Data Mining Methods and Models

By Daniel T. Larose, Ph.D.
Director, Data Mining @CCSU

Preface

What is data mining?

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."

-- David Hand, Heikki Mannila & Padhraic Smyth,
Principles of Data Mining, MIT Press, 2001

Data mining is predicted to be "one of the most revolutionary developments of the next decade", according to the online technology magazine *ZDNET News* (February 8, 2001). In fact, the *MIT Technology Review* chose data mining as one of ten emerging technologies that will change the world.

Because data mining represents such an important field, *Wiley Interscience* and Dr. Daniel T. Larose have teamed up to publish a new series on data mining, initially consisting of three volumes. The first volume in this series, *Discovering Knowledge in Data: An Introduction to Data Mining*, appeared in 2005, and introduced the reader to this rapidly growing field of data mining.

The second volume in the series, *Data Mining Methods and Models*, explores the process of data mining from the point of view of *model building* – the development of complex and powerful predictive models that can deliver actionable results for a wide range of business and research problems.

Why is this book needed?

Data Mining Methods and Models continues the thrust of *Discovering Knowledge in Data*, providing the reader with:

- The models and techniques to uncover hidden nuggets of information,
- The insight into how the data mining algorithms really work, and
- The experience of actually performing data mining on large data sets.

"White Box" Approach: Understanding the Underlying Algorithmic and Model Structures.

The best way to avoid costly errors stemming from a blind black-box approach to data mining, is to instead apply a "white-box" methodology, which emphasizes an understanding of the algorithmic and statistical model structures underlying the software.

***Data Mining Methods and Models* applies this white-box approach by:**

- Walking the reader through the various algorithms,
- Providing examples of the operation of the algorithm on actual large data sets,
- Testing the reader's level of understanding of the concepts and algorithms, and
- Providing an opportunity for the reader to do some real data mining on large data sets.

Algorithm Walk-Throughs

Data Mining Methods and Models walks the reader through the operations and nuances of the various algorithms, using small sample data sets, so that the reader gets a true appreciation of what is really going on inside the algorithm. For example, in Chapter 2, *Regression Modeling*, we observe how a single new data value can seriously alter the model results. Also, in Chapter 6, *Genetic Algorithms*, we proceed step-by-step to find the optimal solution using the selection, crossover, and mutation operators.

Applications of the algorithms and models to large data sets.

Data Mining Methods and Models provides examples of the application of the various algorithms and models on actual large data sets. For example, in Chapter 3, *Multiple Regression and Model Building*, we analytically unlock the relationship between nutrition rating and cereal content using a real-world data set. In Chapter 1, *Dimension Reduction Methods*, we apply principal components analysis to real-world census data about California. All data sets are available from the book series website: www.dataminingconsultant.com.

Chapter exercises: Checking to make sure you understand it.

Data Mining Methods and Models includes over 110 chapter exercises, which allow readers to assess their depth of understanding of the material, as well as have a little fun playing with numbers and data. These include *Clarifying the Concept* exercises, which help to clarify some of the more challenging concepts in data mining, and *Working with the Data* exercises, which challenge the reader to apply the particular data mining algorithm to a small data set, and, step-by-step, to arrive at a computationally sound solution. For example, in Chapter 5, *Naïve Bayes and Bayesian Networks*, readers are asked to find the *maximum a posteriori* classification for the data set and network provided in the chapter.

***Hands-On Analysis*: Learn data mining by doing data mining.**

Chapters 1 – 6 provide the reader with *hands-on analysis problems*, representing an opportunity for the reader to apply his or her newly-acquired data mining expertise to solving real problems using large data sets. Many people learn by doing. *Data Mining Methods and Models* provides a framework where the reader can learn data mining by

doing data mining. For example, in Chapter 4, *Logistic Regression*, readers are challenged to approach a real-world credit approval classification data set, and construct their best possible logistic regression model, using the methods learned in this chapter as possible, providing strong interpretive support for the model, including explanations of derived variables and indicator variables.

The Case Study: Bringing it All Together

Data Mining Methods and Models culminates in a detailed Case Study, *Modeling Response to Direct Mail Marketing*. Here the reader has the opportunity to see how everything he or she has learned is brought all together to create actionable and profitable solutions. The Case Study includes over 50 pages of graphical, exploratory data analysis, predictive modeling, customer profiling, and offers different solutions, depending on the requisites of the client. The models are evaluated using a custom-built cost-benefit table, reflecting the true costs of classification errors, rather than the usual methods such as overall error rate. Thus, the analyst can compare models using the estimated profit per customer contacted, and can predict how much money the models will earn, based on the number of customers contacted.

Data mining as a process.

Data Mining Methods and Models continues the coverage of data mining as a process. The particular standard process used is the *CRISP-DM* framework: the *Cross-Industry Standard Process for Data Mining*. *CRISP-DM* demands that data mining be seen as an entire process, from communication of the business problem, through data collection and management, data preprocessing, model building, model evaluation, and, finally, model deployment. Therefore, this book is not only for analysts and managers, but also for data management professionals, database analysts, and decision makers.

The Software

The software used in this book includes the following:

- *Clementine* data mining software suite,
- *SPSS* statistical software,
- *Minitab* statistical software, and
- *WEKA* open source data mining software.

Clementine (http://www.spss.com/clementine/) is one of the most widely used data mining software suites, and is distributed by *SPSS*, whose base software is also used in this book. *SPSS* is available for download on a trial basis from their website at www.spss.com. *Minitab* is an easy-to-use statistical software package, that is available for download on a trial basis from their website at www.minitab.com.

WEKA: The Open-Source Alternative

The Weka (Waikato Environment for Knowledge Analysis) machine learning workbench is open source software issued under the GNU General Public License, which includes a collection of tools for completing many data mining tasks. *Data Mining Methods and Models* presents several hands-on, step-by-step tutorial examples using Weka 3.4, along with input files available from the book's companion web site www.dataminingconsultant.com. The reader is shown how to carry out the following types of analysis, using WEKA: Logistic Regression (Chapter 4), Naïve Bayes classification (Chapter 5), Bayesian Networks classification (Chapter 5), and Genetic Algorithms (Chapter 6). For more information regarding Weka see http://www.cs.waikato.ac.nz/~ml/. The author is deeply grateful to James Steck for providing these WEKA examples and exercises. James Steck (james_steck@comcast.net) served as Graduate Assistant to the author during the 2004-2005 academic year. He was one of the first students to complete the Master of Science in Data Mining from Central Connecticut State University in 2005 (GPA 4.0), and received the first data mining Graduate Academic Award. James lives with his wife and son in Issaquah, WA.

The Companion Website: www.dataminingconsultant.com

The reader will find supporting materials, both for this book and for the other data mining books written by Daniel Larose for *Wiley Interscience*, at the companion website, www.dataminingconsultant.com. There one may download the many data sets used in the book, so that the reader may develop a hands-on feel for the analytic methods and models encountered throughout the book. Errata are also available, as is a comprehensive set of data mining resources, including links to data sets, data mining groups, and research papers.

However, the real power of the companion website is available to faculty adopters of the textbook, who will have access to the following resources:

- Solutions to all the exercises, including the hands-on analyses,
- Powerpoint® presentations of each chapter, ready for deployment in the classroom,
- Sample data mining course projects, written by the author for use in his own courses, and ready to be adapted for your course,
- Real-world data sets, to be used with the course projects,
- Multiple-choice chapter quizzes, and
- Chapter-by-chapter web resources.

***Data Mining Methods and Models* as a textbook.**

Data Mining Methods and Models naturally fits the role of textbook for an introductory course in data mining. Instructors may appreciate:

- The presentation of data mining as a *process*,
- The "White box" approach, emphasizing an understanding of the underlying algorithmic structures,
    - Algorithm walk-throughs,
    - Application of the algorithms to large data sets,
    - Chapter exercises, and
    - Hands-on analysis,
- The logical presentation, flowing naturally from the *CRISP-DM* standard process and the set of data mining tasks.
- The detailed *Case Study*, bringing together many of the lessons learned from both *Data Mining Methods and Models* and *Discovering Knowledge in Data*.
- The companion website, providing the array of resources for adopters detailed above.

Data Mining Methods and Models is appropriate for advanced undergraduate or graduate-level courses. Some calculus is assumed in a few of the chapters, but the gist of the development can be understood without it. An introductory statistics course would be nice, but is not required. No computer programming or database expertise is required.

Acknowledgements

I wish to thank all the folks at *Wiley*, especially my editor, Val Moliere, for your guidance and support. A heartfelt thanks to James Steck for contributing the WEKA material to this volume.

I also wish to thank Dr. Chun Jin, Dr. Daniel S. Miller, Dr. Roger Bilisoly, and Dr. Darius Dziuda, my colleagues in the Master of Science in Data Mining program at Central Connecticut State University, Dr. Timothy Craine, the chair of the Department of Mathematical Sciences, Dr. Dipak K. Dey, Chair of the Department of Statistics at the University of Connecticut, and Dr. John Judge, Chair of the Department of Mathematics at Westfield State College. Without you, this would have remained a dream.

Thanks to my mom, Irene R. Larose, who passed away this year, and to my dad, Ernest L. Larose. You made all this possible. Thanks to my daughter Chantal for your lovely art work and boundless joy. Thanks to my twin children Tristan and Ravel for sharing the computer, and for sharing your true perspective. Not least, I would like to express my eternal gratitude to my dear wife, Debra J. Larose, for her patience and love, and "For everlasting bond of fellowship."

"Live hand in hand,
and together we'll stand,
on the threshold of a dream ..."
– The Moody Blues

Daniel T. Larose, Ph.D.
Director, Data Mining @CCSU
www.math.ccsu.edu/larose