DANIEL T. LAROSE

WILEY

DATA MINING
METHODS AND MODELS

WWW COMPANION WEB SITE AVAILABLE

DATA MINING METHODS AND MODELS

DANIEL T. LAROSE

Department of Mathematical Sciences
Central Connecticut State University

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format. For more information about Wiley products, visit our web site at www.wiley.com

Library of Congress Cataloging-in-Publication Data:

Larose, Daniel T.
Data mining methods and models / Daniel T. Larose.
p. cm.
Includes bibliographical references.
ISBN-13 978-0-471-66656-1
ISBN-10 0-471-66656-4 (cloth)
1. Data mining. I. Title.
QA76.9.D343L378 2005
005.74–dc22

2005010801

Printed in the United States of America

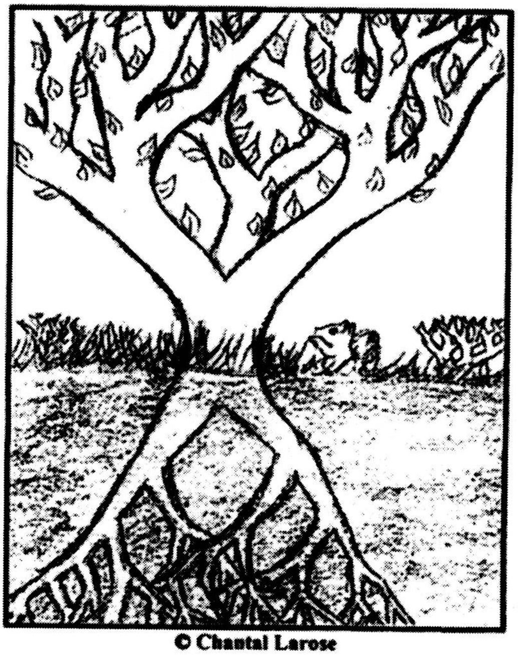
10 9 8 7 6 5 4 3 2 1

DATA MINING METHODS AND MODELS

DEDICATION

To those who have gone before,
including my parents, Ernest Larose (1920–1981)
and Irene Larose (1924–2005),
and my daughter, Ellyriane Soleil Larose (1997–1997);

For those who come after,
including my daughters, Chantal Danielle Larose (1988)
and Ravel Renaissance Larose (1999),
and my son, Tristan Spring Larose (1999).



© Chantal Larose

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, and Padhraic Smyth, *Principles of Data Mining*, MIT Press, Cambridge, MA, 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 10 emerging technologies that will change the world.

Because data mining represents such an important field, Wiley-Interscience and I 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. 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:

- Models and techniques to uncover hidden nuggets of information
- Insight into how the data mining algorithms really work
- 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 the 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
- 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 we observe how a single new data value can seriously alter the model results. Also, in Chapter 6 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 we analytically unlock the relationship between nutrition rating and cereal content using a real-world data set. In Chapter 1 we apply principal components analysis to real-world census data about California. All data sets are available from the book series Web site: `www.dataminingconsultant.com`.

Chapter Exercises: Checking to Make Sure That 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 having 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 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 to 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 by which the reader can learn data mining by doing data mining. For example, in Chapter 4 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 to provide strong interpretive support for the model, including explanations of derived and indicator variables.

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 that 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, and 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.

SOFTWARE

The software used in this book includes the following:

- Clementine data mining software suite
- SPSS statistical software
- Minitab statistical software
- WEKA open-source data mining software

Clementine (`http://www.spss.com/clementine/`), one of the most widely used data mining software suites, is distributed by SPSS, whose base software is also used in this book. SPSS is available for download on a trial basis from their

Web site at www.spss.com. Minitab is an easy-to-use statistical software package, available for download on a trial basis from their Web site at www.minitab.com.

WEKA: 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), naive 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, Washington.

COMPANION WEB SITE: www.dataminingconsultant.com

The reader will find supporting materials for this book and for my other data mining books written for Wiley-Interscience, at the companion Web site, www.dataminingconsultant.com. There one may download the many data sets used in the book, so that the reader may develop a hands-on feeling 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 Web site is available to faculty adopters of the textbook, who 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
- 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 will appreciate the following:

- 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

  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 Web site, 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.

ACKNOWLEDGMENTS

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, 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 book would have remained a dream.

Thanks to my mom, Irene R. Larose, who passed away this year, and to my dad, Ernest L. Larose, who made all this possible. Thanks to my daughter Chantal for her lovely artwork and boundless joy. Thanks to my twin children, Tristan and Ravel, for sharing the computer and for sharing their 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

CONTENTS

CHAPTER 1

DIMENSION REDUCTION METHODS

NEED FOR DIMENSION REDUCTION IN DATA MINING

PRINCIPAL COMPONENTS ANALYSIS

FACTOR ANALYSIS

USER-DEFINED COMPOSITES

NEED FOR DIMENSION REDUCTION IN DATA MINING

The databases typically used in data mining may have millions of records and thousands of variables. It is unlikely that all of the variables are independent, with no correlation structure among them. As mentioned in *Discovering Knowledge in Data: An Introduction to Data Mining* [1], data analysts need to guard against *multicollinearity*, a condition where some of the predictor variables are correlated with each other. Multicollinearity leads to instability in the solution space, leading to possible incoherent results, such as in multiple regression, where a multicollinear set of predictors can result in a regression that is significant overall, even when none of the individual variables are significant. Even if such instability is avoided, inclusion of variables that are highly correlated tends to overemphasize a particular component of the model, since the component is essentially being double counted.

Bellman [2] noted that the sample size needed to fit a multivariate function grows exponentially with the number of variables. In other words, higher-dimension spaces are inherently sparse. For example, the empirical rule tells us that in one dimension, about 68% of normally distributed variates lie between 1 and -1, whereas for a 10-dimensional multivariate normal distribution, only 0.02% of the data lie within the analogous hypersphere.

The use of too many predictor variables to model a relationship with a response variable can unnecessarily complicate the interpretation of the analysis and violates the principle of parsimony: that one should consider keeping the number of predictors

Data Mining Methods and Models By Daniel T. Larose

to a size that could easily be interpreted. Also, retaining too many variables may lead to overfitting, in which the generality of the findings is hindered because the new data do not behave the same as the training data for all the variables.

Further, analysis solely at the variable level might miss the fundamental underlying relationships among predictors. For example, several predictors might fall naturally into a single group (a *factor* or a *component*) that addresses a single aspect of the data. For example, the variables savings account balance, checking account-balance, home equity, stock portfolio value, and 401K balance might all fall together under the single component, *assets*.

In some applications, such as image analysis, retaining full dimensionality would make most problems intractable. For example, a face classification system based on 256 × 256 pixel images could potentially require vectors of dimension 65,536. Humans are endowed innately with visual pattern recognition abilities, which enable us in an intuitive manner to discern patterns in graphic images at a glance, patterns that might elude us if presented algebraically or textually. However, even the most advanced data visualization techniques do not go much beyond five dimensions. How, then, can we hope to visualize the relationship among the hundreds of variables in our massive data sets?

Dimension reduction methods have the goal of using the correlation structure among the predictor variables to accomplish the following:

- To reduce the number of predictor components
- To help ensure that these components are independent
- To provide a framework for interpretability of the results

In this chapter we examine the following dimension reduction methods:

- Principal components analysis
- Factor analysis
- User-defined composites

This chapter calls upon knowledge of matrix algebra. For those of you whose matrix algebra may be rusty, see the book series Web site for review resources. We shall apply all of the following terminology and notation in terms of a concrete example, using real-world data.

PRINCIPAL COMPONENTS ANALYSIS

Principal components analysis (PCA) seeks to explain the correlation structure of a set of predictor variables using a smaller set of linear combinations of these variables. These linear combinations are called *components*. The total variability of a data set produced by the complete set of m variables can often be accounted for primarily by a smaller set of k linear combinations of these variables, which would mean that there is almost as much information in the k components as there is in the original m variables. If desired, the analyst can then replace the original m variables with the $k < m$