

**Wiley Series on Methods and
Applications in Data Mining**

Daniel T. Larose, Series Editor

Second Edition

DISCOVERING KNOWLEDGE IN DATA

An Introduction to Data Mining

Daniel T. Larose • Chantal D. Larose

WILEY SERIES ON METHODS AND APPLICATIONS IN DATA MINING

Series Editor: **Daniel T. Larose**

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition • Daniel T. Larose and Chantal D. Larose

Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data • Darius M. Dziuda

Knowledge Discovery with Support Vector Machines • Lutz Hamel

Data-Mining on the Web: Uncovering Patterns in Web Content, Structure, and Usage • Zdravko Markov and Daniel Larose

Data Mining Methods and Models • Daniel Larose

Practical Text Mining with Perl • Roger Bilisoly

SECOND EDITION

DISCOVERING KNOWLEDGE IN DATA

An Introduction to Data Mining

**DANIEL T. LAROSE
CHANTAL D. LAROSE**



WILEY

Copyright © 2014 by John Wiley & Sons, Inc. All rights reserved.

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

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

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 or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States 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 may not be available in electronic formats. For more information about Wiley products, visit our website at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Larose, Daniel T.

Discovering knowledge in data : an introduction to data mining / Daniel T.

Larose and Chantal D. Larose. – Second edition.

pages cm

Includes index.

ISBN 978-0-470-90874-7 (hardback)

1. Data mining. I. Larose, Chantal D. II. Title.

QA76.9.D343L38 2014

006.3'12-dc23

2013046021

CONTENTS

Preface

What is Data Mining?

Why is This Book Needed?

What's New for the Second Edition?

Danger! Data Mining is Easy to Do Badly.

“White Box” Approach: Understanding
the Underlying Algorithmic and Model Structures

Data Mining as a Process

Graphical Approach, Emphasizing Exploratory Data
Analysis

How The Book is Structured

Acknowledgments

Chapter 1: An Introduction to Data Mining

1.1 What is Data Mining?

1.2 Wanted: Data Miners

1.3 The Need for Human Direction of Data Mining

1.4 The Cross-Industry Standard Practice
for Data Mining

1.5 Fallacies of Data Mining

1.6 What Tasks Can Data Mining Accomplish?

References

Exercises

Note

Chapter 2: Data Preprocessing

2.1 Why do We Need to Preprocess the Data?

2.2 Data Cleaning

[2.3 Handling Missing Data](#)

[2.4 Identifying Misclassifications](#)

[2.5 Graphical Methods for Identifying Outliers](#)

[2.6 Measures of Center and Spread](#)

[2.7 Data Transformation](#)

[2.8 Min-Max Normalization](#)

[2.9 Z-Score Standardization](#)

[2.10 Decimal Scaling](#)

[2.11 Transformations to Achieve Normality](#)

[2.12 Numerical Methods for Identifying Outliers](#)

[2.13 Flag Variables](#)

[2.14 Transforming Categorical Variables into Numerical Variables](#)

[2.15 Binning Numerical Variables](#)

[2.16 Reclassifying Categorical Variables](#)

[2.17 Adding an Index Field](#)

[2.18 Removing Variables that are Not Useful](#)

[2.19 Variables that Should Probably Not Be Removed](#)

[2.20 Removal of Duplicate Records](#)

[2.21 A Word About Id Fields](#)

[References](#)

[Exercises](#)

[Hands-On Analysis](#)

[Notes](#)

[Chapter 3: Exploratory Data Analysis](#)

[3.1 Hypothesis Testing Versus Exploratory Data Analysis](#)

[3.2 Getting to Know the Data Set](#)

[3.3 Exploring Categorical Variables](#)

[3.4 Exploring Numeric Variables](#)

[3.5 Exploring Multivariate Relationships](#)

[3.6 Selecting Interesting Subsets of the Data for Further Investigation](#)

[3.7 Using EDA to Uncover Anomalous Fields](#)

[3.8 Binning Based on Predictive Value](#)

[3.9 Deriving New Variables: Flag Variables](#)

[3.10 Deriving New Variables: Numerical Variables](#)

[3.11 Using EDA to Investigate Correlated Predictor Variables](#)

[3.12 Summary](#)

[Reference](#)

[Exercises](#)

[Hands-On Analysis](#)

[Note](#)

[Chapter 4: Univariate Statistical Analysis](#)

[4.1 Data Mining Tasks in *Discovering Knowledge in Data*](#)

[4.2 Statistical Approaches to Estimation and Prediction](#)

[4.3 Statistical Inference](#)

[4.4 How Confident are We in Our Estimates?](#)

[4.5 Confidence Interval Estimation of the Mean](#)

[4.6 How to Reduce the Margin of Error](#)

[4.7 Confidence Interval Estimation of the Proportion](#)

[4.8 Hypothesis Testing for the Mean](#)

[4.9 Assessing the Strength of Evidence Against the Null Hypothesis](#)

[4.10 Using Confidence Intervals to Perform Hypothesis Tests](#)

[4.11 Hypothesis Testing for the Proportion](#)

[Reference](#)

[Exercises](#)

[Chapter 5: Multivariate Statistics](#)

[5.1 Two-Sample \$t\$ -Test for Difference in Means](#)

[5.2 Two-Sample \$Z\$ -Test for Difference in Proportions](#)

[5.3 Test for Homogeneity of Proportions](#)

[5.4 Chi-Square Test for Goodness of Fit of Multinomial Data](#)

[5.5 Analysis of Variance](#)

[5.6 Regression Analysis](#)

[5.7 Hypothesis Testing in Regression](#)

[5.8 Measuring the Quality of a Regression Model](#)

[5.9 Dangers of Extrapolation](#)

[5.10 Confidence Intervals for the Mean Value of \$y\$ Given \$x\$](#)

[5.11 Prediction Intervals for a Randomly Chosen Value of \$y\$ Given \$x\$](#)

[5.12 Multiple Regression](#)

[5.13 Verifying Model Assumptions](#)

[Reference](#)

[Exercises](#)

[Hands-On Analysis](#)

[Note](#)

[Chapter 6: Preparing to Model the Data](#)

[6.1 Supervised Versus Unsupervised Methods](#)

[6.2 Statistical Methodology and Data Mining Methodology](#)

[6.3 Cross-Validation](#)

[6.4 Overfitting](#)

[6.5 BIAS-Variance Trade-Off](#)

[6.6 Balancing the Training Data Set](#)

[6.7 Establishing Baseline Performance](#)

[Reference](#)

[Exercises](#)

[Chapter 7: *k*-Nearest Neighbor Algorithm](#)

[7.1 Classification Task](#)

[7.2 *k*-Nearest Neighbor Algorithm](#)

[7.3 Distance Function](#)

[7.4 Combination Function](#)

[7.5 Quantifying Attribute Relevance: Stretching the Axes](#)

[7.6 Database Considerations](#)

[7.7 *k*-Nearest Neighbor Algorithm for Estimation and Prediction](#)

[7.8 Choosing *k*](#)

[7.9 Application of *k*-Nearest Neighbor Algorithm Using IBM/SPSS Modeler](#)

[Exercises](#)

[Hands-On Analysis](#)

[Chapter 8: Decision Trees](#)

[8.1 What is a Decision Tree?](#)

[8.2 Requirements for Using Decision Trees](#)

[8.3 Classification and Regression Trees](#)

[8.4 C4.5 Algorithm](#)

[8.5 Decision Rules](#)

[8.6 Comparison of the C5.0 and Cart Algorithms Applied to Real Data](#)

[References](#)

[Exercises](#)

[Hands-On Analysis](#)

[Chapter 9: Neural Networks](#)

[9.1 Input and Output Encoding](#)

[9.2 Neural Networks for Estimation and Prediction](#)

[9.3 Simple Example of a Neural Network](#)

[9.4 Sigmoid Activation Function](#)

[9.5 Back-Propagation](#)

[9.6 Termination Criteria](#)

[9.7 Learning Rate](#)

[9.8 Momentum Term](#)

[9.9 Sensitivity Analysis](#)

[9.10 Application of Neural Network Modeling](#)

[References](#)

[Exercises](#)

[Hands-On Analysis](#)

[Chapter 10: Hierarchical and *k*-Means Clustering](#)

[10.1 The Clustering Task](#)

[10.2 Hierarchical Clustering Methods](#)

[10.3 Single-Linkage Clustering](#)

[10.4 Complete-Linkage Clustering](#)

[10.5 *k*-Means Clustering](#)

[10.6 Example of *k*-Means Clustering at Work](#)

[10.7 Behavior of MSB, MSE, and PSEUDO-*F* as the *k*-Means Algorithm Proceeds](#)

[10.8 Application of *k*-Means Clustering Using SAS Enterprise Miner](#)

[10.9 Using Cluster Membership to Predict Churn](#)

[References](#)

[Exercises](#)

[Hands-On Analysis](#)

[Note](#)

[Chapter 11: Kohonen Networks](#)

[11.1 Self-Organizing Maps](#)

[11.2 Kohonen Networks](#)

[11.3 Example of a Kohonen Network Study](#)

[11.4 Cluster Validity](#)

[11.5 Application of Clustering Using Kohonen Networks](#)

[11.6 Interpreting the Clusters](#)

[11.7 Using Cluster Membership as Input to Downstream Data Mining Models](#)

[References](#)

[Exercises](#)

[Hands-On Analysis](#)

[Chapter 12: Association Rules](#)

[12.1 Affinity Analysis and Market Basket Analysis](#)

[12.2 Support, Confidence, Frequent Itemsets, and the a Priori Property](#)

[12.3 How Does the a Priori Algorithm Work?](#)

[12.4 Extension from Flag Data to General Categorical Data](#)

[12.5 Information-Theoretic Approach: Generalized Rule Induction Method](#)

[12.6 Association Rules are Easy to do Badly](#)

[12.7 How can we Measure the Usefulness of Association Rules?](#)

[12.8 Do Association Rules Represent Supervised or Unsupervised Learning?](#)

[12.9 Local Patterns Versus Global Models](#)

[References](#)

[Exercises](#)

[Hands-On Analysis](#)

[Chapter 13: Imputation of Missing Data](#)

[13.1 Need for Imputation of Missing Data](#)

[13.2 Imputation of Missing Data: Continuous Variables](#)

[13.3 Standard Error of the Imputation](#)

[13.4 Imputation of Missing Data: Categorical Variables](#)

[13.5 Handling Patterns in Missingness](#)

[Reference](#)

[Exercises](#)

[Hands-On Analysis](#)

[Notes](#)

[Chapter 14: Model Evaluation Techniques](#)

[14.1 Model Evaluation Techniques for the Description Task](#)

[14.2 Model Evaluation Techniques for the Estimation and Prediction Tasks](#)

[14.3 Model Evaluation Techniques for the Classification Task](#)

[14.4 Error Rate, False Positives, and False Negatives](#)

[14.5 Sensitivity and Specificity](#)

[14.6 Misclassification Cost Adjustment to Reflect Real-World Concerns](#)

[14.7 Decision Cost/Benefit Analysis](#)

[14.8 Lift Charts and Gains Charts](#)

[14.9 Interweaving Model Evaluation with Model Building](#)

[14.10 Confluence of Results: Applying a Suite of Models](#)

[Reference](#)

[Exercises](#)

[Hands-On Analysis](#)

[Notes](#)

[Appendix: Data Summarization and Visualization](#)

[Part 1 Summarization 1: Building Blocks of Data Analysis](#)

[Part 2 Visualization: Graphs and Tables for Summarizing and Organizing Data](#)

[Part 3 Summarization 2: Measures of Center, Variability, and Position](#)

[Part 4 Summarization and Visualization of Bivariate Relationships](#)

[Index](#)

[End User License Agreement](#)

List of Tables

[Chapter 1](#)

[Table 1.1](#)

[Table 1.2](#)

[Chapter 2](#)

[Table 2.1](#)

[Table 2.2](#)

[Table 2.3](#)

[Chapter 3](#)

[Table 3.1](#)

[Table 3.2](#)

[Table 3.3](#)

[Table 3.4](#)

[Table 3.5](#)

[Table 3.6](#)

[Table 3.7](#)

[Table 3.8](#)

[Table 3.9](#)

[Chapter 4](#)

[Table 4.1](#)

[Table 4.2](#)

[Table 4.3](#)

[Table 4.4](#)

[Table 4.5](#)

[Table 4.6](#)

[Table 4.7](#)

[Table 4.8](#)

[Chapter 5](#)

[Table 5.1](#)

[Table 5.2](#)

[Table 5.3](#)

[Table 5.4](#)

[Table 5.5](#)

[Table 5.6](#)

[Table 5.7](#)

[Table 5.8](#)

[Table 5.9](#)

[Table 5.10](#)

[Table 5.11](#)

[Table 5.12](#)

[Chapter 6](#)

[Table 6.1](#)

[Chapter 7](#)

[Table 7.1](#)

[Table 7.2](#)

[Table 7.3](#)

[Table 7.4](#)

[Table 7.5](#)

[Chapter 8](#)

[Table 8.1](#)

[Table 8.2](#)

[Table 8.3](#)

[Table 8.4](#)

[Table 8.5](#)

[Table 8.6](#)

[Table 8.7](#)

[Table 8.8](#)

[Table 8.9](#)

[Table 8.10](#)

[Table 8.11](#)

[Chapter 9](#)

[Table 9.1](#)

[Chapter 10](#)

[Table 10.1](#)

[Table 10.2](#)

[Table 10.3](#)

[Table 10.4](#)

[Table 10.5](#)

[Chapter 11](#)

[Table 11.1](#)

[Chapter 12](#)

[Table 12.1](#)

[Table 12.2](#)

[Table 12.3](#)

[Table 12.4](#)

[Table 12.5](#)

[Table 12.6](#)

[Table 12.7](#)

[Table 12.8](#)

[Chapter 14](#)

[Table 14.1](#)

[Table 14.2](#)

[Table 14.3](#)

[Table 14.4](#)

[Table 14.5](#)

[Appendix](#)

[Table A.1](#)

[Table A.2](#)

[Table A.3](#)

[Table A.4](#)

[Table A.5](#)

List of Illustrations

[Chapter 1](#)

[Figure 1.1 CRISP-DM is an iterative, adaptive process.](#)

[Figure 1.2 Regression estimates lie on the regression line.](#)

[Figure 1.3 Which drug should be prescribed for which type of patient?](#)

[Chapter 2](#)

[Figure 2.1 Some of our field values are missing.](#)

[Figure 2.2 Replacing missing field values with user-defined constants.](#)

[Figure 2.3 Replacing missing field values with means or modes.](#)

[Figure 2.4 Replacing missing field values with random draws from the distribution of the variable.](#)

Figure 2.5 Histogram of vehicle weights: can you find the outlier?

Figure 2.6 Scatter plot of *mpg* against *Weightlbs* shows two outliers.

Figure 2.7 Statistical summary of *customer service calls*.

Figure 2.8 Summary statistics for *weight*.

Figure 2.9 Standard normal *Z* distribution.

Figure 2.10 Original data.

Figure 2.11 *Z*-Standardized data are still right-skewed, not normally distributed.

Figure 2.12 Right-skewed data have positive skewness.

Figure 2.13 Left-skewed data have negative skewness.

Figure 2.14 Statistics for calculating skewness.

Figure 2.15 Square root transformation somewhat reduces skewness.

Figure 2.16 Natural log transformation reduces skewness even further.

Figure 2.17 Statistics for calculating skewness.

Figure 2.18 The transformation $inverse_sqrt(weight)$ has eliminated the skewness, but is still not normal.

Figure 2.19 Statistics for $inverse_sqrt(weight)$.

Figure 2.20 Normal probability plot of $inverse_sqrt(weight)$ indicates nonnormality.

Figure 2.21 Normal probability plot of normally distributed data.

[Figure 2.22 Illustration of binning methods.](#)

[Chapter 3](#)

[Figure 3.1 Field values of the first 10 records in the *churn* data set.](#)

[Figure 3.2 Summarization and visualization of the *churn* data set.](#)

[Figure 3.3 Churners and non-churners.](#)

[Figure 3.4 Comparison bar chart of churn proportions, by International Plan participation.](#)

[Figure 3.5 Comparison bar chart of churn proportions, by International Plan participation, with equal bar length.](#)

[Figure 3.6 The clustered bar chart is the graphical counterpart of the contingency table.](#)

[Figure 3.7 Comparative pie chart associated with Table 3.2.](#)

[Figure 3.8 Clustered bar chart associated with Table 3.3.](#)

[Figure 3.9 Comparative pie chart associated with Table 3.3.](#)

[Figure 3.10 Those without the Voice Mail Plan are more likely to churn.](#)

[Figure 3.11 Multilayer clustered bar chart.](#)

[Figure 3.12 Statistics for multilayer clustered bar chart.](#)

[Figure 3.13 Directed web graph supports earlier findings.](#)

[Figure 3.14 Histogram of customer service calls with no overlay.](#)

Figure 3.15 Histogram of customer service calls, with churn overlay.

Figure 3.16 “Normalized” histogram of customer service calls, with churn overlay.

Figure 3.17 (a) Nonnormalized histogram of day minutes; (b) normalized histogram of day minutes.

Figure 3.18 (a) Nonnormalized histogram of evening minutes; (b) normalized histogram of evening minutes.

Figure 3.19 (a) Nonnormalized histogram of night minutes; (b) normalized histogram of night minutes.

Figure 3.20 (a) Nonnormalized histogram of *International Calls*; (b) normalized histogram of *International Calls*.

Figure 3.21 *t*-test is significant for difference in mean international calls for churners and non-churners.

Figure 3.22 Customers with both high day minutes and high evening minutes are at greater risk of churning.

Figure 3.23 There is an interaction effect between *customer service calls* and *day minutes*, with respect to churn.

Figure 3.24 Very high proportion of churners for high customer service calls and low day minutes.

Figure 3.25 Much lower proportion of churners for high customer service calls and high day minutes.

Figure 3.26 Only three area codes for all records.

Figure 3.27 Anomaly: three area codes distributed randomly across all 50 states.

Figure 3.28 Binning *evening minutes* helps to tease out a signal from the noise.

Figure 3.29 Use the equation of the line to separate the records, via a flag variable.

Figure 3.30 (a) Nonnormalized histogram of *CSCInternational Z*; (b) normalized histogram of *CSCInternational Z*.

Figure 3.31 Matrix plot of *Day Minutes*, *Day Calls*, and *Day Charge*.

Figure 3.32 Correlations and *p*-values.

Figure 3.33 *Minitab* regression output for *Day Charge* vs. *Day Minutes*.

Figure 3.34 *Account length* is positively correlated with *day calls*.

Chapter 4

Figure 4.1 Summary statistics of customer service calls.

Figure 4.2 Summary statistics of customer service calls for those with both the International Plan and VoiceMail Plan and with more than 200 day minutes.

Figure 4.3 Reject values of μ_0 that would fall outside the equivalent confidence interval.

Figure 4.4 Placing the hypothesized values of μ_0 on the number line in relation to the confidence interval informs us immediately of the conclusion.

Chapter 5

Figure 5.1 Dotplot of groups A, B, and C shows considerable overlap.

[Figure 5.2 Dotplot of Groups D, E, and F shows little overlap.](#)

[Figure 5.3 ANOVA results for \$H_0: \mu_A = \mu_B = \mu_C\$.](#)

[Figure 5.4 ANOVA results for \$H_0: \mu_D = \mu_E = \mu_F\$.](#)

[Figure 5.5 Scatter plot of nutritional rating versus sugar content for 76 cereals.](#)

[Figure 5.6 Regression results for using *sugars* to estimate *rating*.](#)

[Figure 5.7 Dangers of extrapolation.](#)

[Figure 5.8 Multiple regression results.](#)

[Figure 5.9 Plots for verifying regression model assumptions. Note the outlier.](#)

[Figure 5.10 Plots for verifying regression model assumptions, after outlier omitted.](#)

[Chapter 6](#)

[Figure 6.1 The optimal level of model complexity is at the minimum error rate on the test set.](#)

[Figure 6.2 Low complexity separator with high error rate.](#)

[Figure 6.3 High complexity separator with low error rate.](#)

[Figure 6.4 With more data: low complexity separator need not change much; high complexity separator needs much revision.](#)

[Chapter 7](#)

[Figure 7.1 Scatter plot of sodium/potassium ratio against age, with drug overlay.](#)

[Figure 7.2 Close-up of three nearest neighbors to new patient 2.](#)

[Figure 7.3 Close-up of three nearest neighbors to new patient 3.](#)

[Figure 7.4 Euclidean distance.](#)

[Figure 7.5 Modeler \$k\$ -nearest neighbor results.](#)

[Chapter 8](#)

[Figure 8.1 Simple decision tree.](#)

[Figure 8.2 CART decision tree after initial split.](#)

[Figure 8.3 CART decision tree after decision node \$A\$ split.](#)

[Figure 8.4 CART decision tree, fully_grown form.](#)

[Figure 8.5 Modeler's CART decision tree.](#)

[Figure 8.6 C4.5 concurs with CART in choosing assets for the initial partition.](#)

[Figure 8.7 C4.5 Decision tree: fully_grown form.](#)

[Figure 8.8 CART decision tree for the adult data set.](#)

[Figure 8.9 C5.0 decision tree for the adult data set.](#)

[Chapter 9](#)

[Figure 9.1 Real neuron and artificial neuron model.](#)

[Figure 9.2 Simple neural network.](#)

[Figure 9.3 Graph of the sigmoid function \$y = f\(x\) = 1/\(1 + e^{-x}\)\$.](#)

[Figure 9.4 Using the slope of SSE with respect to \$w_1\$ to find weight adjustment direction.](#)

[Figure 9.5 Large \$\eta\$ may cause algorithm to overshoot global minimum.](#)

[Figure 9.6 Small momentum \$\alpha\$ may cause algorithm to undershoot global minimum.](#)

Figure 9.7 Large momentum α may cause algorithm to overshoot global minimum.

Figure 9.8 Neural network for the adult data set generated by Insightful Miner.

Figure 9.9 Some of the neural network weights for the income example.

Figure 9.10 Most important variables: results from sensitivity analysis.

Chapter 10

Figure 10.1 Clusters should have small within-cluster variation compared to the between-cluster variation.

Figure 10.2 Single-linkage agglomerative clustering on the sample data set.

Figure 10.3 Complete-linkage agglomerative clustering on the sample data set.

Figure 10.4 How will k -means partition these data into $k = 2$ clusters?

Figure 10.5 Clusters and centroids Δ after first pass through k -means algorithm.

Figure 10.6 Clusters and centroids Δ after second pass through k -means algorithm.

Figure 10.7 Enterprise Miner profile of International Plan adopters across clusters.

Figure 10.8 VoiceMail Plan adopters and nonadopters are mutually exclusive.

Figure 10.9 Distribution of *customer service calls* is similar across clusters.

Figure 10.10 Churn behavior across clusters for International Plan adopters and nonadopters.

[Figure 10.11 Churn behavior across clusters for VoiceMail Plan adopters and nonadopters.](#)

[Chapter 11](#)

[Figure 11.1 Topology of a simple self-organizing map for clustering records by age and income.](#)

[Figure 11.2 Example: topology of the 2 × 2 Kohonen network.](#)

[Figure 11.3 Topology of 3 × 3 Kohonen network used for clustering the churn data set.](#)

[Figure 11.4 Modeler uncovered six clusters.](#)

[Figure 11.5 International Plan adopters reside exclusively in Clusters 12 and 22.](#)

[Figure 11.6 Similar clusters are closer to each other.](#)

[Figure 11.7 How the variables are distributed among the clusters.](#)

[Figure 11.8 Assessing whether the means across clusters are significantly different.](#)

[Figure 11.9 Proportions of churners among the clusters.](#)

[Figure 11.10 Output of CART decision tree for data set enriched by cluster membership.](#)

[Chapter 12](#)

[Figure 12.1 Association rules for vegetable stand data, generated by Modeler.](#)

[Figure 12.2 Association rules for categorical attributes found by the a priori algorithm.](#)

[Figure 12.3 An association rule that is worse than useless.](#)

Figure 12.4 This association rule is useful, because the posterior probability (0.60029) is much greater than the prior probability (0.3316).

Figure 12.5 Profitable pattern: VoiceMail Plan adopters less likely to churn.

Chapter 13

Figure 13.1 Multiple regression results for imputation of missing potassium values. (The predicted values section of this output is for Almond Delight only.)

Figure 13.2 CART model for imputing the missing value of *maritalstatus*.

Chapter 14

Figure 14.1 Regression results, with MSE and s indicated.

Figure 14.2 Lift chart for model 1: strong lift early, then falls away rapidly.

Figure 14.3 Gains chart for model 1.

Figure 14.4 Combined lift chart for models 1 and 2.

Appendix

Figure A.1 Bar chart for *marital status*.

Figure A.2 Pie chart of *marital status*.

Figure A.3 Histogram of *income*.

Figure A.4 Stem-and-leaf display of *income*.

Figure A.5 Dotplot of *income*.

Figure A.6 Symmetric and skewed curves.

Figure A.7 Boxplot of left-skewed data.

Figure A.8 Clustered bar chart for *risk*, clustered by *mortgage*.

Figure A.9 Individual value plot of *income* versus *risk*.

Figure A.10 Some possible relationships between *x* and *y*.

Preface

What is Data Mining?

According to the Gartner Group,

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.

Today, there are a variety of terms used to describe this process, including *analytics*, *predictive analytics*, *big data*, *machine learning*, and *knowledge discovery in databases*. But these terms all share in common the objective of mining actionable nuggets of knowledge from large data sets. We shall therefore use the term *data mining* to represent this process throughout this text.

Why is This Book Needed?

Humans are inundated with data in most fields. Unfortunately, these valuable data, which cost firms millions to collect and collate, are languishing in warehouses and repositories. *The problem is that there are not enough trained human analysts available who are skilled at translating all of these data into knowledge*, and thence up the taxonomy tree into wisdom. This is why this book is needed.

The McKinsey Global Institute reports:[1](#)

There will be a shortage of talent necessary for organizations to take advantage of big data. A significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data.... We project that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions. ... In addition, we project a need for 1.5 million additional managers and analysts in the United States who can ask the right questions and consume the results of the analysis of big data effectively.

This book is an attempt to help alleviate this critical shortage of data analysts. *Discovering Knowledge in Data: An Introduction to Data Mining* provides readers 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.

Data mining is becoming more widespread everyday, because it empowers companies to uncover profitable patterns and trends from their existing databases. Companies and institutions have spent millions of dollars to collect megabytes and terabytes of data, but are not taking advantage of the valuable and actionable information hidden deep within their data repositories. However, as the practice of data mining becomes more widespread, companies which do not apply these techniques are in danger of falling behind, and losing market share, because their competitors are applying data mining, and thereby gaining the competitive edge.

In *Discovering Knowledge in Data*, the step-by-step, hands-on solutions of real-world business problems, using widely available data mining techniques applied to real-world data sets, will appeal to managers, CIOs, CEOs, CFOs, and others who need to keep abreast of the latest methods for enhancing return-on-investment.

What's New for the Second Edition?

The second edition of *Discovery Knowledge in Data* is enhanced with an abundance of new material and useful features, including:

- Nearly 100 pages of new material.
- Three new chapters:
 - Chapter 5: *Multivariate Statistical Analysis* covers the hypothesis tests used for verifying whether data partitions are valid, along with analysis of variance, multiple regression, and other topics.