Principles of Data Mining

Adaptive Computation and Machine Learning

Thomas Dietterich, Editor
Christopher Bishop, David Heckerman, Michael Jordan, and Michael Kearns, Associate Editors

Bioinformatics: The Machine Learning Approach, Pierre Baldi and Søren Brunak
Reinforcement Learning: An Introduction, Richard S. Sutton and Andrew G. Barto
Graphical Models for Machine Learning, Brendan J. Frey
Learning in Graphical Models, Michael I. Jordan
Causation, Prediction, and Search, second edition, Peter Spirtes, Clark Glymour, and Richard Scheines
Principles of Data Mining, David J. Hand, Heikki Mannila, and Padhraic Smyth

Principles of Data Mining

David Hand
Heikki Mannila
Padhraic Smyth

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

This book was typeset in Palatino by the authors and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Hand, D. J.
Principles of data mining / David Hand, Heikki Mannila, Padhraic Smyth.
p. cm.—(Adaptive computation and machine learning)
Includes bibliographical references and index.
ISBN 0-262-08290-X (hc. : alk. paper)
1. Data Mining. I. Mannila, Heikki. II. Smyth, Padhraic. III. Title. IV. Series.

QA76.9.D343 H38 2001
006.3—dc21 2001032620

To Crista, Aidan, and Cian

To Paula and Elsa

To Shelley, Rachel, and Emily

Brief Contents

Contents

List of Tables

List of Figures

Series Foreword

The rapid growth and integration of databases provides scientists, engineers, and business people with a vast new resource that can be analyzed to make scientific discoveries, optimize industrial systems, and uncover financially valuable patterns. To undertake these large data analysis projects, researchers and practitioners have adopted established algorithms from statistics, machine learning, neural networks, and databases and have also developed new methods targeted at large data mining problems. *Principles of Data Mining* by David Hand, Heikki Mannila, and Padhraic Smyth provides practioners and students with an introduction to the wide range of algorithms and methodologies in this exciting area. The interdisciplinary nature of the field is matched by these three authors, whose expertise spans statistics, databases, and computer science. The result is a book that not only provides the technical details and the mathematical principles underlying data mining methods, but also provides a valuable perspective on the entire enterprise.

Data mining is one component of the exciting area of machine learning and adaptive computation. The goal of building computer systems that can adapt to their envirionments and learn from their experience has attracted researchers from many fields, including computer science, engineering, mathematics, physics, neuroscience, and cognitive science. Out of this research has come a wide variety of learning techniques that have the potential to transform many scientific and industrial fields. Several research communities have converged on a common set of issues surrounding supervised, unsupervised, and reinforcement learning problems. The MIT Press series on Adaptive Computation and Machine Learning seeks to unify the many diverse strands of machine learning research and to foster high quality research and innovative applications.

Thomas Dietterich

Preface

The science of extracting useful information from large data sets or databases is known as data mining. It is a new discipline, lying at the intersection of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and other areas. All of these are concerned with certain aspects of data analysis, so they have much in common—but each also has its own distinct flavor, emphasizing particular problems and types of solution.

Because data mining encompasses a wide variety of topics in computer science and statistics it is impossible to cover all the potentially relevant material in a single text. Given this, we have focused on the topics that we believe are the most fundamental.

From a teaching viewpoint the text is intended for undergraduate students at the senior (final year) level, or first or second-year graduate level, who wish to learn about the basic principles of data mining. The text should also be of value to researchers and practitioners who are interested in gaining a better understanding of data mining methods and techniques. A familiarity with the very basic concepts in probability, calculus, linear algebra, and optimization is assumed—in other words, an undergraduate background in any quantitative discipline such as engineering, computer science, mathematics, economics, etc., should provide a good background for reading and understanding this text.

There are already many other books on data mining on the market. Many are targeted at the business community directly and emphasize specific methods and algorithms (such as decision tree classifiers) rather than general principles (such as parameter estimation or computational complexity). These texts are quite useful in providing general context and case studies, but have limitations in a classroom setting, since the underlying foundational principles are often missing. There are other texts on data mining that have a more academic flavor, but to date these have been written largely from a computer

science viewpoint, specifically from either a database viewpoint (Han and Kamber, 2000), or from a machine learning viewpoint (Witten and Franke, 2000).

This text has a different bias. We have attempted to provide a foundational view of data mining. Rather than discuss specific data mining applications at length (such as, say, collaborative filtering, credit scoring, and fraud detection), we have instead focused on the underlying theory and algorithms that provide the "glue" for such applications. This is not to say that we do not pay attention to the applications. Data mining is fundamentally an applied discipline, and with this in mind we make frequent references to case studies and specific applications where the basic theory can (or has been) applied.

In our view a mastery of data mining requires an understanding of both statistical and computational issues. This requirement to master two different areas of expertise presents quite a challenge for student and teacher alike. For the typical computer scientist, the statistics literature is relatively impenetrable: a litany of jargon, implicit assumptions, asymptotic arguments, and lack of details on how the theoretical and mathematical concepts are actually realized in the form of a data analysis algorithm. The situation is effectively reversed for statisticians: the computer science literature on machine learning and data mining is replete with discussions of algorithms, pseudocode, computational efficiency, and so forth, often with little reference to an underlying model or inference procedure. An important point is that *both* approaches are nonetheless essential when dealing with large data sets. An understanding of both the "mathematical modeling" view, and the "computational algorithm" view are essential to properly grasp the complexities of data mining.

In this text we make an attempt to bridge these two worlds and to explicitly link the notion of statistical modeling (with attendant assumptions, mathematics, and notation) with the "real world" of actual computational methods and algorithms.

With this in mind, we have structured the text in a somewhat unusual manner. We begin with a discussion of the very basic principles of modeling and inference, then introduce a systematic framework that connects models to data via computational methods and algorithms, and finally instantiate these ideas in the context of specific techniques such as classification and regression. Thus, the text can be divided into three general sections:

1. **Fundamentals**: Chapters 1 through 4 focus on the fundamental aspects of data and data analysis: introduction to data mining (chapter 1), mea-

surement (chapter 2), summarizing and visualizing data (chapter 3), and uncertainty and inference (chapter 4).

2. **Data Mining Components**: Chapters 5 through 8 focus on what we term the "components" of data mining algorithms: these are the building blocks that can be used to systematically create and analyze data mining algorithms. In chapter 5 we discuss this systematic approach to algorithm analysis, and argue that this "component-wise" view can provide a useful systematic perspective on what is often a very confusing landscape of data analysis algorithms to the novice student of the topic. In this context, we then delve into broad discussions of each component: model representations in chapter 6, score functions for fitting the models to data in chapter 7, and optimization and search techniques in chapter 8. (Discussion of data management is deferred until chapter 12.)

3. **Data Mining Tasks and Algorithms**: Having discussed the fundamental components in the first 8 chapters of the text, the remainder of the chapters (from 9 through 14) are then devoted to specific data mining tasks and the algorithms used to address them. We organize the basic tasks into density estimation and clustering (chapter 9), classification (chapter 10), regression (chapter 11), pattern discovery (chapter 13), and retrieval by content (chapter 14). In each of these chapters we use the framework of the earlier chapters to provide a general context for the discussion of specific algorithms for each task. For example, for classification we ask: what models and representations are plausible and useful? what score functions should we, or can we, use to train a classifier? what optimization and search techniques are necessary? what is the computational complexity of each approach once we implement it as an actual algorithm? Our hope is that this general approach will provide the reader with a "roadmap" to an understanding that data mining algorithms are based on some very general and systematic principles, rather than simply a cornucopia of seemingly unrelated and exotic algorithms.

In terms of using the text for teaching, as mentioned earlier the target audience for the text is students with a quantitative undergraduate background, such as in computer science, engineering, mathematics, the sciences, and more quantitative business-oriented degrees such as economics. From the instructor's viewpoint, how much of the text should be covered in a course will depend on both the length of the course (e.g., 10 weeks versus 15 weeks) and the familiarity of the students with basic concepts in statistics and ma-

chine learning. For example, for a 10-week course with first-year graduate students who have some exposure to basic statistical concepts, the instructor might wish to move quickly through the early chapters: perhaps covering chapters 3, 4, 5, and 7 fairly rapidly; assigning chapters 1, 2, 6, and 8 as background/review reading; and then spending the majority of the 10 weeks covering chapters 9 through 14 in some depth.

Conversely many students and readers of this text may have little or no formal statistical background. It is unfortunate that in many quantitative disciplines (such as computer science) students at both undergraduate and graduate levels often get only a very limited exposure to statistical thinking in many modern degree programs. Since we take a fairly strong statistical view of data mining in this text, our experience in using draft versions of the text in computer science departments has taught us that mastery of the entire text in a 10-week or 15-week course presents quite a challenge to many students, since to fully absorb the material they must master quite a broad range of statistical, mathematical, and algorithmic concepts in chapters 2 through 8. In this light, a less arduous path is often desirable. For example, chapter 11 on regression is probably the most mathematically challenging in the text and can be omitted without affecting understanding of any of the remaining material. Similarly some of the material in chapter 9 (on mixture models for example) could also be omitted, as could the Bayesian estimation framework in chapter 4. In terms of what is essential reading, most of the material in chapters 1 through 5 and in chapters 7, 8, and 12 we consider to be essential for the students to be able to grasp the modeling and algorithmic ideas that come in the later chapters (chapter 6 contains much useful material on the general concepts of modeling but is quite long and could be skipped in the interests of time). The more "task-specific" chapters of 9, 10, 11, 13, and 14 can be chosen in a "menu-based" fashion, i.e., each can be covered somewhat independently of the others (but they do assume that the student has a good working knowledge of the material in chapters 1 through 8).

An additional suggestion for students with limited statistical exposure is to have them review some of the basic concepts in probability and statistics *before* they get to chapter 4 (on uncertainty) in the text. Unless students are comfortable with basic concepts such as conditional probability and expectation, they will have difficulty following chapter 4 and much of what follows in later chapters. We have included a brief appendix on basic probability and definitions of common distributions, but some students will probably want to go back and review their undergraduate texts on probability and statistics before venturing further.

On the other side of the coin, for readers with substantial statistical background (e.g., statistics students or statisticians with an interest in data mining) much of this text will look quite familiar and the statistical reader may be inclined to say "well, this data mining material seems very similar in many ways to a course in applied statistics!" And this is indeed somewhat correct, in that data mining (as we view it) relies very heavily on statistical models and methodologies. However, there are portions of the text that statisticians will likely find quite informative: the overview of chapter 1, the algorithmic viewpoint of chapter 5, the score function viewpoint of chapter 7, and all of chapters 12 through 14 on database principles, pattern finding, and retrieval by content. In addition, we have tried to include in our presentation of many of the traditional statistical concepts (such as classification, clustering, regression, etc.) additional material on algorithmic and computational issues that would not typically be presented in a statistical textbook. These include statements on computational complexity and brief discussions on how the techniques can be used in various data mining applications. Nonetheless, statisticians will find much familiar material in this text. For views of data mining that are more oriented towards computational and data-management issues see, for example, Han and Kamber (2000), and for a business focus see, for example, Berry and Linoff (2000). These texts could well serve as complementary reading in a course environment.

In summary, this book describes tools for data mining, splitting the tools into their component parts, so that their structure and their relationships to each other can be seen. Not only does this give insight into what the tools are designed to achieve, but it also enables the reader to design tools of their own, suited to the particular problems and opportunities facing them. The book also shows how data mining is a process—not something which one does, and then finishes, but an ongoing voyage of discovery, interpretation, and re-investigation. The book is liberally illustrated with real data applications, many arising from the authors' own research and applications work. For didactic reasons, not all of the data sets discussed are large—it is easier to explain what is going on in a "small" data set. Once the idea has been communicated, it can readily be applied in a realistically large context.

Data mining is, above all, an exciting discipline. Certainly, as with any scientific enterprise, much of the effort will be unrewarded (it is a rare and perhaps rather dull undertaking which gives a guaranteed return). But this is more than compensated for by the times when an exciting discovery—a gem or nugget of valuable information—is unearthed. We hope that you as a reader of this text will be inspired to go forth and discover your own gems!

We would like to gratefully acknowledge Christine McLaren for granting permission to use the red blood cell data as an illustrative example in chapters 9 and 10. Padhraic Smyth's work on this text was supported in part by the National Science Foundation under Grant IRI-9703120.

We would also like to thank Niall Adams for help in producing some of the diagrams, Tom Benton for assisting with proof corrections, and Xianping Ge for formatting the references. Naturally, any mistakes which remain are the responsibility of the authors (though each of the three of us reserves the right to blame the other two).

Finally we would each like to thank our respective wives and families for providing excellent encouragement and support throughout the long and seemingly never-ending saga of "the book"!