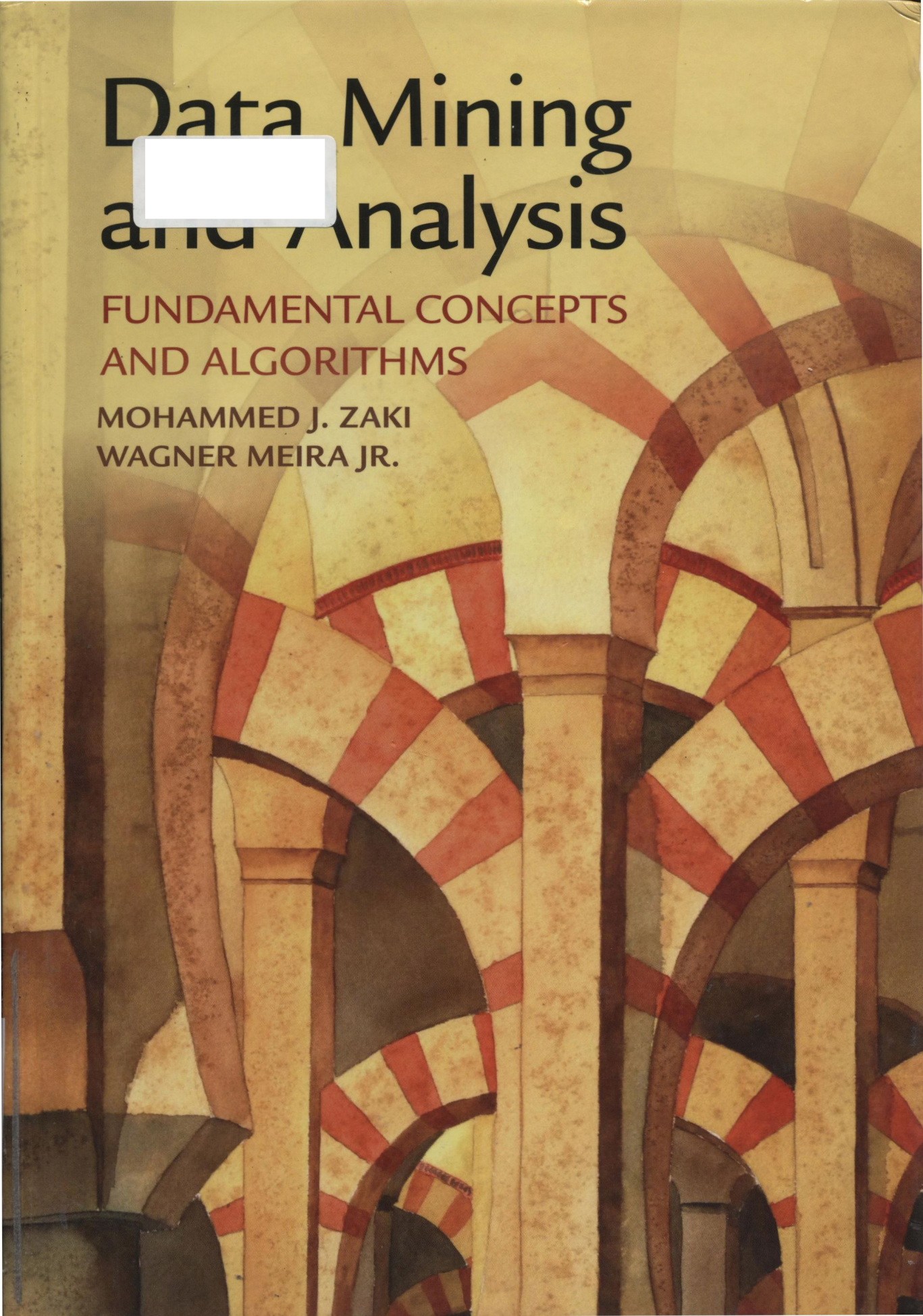


Data Mining and Analysis

FUNDAMENTAL CONCEPTS
AND ALGORITHMS

MOHAMMED J. ZAKI
WAGNER MEIRA JR.



DATA MINING AND ANALYSIS

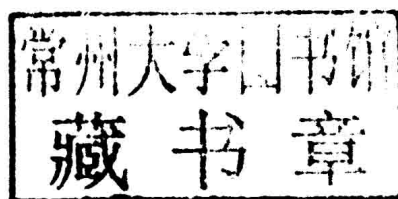
Fundamental Concepts and Algorithms

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

© Mohammed J. Zaki and Wagner Meira Jr. 2014

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

First 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

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

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.

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.

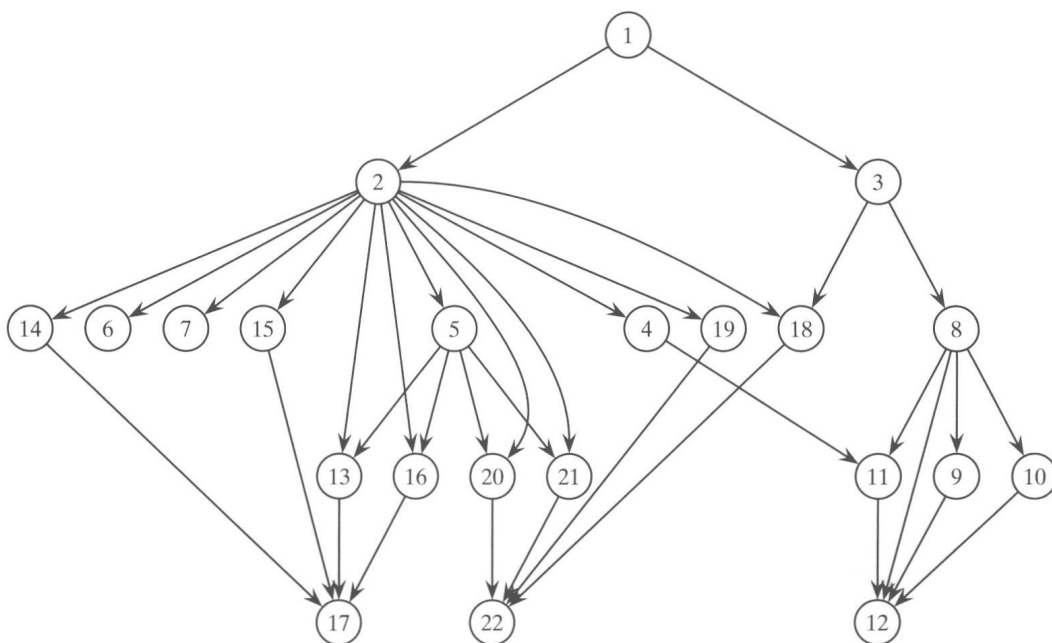


Figure 0.1. Chapter dependencies

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 Dunder, 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.

Contents

Preface	<i>page</i>	ix
1	Data Mining and Analysis	1
1.1	Data Matrix	1
1.2	Attributes	3
1.3	Data: Algebraic and Geometric View	4
1.4	Data: Probabilistic View	14
1.5	Data Mining	25
1.6	Further Reading	30
1.7	Exercises	30
 PART ONE: DATA ANALYSIS FOUNDATIONS		
2	Numeric Attributes	33
2.1	Univariate Analysis	33
2.2	Bivariate Analysis	42
2.3	Multivariate Analysis	48
2.4	Data Normalization	52
2.5	Normal Distribution	54
2.6	Further Reading	60
2.7	Exercises	60
3	Categorical Attributes	63
3.1	Univariate Analysis	63
3.2	Bivariate Analysis	72
3.3	Multivariate Analysis	82
3.4	Distance and Angle	87
3.5	Discretization	89
3.6	Further Reading	91
3.7	Exercises	91
4	Graph Data	93
4.1	Graph Concepts	93
4.2	Topological Attributes	97

4.3	Centrality Analysis	102
4.4	Graph Models	112
4.5	Further Reading	132
4.6	Exercises	132
5	Kernel Methods	134
5.1	Kernel Matrix	138
5.2	Vector Kernels	144
5.3	Basic Kernel Operations in Feature Space	148
5.4	Kernels for Complex Objects	154
5.5	Further Reading	161
5.6	Exercises	161
6	High-dimensional Data	163
6.1	High-dimensional Objects	163
6.2	High-dimensional Volumes	165
6.3	Hypersphere Inscribed within Hypercube	168
6.4	Volume of Thin Hypersphere Shell	169
6.5	Diagonals in Hyperspace	171
6.6	Density of the Multivariate Normal	172
6.7	Appendix: Derivation of Hypersphere Volume	175
6.8	Further Reading	180
6.9	Exercises	180
7	Dimensionality Reduction	183
7.1	Background	183
7.2	Principal Component Analysis	187
7.3	Kernel Principal Component Analysis	202
7.4	Singular Value Decomposition	208
7.5	Further Reading	213
7.6	Exercises	214
PART TWO: FREQUENT PATTERN MINING		
8	Itemset Mining	217
8.1	Frequent Itemsets and Association Rules	217
8.2	Itemset Mining Algorithms	221
8.3	Generating Association Rules	234
8.4	Further Reading	236
8.5	Exercises	237
9	Summarizing Itemsets	242
9.1	Maximal and Closed Frequent Itemsets	242
9.2	Mining Maximal Frequent Itemsets: GenMax Algorithm	245
9.3	Mining Closed Frequent Itemsets: Charm Algorithm	248
9.4	Nonderivable Itemsets	250
9.5	Further Reading	256
9.6	Exercises	256

10	Sequence Mining	259
10.1	Frequent Sequences	259
10.2	Mining Frequent Sequences	260
10.3	Substring Mining via Suffix Trees	267
10.4	Further Reading	277
10.5	Exercises	277
11	Graph Pattern Mining	280
11.1	Isomorphism and Support	280
11.2	Candidate Generation	284
11.3	The gSpan Algorithm	288
11.4	Further Reading	296
11.5	Exercises	297
12	Pattern and Rule Assessment	301
12.1	Rule and Pattern Assessment Measures	301
12.2	Significance Testing and Confidence Intervals	316
12.3	Further Reading	328
12.4	Exercises	328
PART THREE: CLUSTERING		
13	Representative-based Clustering	333
13.1	K-means Algorithm	333
13.2	Kernel K-means	338
13.3	Expectation-Maximization Clustering	342
13.4	Further Reading	360
13.5	Exercises	361
14	Hierarchical Clustering	364
14.1	Preliminaries	364
14.2	Agglomerative Hierarchical Clustering	366
14.3	Further Reading	372
14.4	Exercises and Projects	373
15	Density-based Clustering	375
15.1	The DBSCAN Algorithm	375
15.2	Kernel Density Estimation	379
15.3	Density-based Clustering: DENCLUE	385
15.4	Further Reading	390
15.5	Exercises	391
16	Spectral and Graph Clustering	394
16.1	Graphs and Matrices	394
16.2	Clustering as Graph Cuts	401
16.3	Markov Clustering	416
16.4	Further Reading	422
16.5	Exercises	423

17	Clustering Validation	425
17.1	External Measures	425
17.2	Internal Measures	440
17.3	Relative Measures	448
17.4	Further Reading	461
17.5	Exercises	462
 PART FOUR: CLASSIFICATION		
18	Probabilistic Classification	467
18.1	Bayes Classifier	467
18.2	Naive Bayes Classifier	473
18.3	K Nearest Neighbors Classifier	477
18.4	Further Reading	479
18.5	Exercises	479
 19	 Decision Tree Classifier	 481
19.1	Decision Trees	483
19.2	Decision Tree Algorithm	485
19.3	Further Reading	496
19.4	Exercises	496
 20	 Linear Discriminant Analysis	 498
20.1	Optimal Linear Discriminant	498
20.2	Kernel Discriminant Analysis	505
20.3	Further Reading	511
20.4	Exercises	512
 21	 Support Vector Machines	 514
21.1	Support Vectors and Margins	514
21.2	SVM: Linear and Separable Case	520
21.3	Soft Margin SVM: Linear and Nonseparable Case	524
21.4	Kernel SVM: Nonlinear Case	530
21.5	SVM Training Algorithms	534
21.6	Further Reading	545
21.7	Exercises	546
 22	 Classification Assessment	 548
22.1	Classification Performance Measures	548
22.2	Classifier Evaluation	562
22.3	Bias-Variance Decomposition	572
22.4	Further Reading	581
22.5	Exercises	582
	 Index	 585

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 i th row, which is a d -tuple given as

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

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

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

Depending on the application domain, rows may also be referred to as *entities*, *instances*, *examples*, *records*, *transactions*, *objects*, *points*, *feature-vectors*, *tuples*, and so on. Likewise, columns may also be called *attributes*, *properties*, *features*, *dimensions*, *variables*, *fields*, and so on. The number of instances n is referred to as the *size* of

Table 1.1. Extract from the Iris dataset

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

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

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

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

1.2 ATTRIBUTES

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

Numeric Attributes

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

- *Interval-scaled*: For these kinds of attributes only differences (addition or subtraction) make sense. For example, attribute temperature measured in °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

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

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

Categorical attributes may be of two types:

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

1.3 DATA: ALGEBRAIC AND GEOMETRIC VIEW

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

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

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

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

where T is the *matrix transpose* operator.

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

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

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

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

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

Example 1.2. Consider the Iris data in Table 1.1. If we *project* the entire data onto the first two attributes, then each row can be considered as a point or a vector in 2-dimensional space. For example, the projection of the 5-tuple $\mathbf{x}_1 = (5.9, 3.0, 4.2, 1.5, \text{Iris-versicolor})$ on the first two attributes is shown in Figure 1.1a. 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 :

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