Charu C. Aggarwal

Data Mining

The Textbook

About the Book

This textbook explores the different aspects of data mining from the fundamentals to the complex data types and their applications, capturing the wide diversity of problem domains for data mining issues. It goes beyond the traditional focus on data mining problems to introduce advanced data types such as text, time series, discrete sequences, spatial data, graph data, and social networks. Until now, no single book has addressed all these topics in a comprehensive and integrated way. The chapters of this book fall into one of three categories:

- **Fundamental chapters:** Data mining has four main problems, which correspond to clustering, classification, association pattern mining, and outlier analysis. These chapters comprehensively discuss a wide variety of methods for these problems.
- **Domain chapters:** These chapters discuss the specific methods used for different domains of data such as text data, time-series data, sequence data, graph data, and spatial data.
- **Application chapters:** These chapters study important applications such as stream mining, Web mining, ranking, recommendations, social networks, and privacy preservation. The domain chapters also have an applied flavor.

Appropriate for both introductory and advanced data mining courses, *Data Mining: The Textbook* balances mathematical details and intuition. It contains the necessary mathematical details for professors and researchers, but it is presented in a simple and intuitive style to improve accessibility for students and industrial practitioners (including those with a limited mathematical background). Numerous illustrations, examples, and exercises are included, with an emphasis on semantically interpretable examples.

About the Author

Charu C. Aggarwal is a Distinguished Research Staff Member (DRSM) at the IBM T.J. Watson Research Center in Yorktown Heights, New York. He completed his B.S. from IIT Kanpur in 1993 and his Ph.D. from the Massachusetts Institute of Technology in 1996. He has published more than 250 papers in refereed conferences and journals and authored over 80 patents. He is the author or editor of 14 books, including the first comprehensive book on outlier analysis, which is written from a computer science point of view. Because of the commercial value of his patents, he has thrice been designated a Master Inventor at IBM. He has won multiple awards, chaired conferences, and currently serves on several journal editorial boards. He is a fellow of the ACM and the IEEE, for "contributions to knowledge discovery and data mining algorithms."

Computer Science

ISBN 978-3-319-14141-1

▶ springer.com

Aggarwal

Data Mining

Charu C. Aggarwal

Data Mining

The Textbook

Springer

Data Mining: The Textbook

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, New York

March 8, 2015

To my wife Lata,
and my daughter Sayani

Contents

Preface

"*Data is the new oil.*"– Clive Humby

The field of data mining has seen rapid strides over the past two decades, especially from the perspective of the computer science community. While data analysis has been studied extensively in the conventional field of probability and statistics, *data mining* is a term coined by the computer science-oriented community. For computer scientists, issues such as scalability, usability, and computational implementation are extremely important.

The emergence of data science as a discipline requires the development of a book that goes beyond the traditional focus of books on only the fundamental data mining courses. Recent years have seen the emergence of the job description of "data scientists," who try to glean knowledge from vast amounts of data. In typical applications, the data types are so heterogeneous and diverse that the fundamental methods discussed for a multidimensional data type may not be effective. Therefore, more emphasis needs to be placed on the different data types and the applications which arise in the context of these different data types. A comprehensive data mining book must explore the different aspects of data mining, starting from the fundamentals, and then explore the complex data types, and their relationships with the fundamental techniques. While fundamental techniques form an excellent basis for the further study of data mining, they do not provide a complete picture of the true complexity of data analysis. This book studies these advanced topics without compromising the presentation of fundamental methods. Therefore, this book may be used for both introductory and advanced data mining courses. Until now, no single book has addressed all these topics in a comprehensive and integrated way.

The textbook assumes a basic knowledge of probability, statistics, and linear algebra, which is taught in most undergraduate curricula of science and engineering disciplines. Therefore, the book can also be used by industrial practitioners, who have a working knowledge of these basic skills. While stronger mathematical background is helpful for the more advanced chapters, it is not a pre-requisite. Special chapters are also devoted to different aspects of data mining, such as the text data, time-series data, discrete sequences, and graphs. This kind of specialized treatment is intended to capture the wide diversity of problem domains in which a data mining problem might arise.

The chapters of this book fall into one of three categories:

- **The fundamental chapters:** Data mining has four main "super-problems," which correspond to clustering, classification, association pattern mining, and outlier analysis. These problems are so important because they are used repeatedly as building blocks in the context of a wide variety of data mining applications. As a result, a large

amount of emphasis has been placed by data mining researchers and practitioners to design effective and efficient methods for these problems. These chapters comprehensively discuss the vast diversity of methods used by the data mining community in the context of these super-problems.

- **Domain chapters:** These chapters discuss the specific methods used for different *domains* of data such as text data, time-series data, sequence data, graph data, and spatial data. Many of these chapters can also be considered application chapters, because they explore the specific characteristics of the problem in a particular domain.
- **Application chapters:** Advancements in hardware technology and software platforms, have lead to a number of data-intensive applications such as streaming systems, Web mining, social networks, and privacy-preservation. These topics are studied in detail in these chapters. The domain chapters are also focused on many different kinds of applications that arise in the context of those data types.

Suggestions for the Instructor

The book was specifically written to enable the teaching of both the basic data mining and advanced data mining courses from a single book. It can be used to offer various types of data mining courses with different emphases. Specifically, the courses that could be offered with various chapters are as follows:

- **Basic data mining course and fundamentals:** The basic data mining course should focus on the fundamentals of data mining. Chapters 1, 2, 3, 4, 6, 8, and 10 can be covered. In fact, the material in these chapters is more than what is possible to teach in a single course. Therefore, instructors may need to select topics of their interest from these chapters. Some portions of Chapters 5, 7, 9, and 11 can also be covered, although these chapters are really meant for an advanced course.
- **Advanced course (fundamentals):** Such a course would cover advanced topics on the fundamentals of data mining and assume that the student is already familiar with Chapters 1 through 3, and parts of Chapters 4, 6, 8, and 10. The course can then focus on Chapters 5, 7, 9, and 11. Topics such as ensemble analysis are useful for the advanced course. Furthermore, some topics from Chapters 4, 6, 8, and 10, which were not covered in the basic course, can be used. In addition, Chapter 20 on privacy can be offered.
- **Advanced course (data types):** Advanced topics such as text mining, time series, sequences, graphs, and spatial data may be covered. The material should focus on Chapters 13, 14, 15, 16, and 17. Some parts of Chapter 19 (e.g., graph clustering) and Chapter 12 (data streaming) can also be used.
- **Advanced course (applications):** An application course overlaps with a data type course but has a different focus. For example, the focus in an application-centered course would be more on the modeling aspect than the algorithmic aspect. Therefore, the same materials in Chapters 13, 14, 15, 16, and 17 can be used while skipping specific details of algorithms. With less focus on specific algorithms, these chapters can be covered fairly quickly. The remaining time should be allocated to three very important chapters on data streams (Chapter 12), Web mining (Chapter 18), and social network analysis (Chapter 19).

The book is written in a simple style to make it accessible to undergraduate students and industrial practitioners with a limited mathematical background. Thus, the book will serve

both as an introductory text and as an advanced text for students, industrial practitioners, and researchers.

Throughout this book, a vector or a multidimensional data point (including categorical attributes), is annotated with a bar, such as $\overline{X}$ or $\overline{y}$. A vector or multidimensional point may be denoted by either small letters or capital letters, as long as it has a bar. Vector dot products are denoted by centered dots, such as $\overline{X} \cdot \overline{Y}$. A matrix is denoted in capital letters without a bar, such as R. Throughout the book, the $n \times d$ data matrix is denoted by D, with n points and d dimensions. The individual data points in D are therefore d-dimensional row vectors. On the other hand, vectors with one component for each data point are usually n-dimensional column vectors. An example is the n-dimensional column vector $\overline{y}$ of class variables of n data points.