

DATA MINING AND MACHINE LEARNING

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

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

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

DATA MINING AND MACHINE LEARNING

Fundamental Concepts and Algorithms

MOHAMMED J. ZAKI

Rensselaer Polytechnic Institute

WAGNER MEIRA, JR.

Universidade Federal de Minas Gerais



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India
79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

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

www.cambridge.org

Information on this title: www.cambridge.org/9781108473989

DOI: 10.1017/9781108564175

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

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

First edition published 2014, second edition published 2020

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

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

Library of Congress Cataloging in Publication Data

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

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

Other titles: Data mining and analysis

Description: Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2020. | Revised edition of: Data mining and analysis. 2014. | Includes bibliographical references and index.

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

Subjects: LCSH: Data mining.

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

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

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

ISBN 978-1-108-47398-9 Hardback

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

Contents

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

vi	Contents
4.4	Graph Models 111
4.5	Further Reading 131
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 149
5.4	Kernels for Complex Objects 155
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 167
6.3	Hypersphere Inscribed within Hypercube 170
6.4	Volume of Thin Hypersphere Shell 171
6.5	Diagonals in Hyperspace 172
6.6	Density of the Multivariate Normal 173
6.7	Appendix: Derivation of Hypersphere Volume 177
6.8	Further Reading 181
6.9	Exercises 181
7	Dimensionality Reduction 184
7.1	Background 184
7.2	Principal Component Analysis 188
7.3	Kernel Principal Component Analysis 203
7.4	Singular Value Decomposition 210
7.5	Further Reading 215
7.6	Exercises 215
	PART TWO: FREQUENT PATTERN MINING 217
8	Itemset Mining 219
8.1	Frequent Itemsets and Association Rules 219
8.2	Itemset Mining Algorithms 223
8.3	Generating Association Rules 237
8.4	Further Reading 238
8.5	Exercises 239
9	Summarizing Itemsets 244
9.1	Maximal and Closed Frequent Itemsets 244
9.2	Mining Maximal Frequent Itemsets: GenMax Algorithm 247
9.3	Mining Closed Frequent Itemsets: Charm Algorithm 250
9.4	Nonderivable Itemsets 252
9.5	Further Reading 258
9.6	Exercises 258

Contents	vii
10 Sequence Mining	261
10.1 Frequent Sequences	261
10.2 Mining Frequent Sequences	262
10.3 Substring Mining via Suffix Trees	269
10.4 Further Reading	279
10.5 Exercises	279
11 Graph Pattern Mining	282
11.1 Isomorphism and Support	282
11.2 Candidate Generation	286
11.3 The gSpan Algorithm	290
11.4 Further Reading	298
11.5 Exercises	299
12 Pattern and Rule Assessment	303
12.1 Rule and Pattern Assessment Measures	303
12.2 Significance Testing and Confidence Intervals	318
12.3 Further Reading	330
12.4 Exercises	330
PART THREE: CLUSTERING	332
13 Representative-based Clustering	334
13.1 K-means Algorithm	334
13.2 Kernel K-means	339
13.3 Expectation-Maximization Clustering	343
13.4 Further Reading	360
13.5 Exercises	361
14 Hierarchical Clustering	364
14.1 Preliminaries	364
14.2 Agglomerative Hierarchical Clustering	366
14.3 Further Reading	372
14.4 Exercises	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	417
16.4 Further Reading	422
16.5 Exercises	424

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

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

Preface

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

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

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

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

`http://dataminingbook.info`

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

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

Mohammed J. Zaki and Wagner Meira, Jr.