Journal of Statistical Software

May 2006, Volume 16, Book Review 1. http://www.jstatsoft.org/

Reviewer: Hongquan Xu
University of California, Los Angeles

Discovering Knowledge in Data: An Introduction to Data Mining

Daniel T. Larose
John Wiley & Sons, Hoboken, New Jersey, 2004.
ISBN 0-471-66657-2. 240 pp. USD 69.95 (P).
http://www.dataminingconsultant.com/

This is the first of a planned series of three books on data mining published by Wiley. The other two are *Data Mining Methods and Models*, published in January 2006, and *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*, expected in late 2006 or early 2007. This book covers essential topics on data mining in a simple manner and can serve as a general introduction to data mining for almost everyone, including "managers, CIOs, CEOs, CFOs." This is a book that I am looking for a long time.

This book consists of 11 chapters. The first chapter provides the basics of data mining: what, why, how, as well as the data mining process and common tasks. Five case studies are given to illustrate the data mining process and business applications. The fallacies of data mining Larose discusses, missing from most of the data mining books, are no less important than examples of successful data mining. The next three chapters are essentially introductory statistics, including data preprocessing (Chapter 2), exploratory data analysis (Chapter 3) and statistical approaches to estimation and prediction (Chapter 4). These topics are well understood to statisticians, but might be new to others. Although one may debate whether regression or statistical inference is data mining, Larose made the point clear that statistics is critical for data mining. One example that I like is the discussion of a tiny data set of five transactions (Table 2.1, page 28). Larose articulated well what could go wrong with this tiny data set to illustrate the need of data cleaning and statistical thinking in general.

The next seven chapters "represent the heart of the book" and each chapter describes one specific data mining method or technique: k-nearest neighbor algorithms (Chapter 5), decision trees (Chapter 6), neural networks (Chapter 7), hierarchical and k-means clustering (Chapter 8), Kohonen networks (Chapter 9), association rules (Chapter 10), and model evaluation techniques (Chapter 11). Larose did a great job on providing small sample data sets and walking the reader through some complex algorithms such as classification and regression trees, neural networks and Kohonen networks. In addition, examples of actual large data sets are provided and illustrated using data mining software **Clementine**, **Insightful Miner** or SAS **Enterprise Miner**. The chapter of neural networks discusses momentum learning and sensitivity analysis, two topics that are not seen in many other books.

This book can serve as a textbook for an introductory course in data mining. It is appropriate for undergraduate courses, with or without a prerequisite of an introductory statistics course, and should be subsidized if used as a graduate textbook. The companion Web site contains the following additional material for instructors: answer keys to chapter exercises and hand-on analyses, **PowerPoint** presentations (for the first 10 chapters), data mining projects and data sets, and Web resources and quizzes (yet to be developed). I used this book for Fiat Lux Freshman Seminars at UCLA and liked it very much.

In summary, this is an excellent introductory book on data mining. I recommend it for every one who wants to learn data mining.

Reviewer:

Hongquan Xu
University of California, Los Angeles
Department of Statistics
Los Angeles, CA 90095-1554
United States of America
E-mail: hqxu@stat.ucla.edu
URL: http://www.stat.ucla.edu/~hqxu/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 16, Book Review 1 *Published:* 2006-05-22
May 2006