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

First edition published 2014, second edition published 2020

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

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

Library of Congress Cataloging in Publication Data
Names: Zaki, Mohammed J., 1971- author. | Meira, Wagner, 1967- author.
Title: Data mining and machine learning : fundamental concepts and algorithms / Mohammed J. Zaki, Wagner Meira, Jr.
Other titles: Data mining and analysis
Description: Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2020. | Revised edition of: Data mining and analysis. 2014. | Includes bibliographical references and index.
Identifiers: LCCN 2019037293 (print) | LCCN 2019037294 (ebook) | ISBN 9781108473989 (hardback) | ISBN 9781108564175 (epub)
Subjects: LCSH: Data mining.
Classification: LCC QA76.9.D343 Z36 2020 (print) | LCC QA76.9.D343 (ebook) | DDC 006.3/12–dc23
LC record available at https://lccn.loc.gov/2019037293
LC ebook record available at https://lccn.loc.gov/2019037294

ISBN 978-1-108-47398-9 Hardback

Contents

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.