

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

©2001 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

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

1	<i>Introduction</i>	1
2	<i>Measurement and Data</i>	25
3	<i>Visualizing and Exploring Data</i>	53
4	<i>Data Analysis and Uncertainty</i>	93
5	<i>A Systematic Overview of Data Mining Algorithms</i>	141
6	<i>Models and Patterns</i>	165
7	<i>Score Functions for Data Mining Algorithms</i>	211
8	<i>Search and Optimization Methods</i>	235
9	<i>Descriptive Modeling</i>	271
10	<i>Predictive Modeling for Classification</i>	327
11	<i>Predictive Modeling for Regression</i>	367
12	<i>Data Organization and Databases</i>	399
13	<i>Finding Patterns and Rules</i>	427
14	<i>Retrieval by Content</i>	449

Contents

<i>List of Tables</i>	xvii
<i>List of Figures</i>	xix
<i>Series Foreword</i>	xxv
<i>Preface</i>	xxvii
1 Introduction	1
1.1 Introduction to Data Mining	1
1.2 The Nature of Data Sets	4
1.3 Types of Structure: Models and Patterns	9
1.4 Data Mining Tasks	11
1.5 Components of Data Mining Algorithms	15
1.5.1 Score Functions	16
1.5.2 Optimization and Search Methods	16
1.5.3 Data Management Strategies	17
1.6 The Interacting Roles of Statistics and Data Mining	18
1.7 Data Mining: Dredging, Snooping, and Fishing	22
1.8 Summary	23
1.9 Further Reading	24
2 Measurement and Data	25
2.1 Introduction	25
2.2 Types of Measurement	26
2.3 Distance Measures	31
2.4 Transforming Data	38

2.5	The Form of Data	41
2.6	Data Quality for Individual Measurements	44
2.7	Data Quality for Collections of Data	47
2.8	Conclusion	52
2.9	Further Reading	52
3	<i>Visualizing and Exploring Data</i>	53
3.1	Introduction	53
3.2	Summarizing Data: Some Simple Examples	55
3.3	Tools for Displaying Single Variables	57
3.4	Tools for Displaying Relationships between Two Variables	62
3.5	Tools for Displaying More Than Two Variables	70
3.6	Principal Components Analysis	74
3.7	Multidimensional Scaling	84
3.8	Further Reading	90
4	<i>Data Analysis and Uncertainty</i>	93
4.1	Introduction	93
4.2	Dealing with Uncertainty	94
4.3	Random Variables and Their Relationships	97
4.3.1	Multivariate Random Variables	97
4.4	Samples and Statistical Inference	102
4.5	Estimation	105
4.5.1	Desirable Properties of Estimators	106
4.5.2	Maximum Likelihood Estimation	108
4.5.3	Bayesian Estimation	116
4.6	Hypothesis Testing	124
4.6.1	Classical Hypothesis Testing	124
4.6.2	Hypothesis Testing in Context	130
4.7	Sampling Methods	132
4.8	Conclusion	138
4.9	Further Reading	139
5	<i>A Systematic Overview of Data Mining Algorithms</i>	141
5.1	Introduction	141
5.2	An Example: The CART Algorithm for Building Tree Classifiers	145
5.3	The Reductionist Viewpoint on Data Mining Algorithms	151

5.3.1	Multilayer Perceptrons for Regression and Classification	153
5.3.2	The A Priori Algorithm for Association Rule Learning	157
5.3.3	Vector-Space Algorithms for Text Retrieval	160
5.4	Discussion	162
5.5	Further Reading	164
6	<i>Models and Patterns</i>	165
6.1	Introduction	165
6.2	Fundamentals of Modeling	167
6.3	Model Structures for Prediction	168
6.3.1	Regression Models with Linear Structure	169
6.3.2	Local Piecewise Model Structures for Regression	174
6.3.3	Nonparametric “Memory-Based” Local Models	175
6.3.4	Stochastic Components of Model Structures	178
6.3.5	Predictive Models for Classification	180
6.3.6	An Aside: Selecting a Model of Appropriate Complexity	183
6.4	Models for Probability Distributions and Density Functions	184
6.4.1	General Concepts	184
6.4.2	Mixtures of Parametric Models	185
6.4.3	Joint Distributions for Unordered Categorical Data	188
6.4.4	Factorization and Independence in High Dimensions	188
6.5	The Curse of Dimensionality	193
6.5.1	Variable Selection for High-Dimensional Data	194
6.5.2	Transformations for High-Dimensional Data	195
6.6	Models for Structured Data	197
6.7	Pattern Structures	203
6.7.1	Patterns in Data Matrices	203
6.7.2	Patterns for Strings	206
6.8	Further Reading	208
7	<i>Score Functions for Data Mining Algorithms</i>	211
7.1	Introduction	211
7.2	Scoring Patterns	212
7.3	Predictive versus Descriptive Score Functions	215
7.3.1	Score Functions for Predictive Models	215
7.3.2	Score Functions for Descriptive Models	217

7.4	Scoring Models with Different Complexities	220
7.4.1	General Concepts in Comparing Models	220
7.4.2	Bias-Variance Again	221
7.4.3	Score Functions That Penalize Complexity	224
7.4.4	Score Functions Using External Validation	227
7.5	Evaluation of Models and Patterns	229
7.6	Robust Methods	231
7.7	Further Reading	232
8	<i>Search and Optimization Methods</i>	235
8.1	Introduction	235
8.2	Searching for Models and Patterns	238
8.2.1	Background on Search	238
8.2.2	The State-Space Formulation for Search in Data Mining	241
8.2.3	A Simple Greedy Search Algorithm	243
8.2.4	Systematic Search and Search Heuristics	244
8.2.5	Branch-and-Bound	246
8.3	Parameter Optimization Methods	247
8.3.1	Parameter Optimization: Background	247
8.3.2	Closed Form and Linear Algebra Methods	249
8.3.3	Gradient-Based Methods for Optimizing Smooth Functions	250
8.3.4	Univariate Parameter Optimization	251
8.3.5	Multivariate Parameter Optimization	255
8.3.6	Constrained Optimization	259
8.4	Optimization with Missing Data: The EM Algorithm	260
8.5	Online and Single-Scan Algorithms	265
8.6	Stochastic Search and Optimization Techniques	266
8.7	Further Reading	268
9	<i>Descriptive Modeling</i>	271
9.1	Introduction	271
9.2	Describing Data by Probability Distributions and Densities	272
9.2.1	Introduction	272
9.2.2	Score Functions for Estimating Probability Distributions and Densities	274
9.2.3	Parametric Density Models	275
9.2.4	Mixture Distributions and Densities	279

9.2.5	The EM Algorithm for Mixture Models	281
9.2.6	Nonparametric Density Estimation	284
9.2.7	Joint Distributions for Categorical Data	287
9.3	Background on Cluster Analysis	293
9.4	Partition-Based Clustering Algorithms	296
9.4.1	Score Functions for Partition-Based Clustering	297
9.4.2	Basic Algorithms for Partition-Based Clustering	302
9.5	Hierarchical Clustering	308
9.5.1	Agglomerative Methods	311
9.5.2	Divisive Methods	314
9.6	Probabilistic Model-Based Clustering Using Mixture Models	315
9.7	Further Reading	324
10	<i>Predictive Modeling for Classification</i>	327
10.1	A Brief Overview of Predictive Modeling	327
10.2	Introduction to Classification Modeling	329
10.2.1	Discriminative Classification and Decision Boundaries	330
10.2.2	Probabilistic Models for Classification	331
10.2.3	Building Real Classifiers	335
10.3	The Perceptron	339
10.4	Linear Discriminants	341
10.5	Tree Models	343
10.6	Nearest Neighbor Methods	347
10.7	Logistic Discriminant Analysis	352
10.8	The Naive Bayes Model	353
10.9	Other Methods	356
10.10	Evaluating and Comparing Classifiers	359
10.11	Feature Selection for Classification in High Dimensions	362
10.12	Further Reading	363
11	<i>Predictive Modeling for Regression</i>	367
11.1	Introduction	367
11.2	Linear Models and Least Squares Fitting	368
11.2.1	Computational Issues in Fitting the Model	370
11.2.2	A Probabilistic Interpretation of Linear Regression	372
11.2.3	Interpreting the Fitted Model	375
11.2.4	Inference and Generalization	377
11.2.5	Model Search and Model Building	378

11.2.6	Diagnostics and Model Inspection	381
11.3	Generalized Linear Models	384
11.4	Artificial Neural Networks	391
11.5	Other Highly Parameterized Models	393
11.5.1	Generalized Additive Models	393
11.5.2	Projection Pursuit Regression	395
11.6	Further Reading	397
12	<i>Data Organization and Databases</i>	399
12.1	Introduction	399
12.2	Memory Hierarchy	400
12.3	Index Structures	402
12.3.1	B-trees	402
12.3.2	Hash Indices	403
12.4	Multidimensional Indexing	404
12.5	Relational Databases	405
12.6	Manipulating Tables	409
12.7	The Structured Query Language (SQL)	413
12.8	Query Execution and Optimization	415
12.9	Data Warehousing and Online Analytical Processing (OLAP)	417
12.10	Data Structures for OLAP	419
12.11	String Databases	420
12.12	Massive Data Sets, Data Management, and Data Mining	421
12.12.1	Force the Data into Main Memory	422
12.12.2	Scalable Versions of Data Mining Algorithms	423
12.12.3	Special-Purpose Algorithms for Disk Access	424
12.12.4	Pseudo Data Sets and Sufficient Statistics	425
12.13	Further Reading	426
13	<i>Finding Patterns and Rules</i>	427
13.1	Introduction	427
13.2	Rule Representations	428
13.3	Frequent Itemsets and Association Rules	429
13.3.1	Introduction	429
13.3.2	Finding Frequent Sets and Association Rules	433
13.4	Generalizations	435
13.5	Finding Episodes from Sequences	436

13.6	Selective Discovery of Patterns and Rules	438
13.6.1	Introduction	438
13.6.2	Heuristic Search for Finding Patterns	439
13.6.3	Criteria for Interestingness	440
13.7	From Local Patterns to Global Models	442
13.8	Predictive Rule Induction	443
13.9	Further Reading	447
14	<i>Retrieval by Content</i>	449
14.1	Introduction	449
14.2	Evaluation of Retrieval Systems	452
14.2.1	The Difficulty of Evaluating Retrieval Performance	452
14.2.2	Precision versus Recall	453
14.2.3	Precision and Recall in Practice	456
14.3	Text Retrieval	456
14.3.1	Representation of Text	457
14.3.2	Matching Queries and Documents	461
14.3.3	Latent Semantic Indexing	465
14.3.4	Document and Text Classification	469
14.4	Modeling Individual Preferences	470
14.4.1	Relevance Feedback	470
14.4.2	Automated Recommender Systems	471
14.5	Image Retrieval	472
14.5.1	Image Understanding	473
14.5.2	Image Representation	473
14.5.3	Image Queries	474
14.5.4	Image Invariants	475
14.5.5	Generalizations of Image Retrieval	476
14.6	Time Series and Sequence Retrieval	476
14.6.1	Global Models for Time Series Data	478
14.6.2	Structure and Shape in Time Series	480
14.7	Summary	481
14.8	Further Reading	482
	<i>Appendix: Random Variables</i>	485
	<i>References</i>	491
	<i>Index</i>	525

List of Tables

1.1	Examples of data in Public Use Microdata Sample data sets.	5
2.1	A cross-classification of two binary variables.	37
3.1	Numerical codes, names, and counties for the 25 villages with dialect similarities displayed in figure 3.19.	89
5.1	Three well-known data mining algorithms broken down in terms of their algorithm components.	143
6.1	A simple contingency table for two-dimensional categorical data for a hypothetical data set of medical patients who have been diagnosed for dementia.	187
11.1	The analysis of variance decomposition table for a regression.	377
11.2	The analysis of variance decomposition table for model building.	379
11.3	Analysis of deviance table.	390
14.1	A schematic of the four possible outcomes in a retrieval experiment where documents are labeled as being “relevant” or “not relevant.”	454
14.2	A toy document-term matrix for 10 documents and 6 terms.	458
14.3	TF-IDF document-term matrix.	464
14.4	Distances resulting from a query containing the terms <i>database</i> and <i>index</i> .	464

List of Figures

1.1	A portion of a retail transaction data set displayed as a binary image, with 100 individual customers (rows) and 40 categories of items (columns).	8
2.1	A sample correlation matrix plotted as a pixel image.	34
2.2	A simple nonlinear relationship between variable V_1 and V_2 . (In these and subsequent figures V_1 and V_2 are on the X and Y axes respectively).	39
2.3	The data of figure 2.2 after the simple transformation of V_2 to $1/V_2$.	40
2.4	Another simple nonlinear relationship. Here the variance of V_2 increases as V_1 increases.	41
2.5	The data of figure 2.4 after a simple square root transformation of V_2 . Now the variance of V_2 is relatively constant as V_1 increases.	42
2.6	A plot of 200 points from highly positively correlated bivariate data (from a bivariate normal distribution), with a single easily identifiable outlier.	51
3.1	Histogram of the number of weeks of the year a particular brand of credit card was used.	58
3.2	Histogram of diastolic blood pressure for 768 females of Pima Indian descent.	59
3.3	Kernel estimate of the weights (in kg) of 856 elderly women.	61
3.4	As figure 3.3, but with more smoothing.	62

3.5	Boxplots on four different variables from the Pima Indians diabetes data set.	63
3.6	A standard scatterplot for two banking variables.	64
3.7	A scatterplot of 96,000 cases, with much overprinting. Each data point represents an individual applicant for a loan. The vertical axis shows the age of the applicant, and the horizontal axis indicates the day on which the application was made.	65
3.8	Overprinting conceals the actual strength of the correlation.	66
3.9	A contour plot of the data from figure 3.7.	67
3.10	A plot of the number of credit cards in circulation in the United Kingdom, by year.	68
3.11	Patterns of change over time in the number of miles flown by UK airlines in the 1960s.	69
3.12	Weight changes over time in a group of 10,000 school children in the 1930s. The steplike pattern in the data highlights a problem with the measurement process.	70
3.13	A scatterplot matrix for the computer CPU data.	72
3.14	A trellis plot for the epileptic seizures data.	73
3.15	An example of a star plot.	75
3.16	A parallel coordinates plot for the epileptic seizure data.	76
3.17	Scree plots for the computer CPU data set.	79
3.18	Projection onto the first two principal components.	82
3.19	A multidimensional scaling plot of the village dialect similarities data.	88
4.1	An illustration of the dual roles of probability and statistics in data analysis.	103
4.2	The likelihood function for three hypothetical data sets under a Binomial model.	109
4.3	The likelihood as a function of θ for a sample of 20 data points from a Normal density.	111
4.4	The likelihood function for the same model as in figure 4.3 but with 200 data points.	112
4.5	Means of samples of size 10(a), 100(b), and 1000(c) drawn from a population with a mean of 0.5.	136
5.1	A scatterplot of data showing color intensity versus alcohol content for a set of wines.	145
5.2	A classification tree for the data in figure 5.1.	146

5.3	The decision boundaries from the classification tree in figure 5.2 are superposed on the original data.	148
5.4	A hypothetical plot of misclassification error rates for both training and test data as a function of tree complexity (e.g., number of leaves in the tree).	149
5.5	A diagram of a simple multilayer perceptron (or neural network) model with two hidden nodes ($d_1 = 2$) and a single output node ($d_2 = 1$).	153
5.6	An example of the type of decision boundaries that a neural network model would produce for the two-dimensional wine data of figure 5.2(a).	155
6.1	Fitting of different models to simulated data from a third-order polynomial.	171
6.2	An example of a piecewise linear fit to the data of figure 6.1 with $k = 5$ linear segments.	173
6.3	Nitrous oxide (NO _x) as a function of ethanol (E) using kernel regression with triangular kernels.	177
6.4	An example of linear decision boundaries for the two-dimensional wine classification data set of chapter 5 (See figure 5.1).	181
6.5	An example of <i>piecewise linear</i> decision boundaries for the two-dimensional wine classification data set of chapter 5 (see figure 5.1).	182
6.6	Scatter plots and contour plots of simulated data from a bivariate mixture of three Normal distributions.	186
6.7	A graphical model structure corresponding to a first-order Markov assumption.	190
6.8	A plausible graphical model structure for two variables education and baldness that are conditionally independent given age.	191
6.9	The graphical model structure for a problem with several variables demonstrating conditional independence.	192
6.10	The first-order Bayes graphical model structure, with a single class Y and 6 conditionally independent feature variables X_1, \dots, X_6 .	192
6.11	A graphical model structure corresponding to a first-order hidden Markov assumption.	200

7.1	Classification accuracy of the best model selected on a validation data set from a set of K models, $1 \leq K \leq 100$, where each model is making random predictions.	230
8.1	An example of a simple state-space involving four variables X_1, X_2, X_3, X_4 .	242
8.2	An example of a simple search tree for the state-space of figure 8.1.	245
8.3	An example of a score function $S(\theta)$ of a single univariate parameter θ with both a global minimum and a local minimum.	252
8.4	An example of a situation in which we minimize a score function of two variables and the shape of the score function is a parabolic “bowl.”	257
9.1	Illustration of the density contours for a two-dimensional Normal density function.	276
9.2	The log-likelihood of the red-blood cell data under a two-component Normal mixture model (see figure 9.11) as a function of iteration number.	283
9.3	Density estimates for the variable ethanol (E) using a histogram (top left) and Gaussian kernel estimates with three different bandwidths.	286
9.4	Scatter plot of antenna data.	304
9.5	Example of running the K -means algorithm on the two-dimensional antenna data.	305
9.6	A summary of the trajectories of the three cluster means during the K -means iterations of figure 9.5.	306
9.7	Duration of eruptions versus waiting time between eruptions (in minutes) for the Old Faithful geyser in Yellowstone Park.	309
9.8	Dendrogram resulting from clustering of data in figure 9.7.	310
9.9	Dendrogram of the single link method applied to the data in figure 9.7.	314
9.10	Red blood cell measurements (mean volume and mean hemoglobin concentration) from 182 individuals.	316
9.11	Example of running the EM algorithm on the red blood cell measurements of figure 9.10.	317
9.12	Log-likelihood and BIC score as a function of the number of Normal components fitted to the red blood cell data of figure 9.11.	320

10.1	A simple example illustrating posterior class probabilities for a two-class one-dimensional classification problem.	333
10.2	Posterior probability contours for $p(c_1 \mathbf{x})$ where c_1 is the label for the healthy class for the red blood cell data discussed in chapter 9.	336
10.3	Posterior probability contours for $p(c_1 \mathbf{x})$ where c_1 is the label for the diabetic class for the Pima Indians data of chapter 3.	337
10.4	Decision boundary produced by the Fisher linear discriminant applied to the red blood cell data from chapter 9.	342
10.5	Decision boundary for a decision tree for the red blood cell data from chapter 9.	348
10.6	An illustration of the potential pitfalls of using principal component analysis as a preprocessor for classification.	364
11.1	Expired ventilation plotted against oxygen uptake in a series of trials, with fitted straight line.	372
11.2	The data from figure 11.1 with a model that includes a term in x^2 .	373
11.3	A plot of record time (in minutes) against distance (in miles) for 35 Scottish hill races from 1984.	375
11.4	Temperature (degrees F) against latitude (degrees N) for 56 cities in the United States.	381
11.5	A plot of tensile strength of paper against the percentage of hardwood in the pulp.	383
11.6	Number of O-rings damaged (vertical axis) against temperature on day of flight, (a) data examined before the flight, and (b) the complete data.	387
11.7	The transformation function of Log(dose) in the model for predicting time for blood pressure to revert to normal.	395
11.8	The transformation function of blood pressure during administration in the model for predicting time for blood pressure to revert to normal.	396
12.1	Representing market basket data as a table with an attribute for each product.	406
12.2	A more realistic representation of market basket data.	407
12.3	Representing prices of products.	407
12.4	Representing the hierarchy of products as a table.	408

12.5	The concept of data mining algorithms which operate on an approximate version of the full data set.	424
13.1	An artificial example of basket data.	430
13.2	Episodes α , β , and γ .	437
14.1	A simple (synthetic) example of precision-recall curves for three hypothetical query algorithms.	455
14.2	Pairwise document distances for the toy document-term matrix in the text.	460
14.3	Projected locations of the 10 documents (from table 14.2) in the two dimensional plane spanned by the first two principal components of the document-term matrix M .	467

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 practitioners 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 environments 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"!