mapping

$$\mathbf{X}\big((v_1, v_2)^T\big) = \begin{pmatrix} \mathbf{X}_1(v_1) \\ \mathbf{X}_2(v_2) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_{1i} \\ \mathbf{e}_{2j} \end{pmatrix}$$

provided that $v_1 = a_{1i}$ and $v_2 = a_{2j}$. The range of $\mathbf{X}$ thus consists of $m_1 \times m_2$ distinct pairs of vector values $\big\{(\mathbf{e}_{1i}, \mathbf{e}_{2j})^T\big\}$, with $1 \le i \le m_1$ and $1 \le j \le m_2$. The joint PMF of $\mathbf{X}$ is given as

$$P\big(\mathbf{X} = (\mathbf{e}_{1i}, \mathbf{e}_{2j})^T\big) = f(\mathbf{e}_{1i}, \mathbf{e}_{2j}) = p_{ij} = \prod_{r=1}^{m_1} \prod_{s=1}^{m_2} p_{ij}^{e_{ir}^1 \cdot e_{js}^2}$$

where p_{ij} the probability of observing the symbol pair (a_{1i}, a_{2j}). These probability parameters must satisfy the condition $\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} p_{ij} = 1$. The joint PMF for $\mathbf{X}$ can be expressed as the $m_1 \times m_2$ matrix

$$\mathbf{P}_{12} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m_2} \\ p_{21} & p_{22} & \cdots & p_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m_1 1} & p_{m_1 2} & \cdots & p_{m_1 m_2} \end{pmatrix} \tag{3.10}$$

Example 3.7. Consider the discretized `sepal length` attribute (X_1) in Table 3.1. We also discretize the `sepal width` attribute (X_2) into three values as shown in Table 3.3. We thus have

$$dom(X_1) = \{a_{11} = \texttt{VeryShort}, a_{12} = \texttt{Short}, a_{13} = \texttt{Long}, a_{14} = \texttt{VeryLong}\}$$

$$dom(X_2) = \{a_{21} = \texttt{Short}, a_{22} = \texttt{Medium}, a_{23} = \texttt{Long}\}$$

The symbolic point $\mathbf{x} = (\texttt{Short}, \texttt{Long}) = (a_{12}, a_{23})$, is mapped to the vector

$$\mathbf{X}(\mathbf{x}) = \begin{pmatrix} \mathbf{e}_{12} \\ \mathbf{e}_{23} \end{pmatrix} = (0, 1, 0, 0 \mid 0, 0, 1)^T \in \mathbb{R}^7$$

Table 3.3. Discretized `sepal width` attribute

Bins	Domain	Counts
[2.0, 2.8]	Short (a_1)	47
(2.8, 3.6]	Medium (a_2)	88
(3.6, 4.4]	Long (a_3)	15

where we use | to demarcate the two subvectors $\mathbf{e}_{12} = (0,1,0,0)^T \in \mathbb{R}^4$ and $\mathbf{e}_{23} = (0,0,1)^T \in \mathbb{R}^3$, corresponding to symbolic attributes `sepal length` and `sepal width`, respectively. Note that $\mathbf{e}_{12}$ is the second standard basis vector in $\mathbb{R}^4$ for $\mathbf{X}_1$, and $\mathbf{e}_{23}$ is the third standard basis vector in $\mathbb{R}^3$ for $\mathbf{X}_2$.

Mean

The bivariate mean can easily be generalized from Eq. (3.5), as follows:

$$\boldsymbol{\mu} = E[\mathbf{X}] = E\left[\begin{pmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{pmatrix}\right] = \begin{pmatrix}E[\mathbf{X}_1] \\ E[\mathbf{X}_2]\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix} = \begin{pmatrix}\mathbf{p}_1 \\ \mathbf{p}_2\end{pmatrix}$$

where $\boldsymbol{\mu}_1 = \mathbf{p}_1 = (p_1^1, \ldots, p_{m_1}^1)^T$ and $\boldsymbol{\mu}_2 = \mathbf{p}_2 = (p_1^2, \ldots, p_{m_2}^2)^T$ are the mean vectors for $\mathbf{X}_1$ and $\mathbf{X}_2$. The vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ also represent the probability mass functions for $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

Sample Mean

The sample mean can also be generalized from Eq. (3.6), by placing a probability mass of $\frac{1}{n}$ at each point:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \frac{1}{n}\begin{pmatrix}\sum_{i=1}^{m_1} n_i^1 \mathbf{e}_{1i} \\ \sum_{j=1}^{m_2} n_j^2 \mathbf{e}_{2j}\end{pmatrix} = \frac{1}{n}\begin{pmatrix}n_1^1 \\ \vdots \\ n_{m_1}^1 \\ n_1^2 \\ \vdots \\ n_{m_2}^2\end{pmatrix} = \begin{pmatrix}\hat{p}_1^1 \\ \vdots \\ \hat{p}_{m_1}^1 \\ \hat{p}_1^2 \\ \vdots \\ \hat{p}_{m_2}^2\end{pmatrix} = \begin{pmatrix}\hat{\mathbf{p}}_1 \\ \hat{\mathbf{p}}_2\end{pmatrix} = \begin{pmatrix}\hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2\end{pmatrix}$$

where n_j^i is the observed frequency of symbol a_{ij} in the sample of size n, and $\hat{\boldsymbol{\mu}}_i = \hat{\mathbf{p}}_i = (p_1^i, p_2^i, \ldots, p_{m_i}^i)^T$ is the sample mean vector for $\mathbf{X}_i$, which is also the empirical PMF for attribute $\mathbf{X}_i$.

Covariance Matrix

The covariance matrix for $\mathbf{X}$ is the $d' \times d' = (m_1 + m_2) \times (m_1 + m_2)$ matrix given as

$$\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22}\end{pmatrix} \tag{3.11}$$

where $\boldsymbol{\Sigma}_{11}$ is the $m_1 \times m_1$ covariance matrix for $\mathbf{X}_1$, and $\boldsymbol{\Sigma}_{22}$ is the $m_2 \times m_2$ covariance matrix for $\mathbf{X}_2$, which can be computed using Eq. (3.8). That is,

$$\boldsymbol{\Sigma}_{11} = \mathbf{P}_1 - \mathbf{p}_1\mathbf{p}_1^T$$

$$\boldsymbol{\Sigma}_{22} = \mathbf{P}_2 - \mathbf{p}_2\mathbf{p}_2^T$$

where $\mathbf{P}_1 = \text{diag}(\mathbf{p}_1)$ and $\mathbf{P}_2 = \text{diag}(\mathbf{p}_2)$. Further, $\mathbf{\Sigma}_{12}$ is the $m_1 \times m_2$ covariance matrix between variables $\mathbf{X}_1$ and $\mathbf{X}_2$, given as

$$\begin{aligned}
\mathbf{\Sigma}_{12} &= E[(\mathbf{X}_1 - \boldsymbol{\mu}_1)(\mathbf{X}_2 - \boldsymbol{\mu}_2)^T] \\
&= E[\mathbf{X}_1\mathbf{X}_2^T] - E[\mathbf{X}_1]E[\mathbf{X}_2]^T \\
&= \mathbf{P}_{12} - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^T \\
&= \mathbf{P}_{12} - \mathbf{p}_1\mathbf{p}_2^T \\
&= \begin{pmatrix} p_{11} - p_1^1 p_1^2 & p_{12} - p_1^1 p_2^2 & \cdots & p_{1m_2} - p_1^1 p_{m_2}^2 \\ p_{21} - p_2^1 p_1^2 & p_{22} - p_2^1 p_2^2 & \cdots & p_{2m_2} - p_2^1 p_{m_2}^2 \\ \vdots & \vdots & \ddots & \vdots \\ p_{m_1 1} - p_{m_1}^1 p_1^2 & p_{m_1 2} - p_{m_1}^1 p_2^2 & \cdots & p_{m_1 m_2} - p_{m_1}^1 p_{m_2}^2 \end{pmatrix}
\end{aligned}$$

where $\mathbf{P}_{12}$ represents the joint PMF for $\mathbf{X}$ given in Eq. (3.10).

Incidentally, each row and each column of $\mathbf{\Sigma}_{12}$ sums to zero. For example, consider row i and column j:

$$\sum_{k=1}^{m_2}(p_{ik} - p_i^1 p_k^2) = \left(\sum_{k=1}^{m_2} p_{ik}\right) - p_i^1 = p_i^1 - p_i^1 = 0$$

$$\sum_{k=1}^{m_1}(p_{kj} - p_k^1 p_j^2) = \left(\sum_{k=1}^{m_1} p_{kj}\right) - p_j^2 = p_j^2 - p_j^2 = 0$$

which follows from the fact that summing the joint mass function over all values of $\mathbf{X}_2$, yields the marginal distribution of $\mathbf{X}_1$, and summing it over all values of $\mathbf{X}_1$ yields the marginal distribution for $\mathbf{X}_2$. Note that p_j^2 is the probability of observing symbol a_{2j}; it should not be confused with the square of p_j. Combined with the fact that $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$ also have row and column sums equal to zero via Eq. (3.7), the full covariance matrix $\mathbf{\Sigma}$ has rows and columns that sum up to zero.

Sample Covariance Matrix

The sample covariance matrix is given as

$$\widehat{\mathbf{\Sigma}} = \begin{pmatrix} \widehat{\mathbf{\Sigma}}_{11} & \widehat{\mathbf{\Sigma}}_{12} \\ \widehat{\mathbf{\Sigma}}_{12}^T & \widehat{\mathbf{\Sigma}}_{22} \end{pmatrix} \tag{3.12}$$

where

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_{11} &= \widehat{\mathbf{P}}_1 - \hat{\mathbf{p}}_1\hat{\mathbf{p}}_1^T \\
\widehat{\mathbf{\Sigma}}_{22} &= \widehat{\mathbf{P}}_2 - \hat{\mathbf{p}}_2\hat{\mathbf{p}}_2^T \\
\widehat{\mathbf{\Sigma}}_{12} &= \widehat{\mathbf{P}}_{12} - \hat{\mathbf{p}}_1\hat{\mathbf{p}}_2^T
\end{aligned}$$

Here $\widehat{\mathbf{P}}_1 = \text{diag}(\hat{\mathbf{p}}_1)$ and $\widehat{\mathbf{P}}_2 = \text{diag}(\hat{\mathbf{p}}_2)$, and $\hat{\mathbf{p}}_1$ and $\hat{\mathbf{p}}_2$ specify the empirical probability mass functions for $\mathbf{X}_1$, and $\mathbf{X}_2$, respectively. Further, $\widehat{\mathbf{P}}_{12}$ specifies the empirical joint PMF for $\mathbf{X}_1$ and $\mathbf{X}_2$, given as

$$\widehat{\mathbf{P}}_{12}(i, j) = \hat{f}(\mathbf{e}_{1i}, \mathbf{e}_{2j}) = \frac{1}{n}\sum_{k=1}^{n} I_{ij}(\mathbf{x}_k) = \frac{n_{ij}}{n} = \hat{p}_{ij} \tag{3.13}$$

where I_{ij} is the indicator variable

$$I_{ij}(\mathbf{x}_k) = \begin{cases} 1 & \text{if } \mathbf{x}_{k1} = \mathbf{e}_{1i} \text{ and } \mathbf{x}_{k2} = \mathbf{e}_{2j} \\ 0 & \text{otherwise} \end{cases}$$

Taking the sum of $I_{ij}(\mathbf{x}_k)$ over all the n points in the sample yields the number of occurrences, n_{ij}, of the symbol pair (a_{1i}, a_{2j}) in the sample. One issue with the cross-attribute covariance matrix $\widehat{\boldsymbol{\Sigma}}_{12}$ is the need to estimate a quadratic number of parameters. That is, we need to obtain reliable counts n_{ij} to estimate the parameters p_{ij}, for a total of $O(m_1 \times m_2)$ parameters that have to be estimated, which can be a problem if the categorical attributes have many symbols. On the other hand, estimating $\widehat{\boldsymbol{\Sigma}}_{11}$ and $\widehat{\boldsymbol{\Sigma}}_{22}$ requires that we estimate m_1 and m_2 parameters, corresponding to p_i^1 and p_j^2, respectively. In total, computing $\boldsymbol{\Sigma}$ requires the estimation of $m_1 m_2 + m_1 + m_2$ parameters.

Example 3.8. We continue with the bivariate categorical attributes X_1 and X_2 in Example 3.7. From Example 3.4, and from the occurrence counts for each of the values of `sepal width` in Table 3.3, we have

$$\hat{\boldsymbol{\mu}}_1 = \hat{\mathbf{p}}_1 = \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} \qquad \hat{\boldsymbol{\mu}}_2 = \hat{\mathbf{p}}_2 = \frac{1}{150}\begin{pmatrix} 47 \\ 88 \\ 15 \end{pmatrix} = \begin{pmatrix} 0.313 \\ 0.587 \\ 0.1 \end{pmatrix}$$

Thus, the mean for $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$ is given as

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{p}}_1 \\ \hat{\mathbf{p}}_2 \end{pmatrix} = (0.3, 0.333, 0.287, 0.08 \mid 0.313, 0.587, 0.1)^T$$

From Example 3.5 we have

$$\widehat{\boldsymbol{\Sigma}}_{11} = \begin{pmatrix} 0.21 & -0.1 & -0.086 & -0.024 \\ -0.1 & 0.222 & -0.096 & -0.027 \\ -0.086 & -0.096 & 0.204 & -0.023 \\ -0.024 & -0.027 & -0.023 & 0.074 \end{pmatrix}$$

In a similar manner we can obtain

$$\widehat{\boldsymbol{\Sigma}}_{22} = \begin{pmatrix} 0.215 & -0.184 & -0.031 \\ -0.184 & 0.242 & -0.059 \\ -0.031 & -0.059 & 0.09 \end{pmatrix}$$

Next, we use the observed counts in Table 3.4 to obtain the empirical joint PMF for $\mathbf{X}_1$ and $\mathbf{X}_2$ using Eq. (3.13), as plotted in Figure 3.2. From these probabilities we get

$$E[\mathbf{X}_1 \mathbf{X}_2^T] = \widehat{\mathbf{P}}_{12} = \frac{1}{150}\begin{pmatrix} 7 & 33 & 5 \\ 24 & 18 & 8 \\ 13 & 30 & 0 \\ 3 & 7 & 2 \end{pmatrix} = \begin{pmatrix} 0.047 & 0.22 & 0.033 \\ 0.16 & 0.12 & 0.053 \\ 0.087 & 0.2 & 0 \\ 0.02 & 0.047 & 0.013 \end{pmatrix}$$

Table 3.4. Observed Counts (n_{ij}): `sepal length` and `sepal width`

		X_2		
		`Short` ($\mathbf{e}_{21}$)	`Medium` ($\mathbf{e}_{22}$)	`Long` ($\mathbf{e}_{23}$)
X_1	`Very Short` ($\mathbf{e}_{11}$)	7	33	5
	`Short` ($\mathbf{e}_{22}$)	24	18	8
	`Long` ($\mathbf{e}_{13}$)	13	30	0
	`Very Long` ($\mathbf{e}_{14}$)	3	7	2

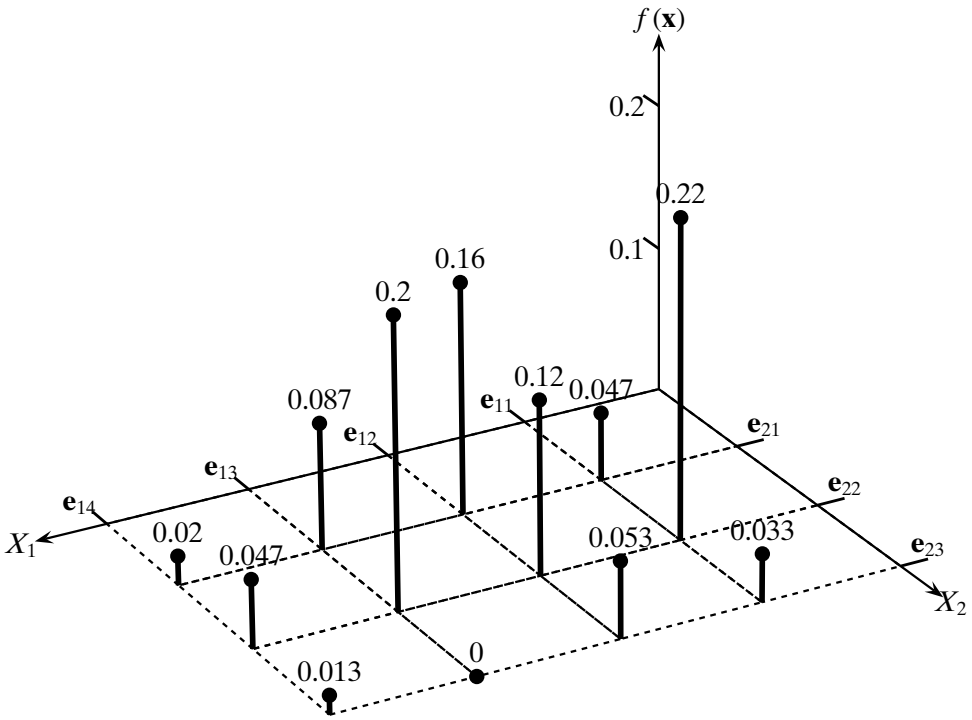


Figure 3.2. Empirical joint probability mass function: `sepal length` and `sepal width`.

Further, we have

$$E[\mathbf{X}_1]E[\mathbf{X}_2]^T = \hat{\boldsymbol{\mu}}_1\hat{\boldsymbol{\mu}}_2^T = \hat{\mathbf{p}}_1\hat{\mathbf{p}}_2^T$$

$$= \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} \begin{pmatrix} 0.313 & 0.587 & 0.1 \end{pmatrix}$$

$$= \begin{pmatrix} 0.094 & 0.176 & 0.03 \\ 0.104 & 0.196 & 0.033 \\ 0.09 & 0.168 & 0.029 \\ 0.025 & 0.047 & 0.008 \end{pmatrix}$$

We can now compute the across-attribute sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{12}$ for $\mathbf{X}_1$ and $\mathbf{X}_2$ using Eq. (3.11), as follows:

$$\widehat{\boldsymbol{\Sigma}}_{12} = \widehat{\mathbf{P}}_{12} - \hat{\mathbf{p}}_1 \hat{\mathbf{p}}_2^T = \begin{pmatrix} -0.047 & 0.044 & 0.003 \\ 0.056 & -0.076 & 0.02 \\ -0.003 & 0.032 & -0.029 \\ -0.005 & 0 & 0.005 \end{pmatrix}$$

One can observe that each row and column in $\widehat{\boldsymbol{\Sigma}}_{12}$ sums to zero. Putting it all together, from $\widehat{\boldsymbol{\Sigma}}_{11}$, $\widehat{\boldsymbol{\Sigma}}_{22}$ and $\widehat{\boldsymbol{\Sigma}}_{12}$ we obtain the sample covariance matrix as follows

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{11} & \widehat{\boldsymbol{\Sigma}}_{12} \\ \widehat{\boldsymbol{\Sigma}}_{12}^T & \widehat{\boldsymbol{\Sigma}}_{22} \end{pmatrix} = \left(\begin{array}{cccc|ccc} 0.21 & -0.1 & -0.086 & -0.024 & -0.047 & 0.044 & 0.003 \\ -0.1 & 0.222 & -0.096 & -0.027 & 0.056 & -0.076 & 0.02 \\ -0.086 & -0.096 & 0.204 & -0.023 & -0.003 & 0.032 & -0.029 \\ -0.024 & -0.027 & -0.023 & 0.074 & -0.005 & 0 & 0.005 \\ \hline -0.047 & 0.056 & -0.003 & -0.005 & 0.215 & -0.184 & -0.031 \\ 0.044 & -0.076 & 0.032 & 0 & -0.184 & 0.242 & -0.059 \\ 0.003 & 0.02 & -0.029 & 0.005 & -0.031 & -0.059 & 0.09 \end{array}\right)$$

In $\widehat{\boldsymbol{\Sigma}}$, each row and column also sums to zero.

3.2.1 Attribute Dependence: Contingency Analysis

Testing for the independence of the two categorical random variables X_1 and X_2 can be done via *contingency table analysis*. The main idea is to set up a hypothesis testing framework, where the null hypothesis H_0 is that $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent, and the alternative hypothesis H_1 is that they are dependent. We then compute the value of the chi-square statistic χ^2 under the null hypothesis. Depending on the p-value, we either accept or reject the null hypothesis; in the latter case the attributes are considered to be dependent.

Contingency Table

A contingency table for $\mathbf{X}_1$ and $\mathbf{X}_2$ is the $m_1 \times m_2$ matrix of observed counts n_{ij} for all pairs of values $(\mathbf{e}_{1i}, \mathbf{e}_{2j})$ in the given sample of size n, defined as

$$\mathbf{N}_{12} = n \cdot \widehat{\mathbf{P}}_{12} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1m_2} \\ n_{21} & n_{22} & \cdots & n_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m_1 1} & n_{m_1 2} & \cdots & n_{m_1 m_2} \end{pmatrix}$$

Table 3.5. Contingency table: `sepal length` vs. `sepal width`

Sepal length (X_1)	Sepal width (X_2)			
	Short a_{21}	Medium a_{22}	Long a_{23}	Row Counts
Very Short (a_{11})	7	33	5	$n_1^1 = 45$
Short (a_{12})	24	18	8	$n_2^1 = 50$
Long (a_{13})	13	30	0	$n_3^1 = 43$
Very Long (a_{14})	3	7	2	$n_4^1 = 12$
Column Counts	$n_1^2 = 47$	$n_2^2 = 88$	$n_3^2 = 15$	$n = 150$

where $\widehat{\mathbf{P}}_{12}$ is the empirical joint PMF for $\mathbf{X}_1$ and $\mathbf{X}_2$, computed via Eq. (3.13). The contingency table is then augmented with row and column marginal counts, as follows:

$$\mathbf{N}_1 = n \cdot \hat{\mathbf{p}}_1 = \begin{pmatrix} n_1^1 \\ \vdots \\ n_{m_1}^1 \end{pmatrix} \qquad \mathbf{N}_2 = n \cdot \hat{\mathbf{p}}_2 = \begin{pmatrix} n_1^2 \\ \vdots \\ n_{m_2}^2 \end{pmatrix}$$

Note that the marginal row and column entries and the sample size satisfy the following constraints:

$$n_i^1 = \sum_{j=1}^{m_2} n_{ij} \qquad n_j^2 = \sum_{i=1}^{m_1} n_{ij} \qquad n = \sum_{i=1}^{m_1} n_i^1 = \sum_{j=1}^{m_2} n_j^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} n_{ij}$$

It is worth noting that both $\mathbf{N}_1$ and $\mathbf{N}_2$ have a multinomial distribution with parameters $\mathbf{p}_1 = (p_1^1, \ldots, p_{m_1}^1)$ and $\mathbf{p}_2 = (p_1^2, \ldots, p_{m_2}^2)$, respectively. Further, $\mathbf{N}_{12}$ also has a multinomial distribution with parameters $\mathbf{P}_{12} = \{p_{ij}\}$, for $1 \le i \le m_1$ and $1 \le j \le m_2$.

Example 3.9 (Contingency Table). Table 3.4 shows the observed counts for the discretized `sepal length` (X_1) and `sepal width` (X_2) attributes. Augmenting the table with the row and column marginal counts and the sample size yields the final contingency table shown in Table 3.5.

χ^2 Statistic and Hypothesis Testing

Under the null hypothesis $\mathbf{X}_1$ and $\mathbf{X}_2$ are assumed to be independent, which means that their joint probability mass function is given as

$$\hat{p}_{ij} = \hat{p}_i^1 \cdot \hat{p}_j^2$$

Under this independence assumption, the expected frequency for each pair of values is given as

$$e_{ij} = n \cdot \hat{p}_{ij} = n \cdot \hat{p}_i^1 \cdot \hat{p}_j^2 = n \cdot \frac{n_i^1}{n} \cdot \frac{n_j^2}{n} = \frac{n_i^1 n_j^2}{n} \tag{3.14}$$

However, from the sample we already have the observed frequency of each pair of values, n_{ij}. We would like to determine whether there is a significant difference in the observed and expected frequencies for each pair of values. If there is no

significant difference, then the independence assumption is valid and we accept the null hypothesis that the attributes are independent. On the other hand, if there is a significant difference, then the null hypothesis should be rejected and we conclude that the attributes are dependent.

The χ^2 statistic quantifies the difference between observed and expected counts for each pair of values; it is defined as follows:

$$\chi^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{3.15}$$

At this point, we need to determine the probability of obtaining the computed χ^2 value. In general, this can be rather difficult if we do not know the sampling distribution of a given statistic. Fortunately, for the χ^2 statistic it is known that its sampling distribution follows the *chi-squared* density function with q degrees of freedom:

$$f(x|q) = \frac{1}{2^{q/2}\Gamma(q/2)} x^{\frac{q}{2}-1} \mathrm{e}^{-\frac{x}{2}} \tag{3.16}$$

where the gamma function Γ is defined as

$$\Gamma(k > 0) = \int_0^\infty x^{k-1} \mathrm{e}^{-x} dx \tag{3.17}$$

The degrees of freedom, q, represent the number of independent parameters. In the contingency table there are $m_1 \times m_2$ observed counts n_{ij}. However, note that each row i and each column j must sum to n_i^1 and n_j^2, respectively. Further, the sum of the row and column marginals must also add to n; thus we have to remove $(m_1 + m_2)$ parameters from the number of independent parameters. However, doing this removes one of the parameters, say $n_{m_1 m_2}$, twice, so we have to add back one to the count. The total degrees of freedom is therefore

$$\begin{aligned} q &= |dom(X_1)| \times |dom(X_2)| - (|dom(X_1)| + |dom(X_2)|) + 1 \\ &= m_1 m_2 - m_1 - m_2 + 1 \\ &= (m_1 - 1)(m_2 - 1) \end{aligned}$$

p-value

The *p-value* of a statistic θ is defined as the probability of obtaining a value at least as extreme as the observed value, say z, under the null hypothesis, defined as

$$p\text{-value}(z) = P(\theta \geq z) = 1 - F(\theta)$$

where $F(\theta)$ is the cumulative probability distribution for the statistic.

The p-value gives a measure of how surprising is the observed value of the statistic. If the observed value lies in a low-probability region, then the value is more surprising. In general, the lower the p-value, the more surprising the observed value, and the

Table 3.6. Expected counts

		X_2		
		Short (a_{21})	Medium (a_{22})	Short (a_{23})
X_1	Very Short (a_{11})	14.1	26.4	4.5
	Short (a_{12})	15.67	29.33	5.0
	Long (a_{13})	13.47	25.23	4.3
	Very Long (a_{14})	3.76	7.04	1.2

more the grounds for rejecting the null hypothesis. The null hypothesis is rejected if the p-value is below some *significance level*, α. For example, if $\alpha = 0.01$, then we reject the null hypothesis if $p\text{-value}(z) \leq \alpha$. The significance level α corresponds to the probability of rejecting the null hypothesis when it is true. For a given significance level α, the value of the test statistic, say z, with a p-value of $p\text{-value}(z) = \alpha$, is called a *critical value*. An alternative test for rejection of the null hypothesis is to check if $\chi^2 > z$, as in that case the p-value of the observed χ^2 value is bounded by α, that is, $p\text{-value}(\chi^2) \leq p\text{-value}(z) = \alpha$. The value $1 - \alpha$ is also called the *confidence level*.

Example 3.10. Consider the contingency table for `sepal length` and `sepal width` in Table 3.5. We compute the expected counts using Eq. (3.14); these counts are shown in Table 3.6. For example, we have

$$e_{11} = \frac{n_1^1 n_1^2}{n} = \frac{45 \cdot 47}{150} = \frac{2115}{150} = 14.1$$

Next we use Eq. (3.15) to compute the value of the χ^2 statistic, which is given as $\chi^2 = 21.8$.

Further, the number of degrees of freedom is given as

$$q = (m_1 - 1) \cdot (m_2 - 1) = 3 \cdot 2 = 6$$

The plot of the chi-squared density function with 6 degrees of freedom is shown in Figure 3.3. From the cumulative chi-squared distribution, we obtain

$$p\text{-value}(21.8) = 1 - F(21.8|6) = 1 - 0.9987 = 0.0013$$

At a significance level of $\alpha = 0.01$, we would certainly be justified in rejecting the null hypothesis because the large value of the χ^2 statistic is indeed surprising. Further, at the 0.01 significance level, the critical value of the statistic is

$$z = F^{-1}(1 - 0.01|6) = F^{-1}(0.99|6) = 16.81$$

This critical value is also shown in Figure 3.3, and we can clearly see that the observed value of 21.8 is in the rejection region, as $21.8 > z = 16.81$. In effect, we reject the null hypothesis that `sepal length` and `sepal width` are independent, and accept the alternative hypothesis that they are dependent.

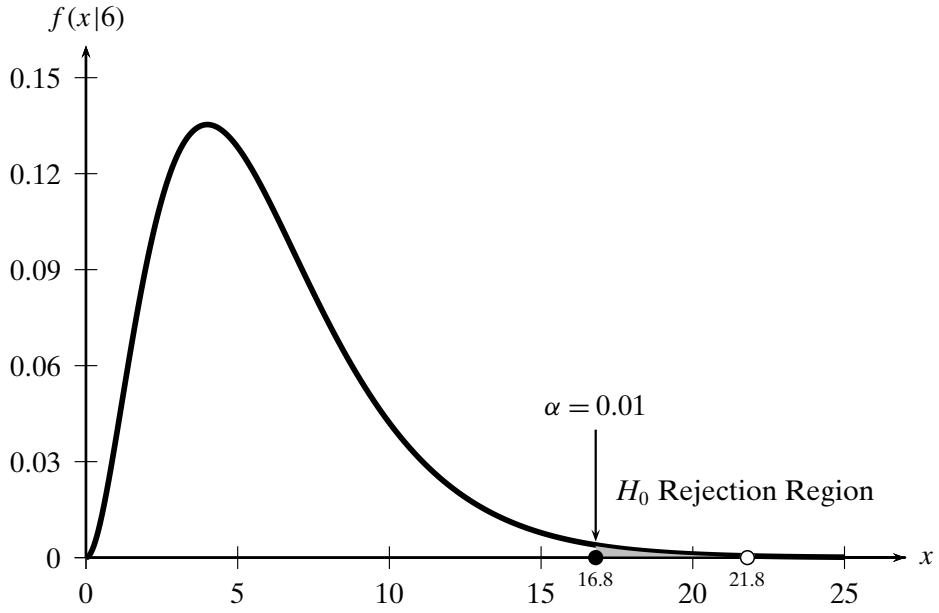


Figure 3.3. Chi-squared distribution ($q = 6$).

3.3 MULTIVARIATE ANALYSIS

Assume that the dataset comprises d categorical attributes X_j $(1 \leq j \leq d)$ with $dom(X_j) = \{a_{j1}, a_{j2}, \ldots, a_{jm_j}\}$. We are given n categorical points of the form $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ with $x_{ij} \in dom(X_j)$. The dataset is thus an $n \times d$ symbolic matrix

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ \hline x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Each attribute X_i is modeled as an m_i-dimensional multivariate Bernoulli variable $\mathbf{X}_i$, and their joint distribution is modeled as a $d' = \sum_{j=1}^{d} m_j$ dimensional vector random variable

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_d \end{pmatrix}$$

Each categorical data point $\mathbf{v} = (v_1, v_2, \ldots, v_d)^T$ is therefore represented as a d'-dimensional binary vector

$$\mathbf{X}(\mathbf{v}) = \begin{pmatrix} \mathbf{X}_1(v_1) \\ \vdots \\ \mathbf{X}_d(v_d) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_{1k_1} \\ \vdots \\ \mathbf{e}_{dk_d} \end{pmatrix}$$

provided $v_i = a_{ik_i}$, the k_ith symbol of X_i. Here $\mathbf{e}_{ik_i}$ is the k_ith standard basis vector in $\mathbb{R}^{m_i}$.

Mean

Generalizing from the bivariate case, the mean and sample mean for $\mathbf{X}$ are given as

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_d \end{pmatrix} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_d \end{pmatrix} \qquad \hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 \\ \vdots \\ \hat{\boldsymbol{\mu}}_d \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{p}}_1 \\ \vdots \\ \hat{\mathbf{p}}_d \end{pmatrix}$$

where $\mathbf{p}_i = (p_1^i, \ldots, p_{m_i}^i)^T$ is the PMF for $\mathbf{X}_i$, and $\hat{\mathbf{p}}_i = (\hat{p}_1^i, \ldots, \hat{p}_{m_i}^i)^T$ is the empirical PMF for $\mathbf{X}_i$.

Covariance Matrix

The covariance matrix for $\mathbf{X}$, and its estimate from the sample, are given as the $d' \times d'$ matrices:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1d} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2d} \\ \cdots & \cdots & \ddots & \cdots \\ \boldsymbol{\Sigma}_{1d}^T & \boldsymbol{\Sigma}_{2d}^T & \cdots & \boldsymbol{\Sigma}_{dd} \end{pmatrix} \qquad \widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{11} & \widehat{\boldsymbol{\Sigma}}_{12} & \cdots & \widehat{\boldsymbol{\Sigma}}_{1d} \\ \widehat{\boldsymbol{\Sigma}}_{12}^T & \widehat{\boldsymbol{\Sigma}}_{22} & \cdots & \widehat{\boldsymbol{\Sigma}}_{2d} \\ \cdots & \cdots & \ddots & \cdots \\ \widehat{\boldsymbol{\Sigma}}_{1d}^T & \widehat{\boldsymbol{\Sigma}}_{2d}^T & \cdots & \widehat{\boldsymbol{\Sigma}}_{dd} \end{pmatrix}$$

where $d' = \sum_{i=1}^{d} m_i$, and $\boldsymbol{\Sigma}_{ij}$ (and $\widehat{\boldsymbol{\Sigma}}_{ij}$) is the $m_i \times m_j$ covariance matrix (and its estimate) for attributes $\mathbf{X}_i$ and $\mathbf{X}_j$:

$$\boldsymbol{\Sigma}_{ij} = \mathbf{P}_{ij} - \mathbf{p}_i \mathbf{p}_j^T \qquad \widehat{\boldsymbol{\Sigma}}_{ij} = \widehat{\mathbf{P}}_{ij} - \hat{\mathbf{p}}_i \hat{\mathbf{p}}_j^T \tag{3.18}$$

Here $\mathbf{P}_{ij}$ is the joint PMF and $\widehat{\mathbf{P}}_{ij}$ is the empirical joint PMF for $\mathbf{X}_i$ and $\mathbf{X}_j$, which can be computed using Eq. (3.13).

Example 3.11 (Multivariate Analysis). Let us consider the 3-dimensional subset of the Iris dataset, with the discretized attributes `sepal length` (X_1) and `sepal width` (X_2), and the categorical attribute `class` (X_3). The domains for X_1 and X_2 are given in Table 3.1 and Table 3.3, respectively, and $dom(X_3) = \{\texttt{iris-versicolor}, \texttt{iris-setosa}, \texttt{iris-virginica}\}$. Each value of X_3 occurs 50 times.

The categorical point $\mathbf{x} = (\texttt{Short}, \texttt{Medium}, \texttt{iris-versicolor})$ is modeled as the vector

$$\mathbf{X}(\mathbf{x}) = \begin{pmatrix} \mathbf{e}_{12} \\ \mathbf{e}_{22} \\ \mathbf{e}_{31} \end{pmatrix} = (0, 1, 0, 0 \mid 0, 1, 0 \mid 1, 0, 0)^T \in \mathbb{R}^{10}$$

From Example 3.8 and the fact that each value in $dom(X_3)$ occurs 50 times in a sample of $n = 150$, the sample mean is given as

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2 \\ \hat{\boldsymbol{\mu}}_3 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{p}}_1 \\ \hat{\mathbf{p}}_2 \\ \hat{\mathbf{p}}_3 \end{pmatrix} = (0.3, 0.333, 0.287, 0.08 \mid 0.313, 0.587, 0.1 \mid 0.33, 0.33, 0.33)^T$$

Using $\hat{\mathbf{p}}_3 = (0.33, 0.33, 0.33)^T$ we can compute the sample covariance matrix for X_3 using Eq. (3.9):

$$\widehat{\boldsymbol{\Sigma}}_{33} = \begin{pmatrix} 0.222 & -0.111 & -0.111 \\ -0.111 & 0.222 & -0.111 \\ -0.111 & -0.111 & 0.222 \end{pmatrix}$$

Using Eq. (3.18) we obtain

$$\widehat{\boldsymbol{\Sigma}}_{13} = \begin{pmatrix} -0.067 & 0.16 & -0.093 \\ 0.082 & -0.038 & -0.044 \\ 0.011 & -0.096 & 0.084 \\ -0.027 & -0.027 & 0.053 \end{pmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}_{23} = \begin{pmatrix} 0.076 & -0.098 & 0.022 \\ -0.042 & 0.044 & -0.002 \\ -0.033 & 0.053 & -0.02 \end{pmatrix}$$

Combined with $\widehat{\boldsymbol{\Sigma}}_{11}$, $\widehat{\boldsymbol{\Sigma}}_{22}$ and $\widehat{\boldsymbol{\Sigma}}_{12}$ from Example 3.8, the final sample covariance matrix is the 10×10 symmetric matrix given as

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{11} & \widehat{\boldsymbol{\Sigma}}_{12} & \widehat{\boldsymbol{\Sigma}}_{13} \\ \widehat{\boldsymbol{\Sigma}}_{12}^T & \widehat{\boldsymbol{\Sigma}}_{22} & \widehat{\boldsymbol{\Sigma}}_{23} \\ \widehat{\boldsymbol{\Sigma}}_{13}^T & \widehat{\boldsymbol{\Sigma}}_{23}^T & \widehat{\boldsymbol{\Sigma}}_{33} \end{pmatrix}$$

3.3.1 Multiway Contingency Analysis

For multiway dependence analysis, we have to first determine the empirical joint probability mass function for $\mathbf{X}$:

$$\hat{f}(\mathbf{e}_{1i_1}, \mathbf{e}_{2i_2}, \ldots, \mathbf{e}_{di_d}) = \frac{1}{n} \sum_{k=1}^{n} I_{i_1 i_2 \ldots i_d}(\mathbf{x}_k) = \frac{n_{i_1 i_2 \ldots i_d}}{n} = \hat{p}_{i_1 i_2 \ldots i_d}$$

where $I_{i_1 i_2 \ldots i_d}$ is the indicator variable

$$I_{i_1 i_2 \ldots i_d}(\mathbf{x}_k) = \begin{cases} 1 & \text{if } x_{k1} = \mathbf{e}_{1i_1}, x_{k2} = \mathbf{e}_{2i_2}, \ldots, x_{kd} = \mathbf{e}_{di_d} \\ 0 & \text{otherwise} \end{cases}$$

The sum of $I_{i_1 i_2 \ldots i_d}$ over all the n points in the sample yields the number of occurrences, $n_{i_1 i_2 \ldots i_d}$, of the symbolic vector $(a_{1i_1}, a_{2i_2}, \ldots, a_{di_d})$. Dividing the occurrences by the sample size results in the probability of observing those symbols. Using the notation $\mathbf{i} = (i_1, i_2, \ldots, i_d)$ to denote the index tuple, we can write the joint empirical PMF as the d-dimensional matrix $\widehat{\mathbf{P}}$ of size $m_1 \times m_2 \times \cdots \times m_d = \prod_{i=1}^{d} m_i$, given as

$$\widehat{\mathbf{P}}(\mathbf{i}) = \left\{\hat{p}_{\mathbf{i}}\right\} \text{ for all index tuples } \mathbf{i}, \text{ with } 1 \le i_1 \le m_1, \ldots, 1 \le i_d \le m_d$$

where $\hat{p}_{\mathbf{i}} = \hat{p}_{i_1 i_2 \ldots i_d}$. The d-dimensional contingency table is then given as

$$\mathbf{N} = n \times \widehat{\mathbf{P}} = \left\{n_{\mathbf{i}}\right\} \text{ for all index tuples } \mathbf{i}, \text{ with } 1 \le i_1 \le m_1, \ldots, 1 \le i_d \le m_d$$

where $n_{\mathbf{i}} = n_{i_1 i_2 \ldots i_d}$. The contingency table is augmented with the marginal count vectors $\mathbf{N}_i$ for all d attributes $\mathbf{X}_i$:

$$\mathbf{N}_i = n\hat{\mathbf{p}}_i = \begin{pmatrix} n_1^i \\ \vdots \\ n_{m_i}^i \end{pmatrix}$$

where $\hat{\mathbf{p}}_i$ is the empirical PMF for $\mathbf{X}_i$.

χ^2-Test

We can test for a d-way dependence between the d categorical attributes using the null hypothesis H_0 that they are d-way independent. The alternative hypothesis H_1 is that they are not d-way independent, that is, they are dependent in some way. Note that d-dimensional contingency analysis indicates whether all d attributes taken together are independent or not. In general we may have to conduct k-way contingency analysis to test if any subset of $k \leq d$ attributes are independent or not.

Under the null hypothesis, the expected number of occurrences of the symbol tuple $(a_{1i_1}, a_{2i_2}, \ldots, a_{di_d})$ is given as

$$e_{\mathbf{i}} = n \cdot \hat{p}_{\mathbf{i}} = n \cdot \prod_{j=1}^{d} \hat{p}_{i_j}^j = \frac{n_{i_1}^1 n_{i_2}^2 \ldots n_{i_d}^d}{n^{d-1}} \tag{3.19}$$

The chi-squared statistic measures the difference between the observed counts $n_{\mathbf{i}}$ and the expected counts $e_{\mathbf{i}}$:

$$\chi^2 = \sum_{\mathbf{i}} \frac{(n_{\mathbf{i}} - e_{\mathbf{i}})^2}{e_{\mathbf{i}}} = \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \cdots \sum_{i_d=1}^{m_d} \frac{(n_{i_1,i_2,\ldots,i_d} - e_{i_1,i_2,\ldots,i_d})^2}{e_{i_1,i_2,\ldots,i_d}} \tag{3.20}$$

The χ^2 statistic follows a chi-squared density function with q degrees of freedom. For the d-way contingency table we can compute q by noting that there are ostensibly $\prod_{i=1}^{d} |dom(X_i)|$ independent parameters (the counts). However, we have to remove $\sum_{i=1}^{d} |dom(X_i)|$ degrees of freedom because the marginal count vector along each dimension $\mathbf{X}_i$ must equal $\mathbf{N}_i$. However, doing so removes one of the parameters d times, so we need to add back $d-1$ to the free parameters count. The total number of degrees of freedom is given as

$$\begin{aligned} q &= \prod_{i=1}^{d} |dom(X_i)| - \sum_{i=1}^{d} |dom(X_i)| + (d-1) \\ &= \left(\prod_{i=1}^{d} m_i\right) - \left(\sum_{i=1}^{d} m_i\right) + d - 1 \end{aligned} \tag{3.21}$$

To reject the null hypothesis, we have to check whether the p-value of the observed χ^2 value is smaller than the desired significance level α (say $\alpha = 0.01$) using the chi-squared density with q degrees of freedom [Eq. (3.16)].

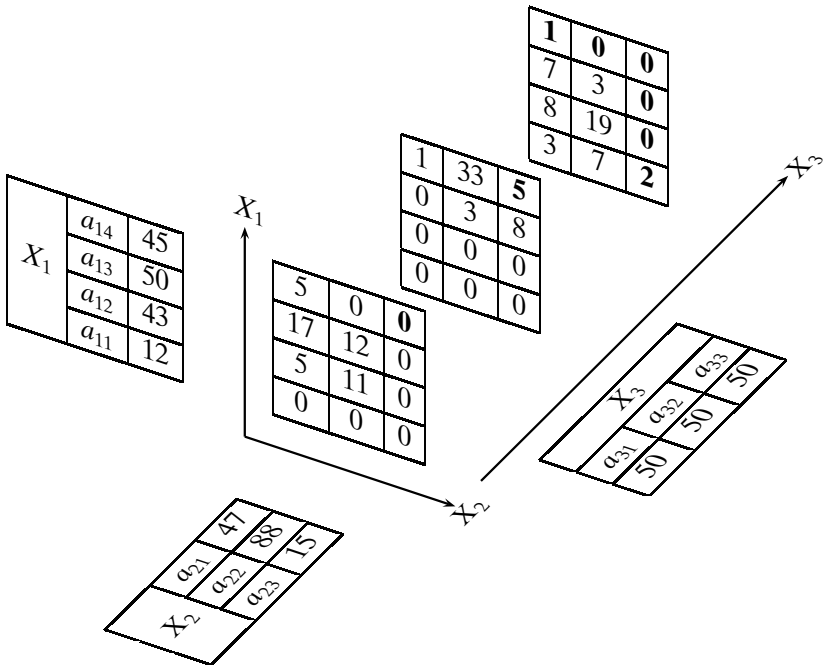


Figure 3.4. 3-Way contingency table (with marginal counts along each dimension).

Table 3.7. 3-Way expected counts

		$X_3(a_{31}/a_{32}/a_{33})$		
		X_2		
		a_{21}	a_{22}	a_{23}
X_1	a_{11}	1.25	2.35	0.40
	a_{12}	4.49	8.41	1.43
	a_{13}	5.22	9.78	1.67
	a_{14}	4.70	8.80	1.50

Example 3.12. Consider the 3-way contingency table in Figure 3.4. It shows the observed counts for each tuple of symbols (a_{1i}, a_{2j}, a_{3k}) for the three attributes `sepal length` (X_1), `sepal width` (X_2), and `class` (X_3). From the marginal counts for X_1 and X_2 in Table 3.5, and the fact that all three values of X_3 occur 50 times, we can compute the expected counts [Eq. (3.19)] for each cell. For instance,

$$e_{(4,1,1)} = \frac{n_4^1 \cdot n_1^2 \cdot n_1^3}{150^2} = \frac{45 \cdot 47 \cdot 50}{150 \cdot 150} = 4.7$$

The expected counts are the same for all three values of X_3 and are given in Table 3.7.

The value of the χ^2 statistic [Eq. (3.20)] is given as

$$\chi^2 = 231.06$$

Using Eq. (3.21), the number of degrees of freedom is given as

$$q = 4 \cdot 3 \cdot 3 - (4+3+3) + 2 = 36 - 10 + 2 = 28$$

In Figure 3.4 the counts in bold are the dependent parameters. All other counts are independent. In fact, any eight distinct cells could have been chosen as the dependent parameters.

For a significance level of $\alpha = 0.01$, the critical value of the chi-square distribution is $z = 48.28$. The observed value of $\chi^2 = 231.06$ is much greater than z, and it is thus extremely unlikely to happen under the null hypothesis. We conclude that the three attributes are not 3-way independent, but rather there is some dependence between them. However, this example also highlights one of the pitfalls of multiway contingency analysis. We can observe in Figure 3.4 that many of the observed counts are zero. This is due to the fact that the sample size is small, and we cannot reliably estimate all the multiway counts. Consequently, the dependence test may not be reliable as well.

3.4 DISTANCE AND ANGLE

With the modeling of categorical attributes as multivariate Bernoulli variables, it is possible to compute the distance or the angle between any two points $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{e}_{1i_1} \\ \vdots \\ \mathbf{e}_{d\,i_d} \end{pmatrix} \qquad \mathbf{x}_j = \begin{pmatrix} \mathbf{e}_{1j_1} \\ \vdots \\ \mathbf{e}_{d\,j_d} \end{pmatrix}$$

The different measures of distance and similarity rely on the number of matching and mismatching values (or symbols) across the d attributes $\mathbf{X}_k$. For instance, we can compute the number of matching values s via the dot product:

$$s = \mathbf{x}_i^T \mathbf{x}_j = \sum_{k=1}^{d} (\mathbf{e}_{ki_k})^T \mathbf{e}_{kj_k}$$

On the other hand, the number of mismatches is simply $d - s$. Also useful is the norm of each point:

$$\|\mathbf{x}_i\|^2 = \mathbf{x}_i^T \mathbf{x}_i = d$$

Euclidean Distance

The Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is given as

$$\delta(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i \mathbf{x}_j + \mathbf{x}_j^T \mathbf{x}_j} = \sqrt{2(d-s)}$$

Thus, the maximum Euclidean distance between any two points is $\sqrt{2d}$, which happens when there are no common symbols between them, that is, when $s = 0$.

Hamming Distance

The *Hamming distance* between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as the number of mismatched values:

$$\delta_H(\mathbf{x}_i, \mathbf{x}_j) = d - s = \frac{1}{2}\delta(\mathbf{x}_i, \mathbf{x}_j)^2$$

Hamming distance is thus equivalent to half the squared Euclidean distance.

Cosine Similarity

The cosine of the angle between $\mathbf{x}_i$ and $\mathbf{x}_j$ is given as

$$\cos\theta = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|} = \frac{s}{d}$$

Jaccard Coefficient

The *Jaccard Coefficient* is a commonly used similarity measure between two categorical points. It is defined as the ratio of the number of matching values to the number of distinct values that appear in both $\mathbf{x}_i$ and $\mathbf{x}_j$, across the d attributes:

$$J(\mathbf{x}_i, \mathbf{x}_j) = \frac{s}{2(d-s)+s} = \frac{s}{2d-s}$$

where we utilize the observation that when the two points do not match for dimension k, they contribute 2 to the distinct symbol count; otherwise, if they match, the number of distinct symbols increases by 1. Over the $d - s$ mismatches and s matches, the number of distinct symbols is $2(d-s)+s$.

Example 3.13. Consider the 3-dimensional categorical data from Example 3.11. The symbolic point (`Short`, `Medium`, `iris-versicolor`) is modeled as the vector

$$\mathbf{x}_1 = \begin{pmatrix} \mathbf{e}_{12} \\ \mathbf{e}_{22} \\ \mathbf{e}_{31} \end{pmatrix} = (0,1,0,0 \mid 0,1,0 \mid 1,0,0)^T \in \mathbb{R}^{10}$$

and the symbolic point (`VeryShort`, `Medium`, `iris-setosa`) is modeled as

$$\mathbf{x}_2 = \begin{pmatrix} \mathbf{e}_{11} \\ \mathbf{e}_{22} \\ \mathbf{e}_{32} \end{pmatrix} = (1,0,0,0 \mid 0,1,0 \mid 0,1,0)^T \in \mathbb{R}^{10}$$

The number of matching symbols is given as

$$\begin{aligned} s = \mathbf{x}_1^T \mathbf{x}_2 &= (\mathbf{e}_{12})^T \mathbf{e}_{11} + (\mathbf{e}_{22})^T \mathbf{e}_{22} + (\mathbf{e}_{31})^T \mathbf{e}_{32} \\ &= \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \\ &= 0 + 1 + 0 = 1 \end{aligned}$$

The Euclidean and Hamming distances are given as

$$\delta(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{2(d-s)} = \sqrt{2 \cdot 2} = \sqrt{4} = 2$$
$$\delta_H(\mathbf{x}_1, \mathbf{x}_2) = d - s = 3 - 1 = 2$$

The cosine and Jaccard similarity are given as

$$\cos\theta = \frac{s}{d} = \frac{1}{3} = 0.333$$
$$J(\mathbf{x}_1, \mathbf{x}_2) = \frac{s}{2d-s} = \frac{1}{5} = 0.2$$

3.5 DISCRETIZATION

Discretization, also called *binning*, converts numeric attributes into categorical ones. It is usually applied for data mining methods that cannot handle numeric attributes. It can also help in reducing the number of values for an attribute, especially if there is noise in the numeric measurements; discretization allows one to ignore small and irrelevant differences in the values.

Formally, given a numeric attribute X, and a random sample $\{x_i\}_{i=1}^n$ of size n drawn from X, the discretization task is to divide the value range of X into k consecutive intervals, also called *bins*, by finding $k-1$ boundary values $v_1, v_2, \ldots, v_{k-1}$ that yield the k intervals:

$$[x_{\min}, v_1],\ (v_1, v_2],\ \ldots,\ (v_{k-1}, x_{\max}]$$

where the extremes of the range of X are given as

$$x_{\min} = \min_i\{x_i\} \qquad\qquad x_{\max} = \max_i\{x_i\}$$

The resulting k intervals or bins, which span the entire range of X, are usually mapped to symbolic values that comprise the domain for the new categorical attribute X.

Equal-Width Intervals

The simplest binning approach is to partition the range of X into k *equal-width* intervals. The interval width is simply the range of X divided by k:

$$w = \frac{x_{\max} - x_{\min}}{k}$$

Thus, the ith interval boundary is given as

$$v_i = x_{\min} + i\,w,\ \text{for } i = 1, \ldots, k-1$$

Equal-Frequency Intervals

In *equal-frequency* binning we divide the range of X into intervals that contain (approximately) equal number of points; equal frequency may not be possible due to repeated values. The intervals can be computed from the empirical quantile or

inverse cumulative distribution function $\hat{F}^{-1}(q)$ for X [Eq. (2.2)]. Recall that $\hat{F}^{-1}(q) = \min\{x \mid P(X \leq x) \geq q\}$, for $q \in [0,1]$. In particular, we require that each interval contain $1/k$ of the probability mass; therefore, the interval boundaries are given as follows:

$$v_i = \hat{F}^{-1}(i/k) \text{ for } i = 1, \ldots, k-1$$

Example 3.14. Consider the `sepal length` attribute in the Iris dataset. Its minimum and maximum values are

$$x_{\min} = 4.3 \qquad x_{\max} = 7.9$$

We discretize it into $k = 4$ bins using equal-width binning. The width of an interval is given as

$$w = \frac{7.9 - 4.3}{4} = \frac{3.6}{4} = 0.9$$

and therefore the interval boundaries are

$$v_1 = 4.3 + 0.9 = 5.2 \qquad v_2 = 4.3 + 2 \cdot 0.9 = 6.1 \qquad v_3 = 4.3 + 3 \cdot 0.9 = 7.0$$

The four resulting bins for `sepal length` are shown in Table 3.1, which also shows the number of points n_i in each bin, which are not balanced among the bins.

For equal-frequency discretization, consider the empirical inverse cumulative distribution function (CDF) for `sepal length` shown in Figure 3.5. With $k = 4$ bins, the bin boundaries are the quartile values (which are shown as dashed lines):

$$v_1 = \hat{F}^{-1}(0.25) = 5.1 \qquad v_2 = \hat{F}^{-1}(0.50) = 5.8 \qquad v_3 = \hat{F}^{-1}(0.75) = 6.4$$

The resulting intervals are shown in Table 3.8. We can see that although the interval widths vary, they contain a more balanced number of points. We do not get identical

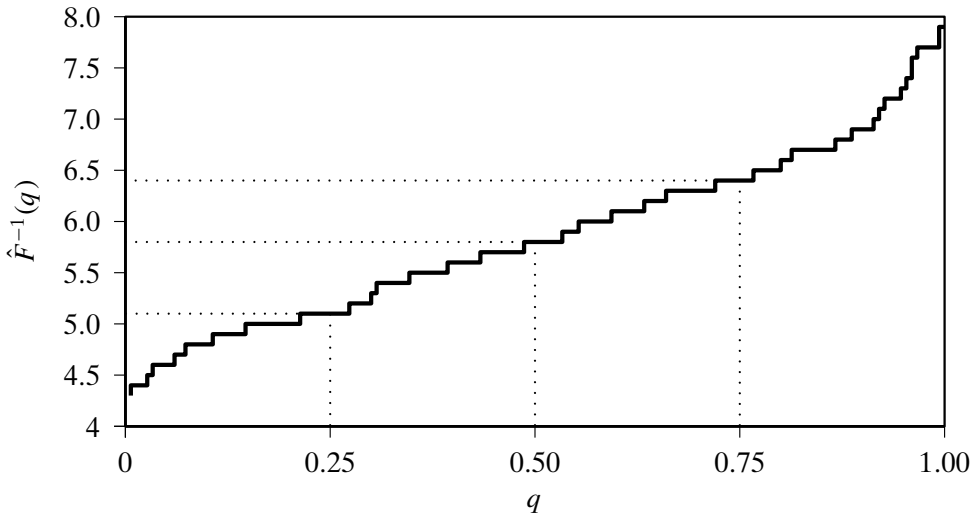


Figure 3.5. Empirical inverse CDF: `sepal length`.

Table 3.8. Equal-frequency discretization: `sepal length`

Bin	Width	Count
[4.3, 5.1]	0.8	$n_1 = 41$
(5.1, 5.8]	0.7	$n_2 = 39$
(5.8, 6.4]	0.6	$n_3 = 35$
(6.4, 7.9]	1.5	$n_4 = 35$

counts for all the bins because many values are repeated; for instance, there are nine points with value 5.1 and there are seven points with value 5.8.

3.6 FURTHER READING

For a comprehensive introduction to categorical data analysis see Agresti (2012). Some aspects also appear in Wasserman (2004). For an entropy-based supervised discretization method that takes the class attribute into account see Fayyad and Irani (1993).

Agresti, A. (2012). *Categorical Data Analysis*, 3rd ed. Hoboken, NJ: John Wiley & Sons.

Fayyad, U. M. and Irani, K. B. (1993). Multi-interval Discretization of Continuous-valued Attributes for Classification Learning. *In Proceedings of the 13th International Joint Conference on Artificial Intelligence*. Morgan-Kaufmann, pp. 1022–1027.

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

3.7 EXERCISES

Q1. Show that for categorical points, the cosine similarity between any two vectors in lies in the range $\cos\theta \in [0, 1]$, and consequently $\theta \in [0^\circ, 90^\circ]$.

Q2. Prove that $E[(\mathbf{X}_1 - \boldsymbol{\mu}_1)(\mathbf{X}_2 - \boldsymbol{\mu}_2)^T] = E[\mathbf{X}_1\mathbf{X}_2^T] - E[\mathbf{X}_1]E[\mathbf{X}_2]^T$.

Table 3.9. Contingency table for Q3

	$Z = f$		$Z = g$	
	$Y = d$	$Y = e$	$Y = d$	$Y = e$
$X = a$	5	10	10	5
$X = b$	15	5	5	20
$X = c$	20	10	25	10

Table 3.10. χ^2 Critical values for different *p-values* for different degrees of freedom (q): For example, for $q = 5$ degrees of freedom, the critical value of $\chi^2 = 11.070$ has *p-value* $= 0.05$.

q	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548

Q3. Consider the 3-way contingency table for attributes X, Y, Z shown in Table 3.9. Compute the χ^2 metric for the correlation between Y and Z. Are they dependent or independent at the 95% confidence level? See Table 3.10 for χ^2 values.

Q4. Consider the "mixed" data given in Table 3.11. Here X_1 is a numeric attribute and X_2 is a categorical one. Assume that the domain of X_2 is given as $dom(X_2) = \{a, b\}$. Answer the following questions.

(a) What is the mean vector for this dataset?

(b) What is the covariance matrix?

Q5. In Table 3.11, assuming that X_1 is discretized into three bins, as follows:

$$c_1 = (-2, -0.5]$$
$$c_2 = (-0.5, 0.5]$$
$$c_3 = (0.5, 2]$$

Answer the following questions:

(a) Construct the contingency table between the discretized X_1 and X_2 attributes. Include the marginal counts.

(b) Compute the χ^2 statistic between them.

(c) Determine whether they are dependent or not at the 5% significance level. Use the χ^2 critical values from Table 3.10.

Table 3.11. Dataset for Q4 and Q5

X_1	X_2
0.3	a
−0.3	b
0.44	a
−0.60	a
0.40	a
1.20	b
−0.12	a
−1.60	b
1.60	b
−1.32	a

CHAPTER 4 Graph Data

The traditional paradigm in data analysis typically assumes that each data instance is independent of another. However, often data instances may be connected or linked to other instances via various types of relationships. The instances themselves may be described by various attributes. What emerges is a network or graph of instances (or nodes), connected by links (or edges). Both the nodes and edges in the graph may have several attributes that may be numerical or categorical, or even more complex (e.g., time series data). Increasingly, today's massive data is in the form of such graphs or networks. Examples include the World Wide Web (with its Web pages and hyperlinks), social networks (wikis, blogs, tweets, and other social media data), semantic networks (ontologies), biological networks (protein interactions, gene regulation networks, metabolic pathways), citation networks for scientific literature, and so on. In this chapter we look at the analysis of the link structure in graphs that arise from these kinds of networks. We will study basic topological properties as well as models that give rise to such graphs.

4.1 GRAPH CONCEPTS

Graphs

Formally, a *graph* $G = (V, E)$ is a mathematical structure consisting of a finite nonempty set V of *vertices* or *nodes*, and a set $E \subseteq V \times V$ of *edges* consisting of *unordered* pairs of vertices. An edge from a node to itself, (v_i, v_i), is called a *loop*. An undirected graph without loops is called a *simple graph*. Unless mentioned explicitly, we will consider a graph to be simple. An edge $e = (v_i, v_j)$ between v_i and v_j is said to be *incident with* nodes v_i and v_j; in this case we also say that v_i and v_j are *adjacent* to one another, and that they are *neighbors*. The number of nodes in the graph G, given as $|V| = n$, is called the *order* of the graph, and the number of edges in the graph, given as $|E| = m$, is called the *size* of G.

A *directed graph* or *digraph* has an edge set E consisting of *ordered* pairs of vertices. A directed edge (v_i, v_j) is also called an *arc*, and is said to be *from* v_i *to* v_j. We also say that v_i is the *tail* and v_j the *head* of the arc.