

Contents

Preface	xi
Acknowledgments	xiii
List of Figures	xv
List of Tables	xix
1 Introduction	1
1.1 How to Read This Book	2
1.2 Reproducibility	3
I R and Data Mining	5
2 Introduction to R	7
2.1 Starting with R	7
2.2 Basic Interaction with the R Console	9
2.3 R Objects and Variables	10
2.4 R Functions	12
2.5 Vectors	16
2.6 Vectorization	18
2.7 Factors	19
2.8 Generating Sequences	22
2.9 Sub-Setting	24
2.10 Matrices and Arrays	26
2.11 Lists	30
2.12 Data Frames	32
2.13 Useful Extensions to Data Frames	36
2.14 Objects, Classes, and Methods	40
2.15 Managing Your Sessions	41
3 Introduction to Data Mining	43
3.1 A Bird's Eye View on Data Mining	43
3.2 Data Collection and Business Understanding	45
3.2.1 Data and Datasets	45
3.2.2 Importing Data into R	46
3.2.2.1 Text Files	47
3.2.2.2 Databases	49
3.2.2.3 Spreadsheets	52

3.2.2.4	Other Formats	52
3.3	Data Pre-Processing	53
3.3.1	Data Cleaning	53
3.3.1.1	Tidy Data	53
3.3.1.2	Handling Dates	56
3.3.1.3	String Processing	58
3.3.1.4	Dealing with Unknown Values	60
3.3.2	Transforming Variables	62
3.3.2.1	Handling Different Scales of Variables	62
3.3.2.2	Discretizing Variables	63
3.3.3	Creating Variables	65
3.3.3.1	Handling Case Dependencies	65
3.3.3.2	Handling Text Datasets	74
3.3.4	Dimensionality Reduction	78
3.3.4.1	Sampling Rows	78
3.3.4.2	Variable Selection	82
3.4	Modeling	87
3.4.1	Exploratory Data Analysis	87
3.4.1.1	Data Summarization	87
3.4.1.2	Data Visualization	96
3.4.2	Dependency Modeling using Association Rules	110
3.4.3	Clustering	119
3.4.3.1	Measures of Dissimilarity	119
3.4.3.2	Clustering Methods	120
3.4.4	Anomaly Detection	131
3.4.4.1	Univariate Outlier Detection Methods	132
3.4.4.2	Multi-Variate Outlier Detection Methods	133
3.4.5	Predictive Analytics	140
3.4.5.1	Evaluation Metrics	141
3.4.5.2	Tree-Based Models	145
3.4.5.3	Support Vector Machines	151
3.4.5.4	Artificial Neural Networks and Deep Learning	158
3.4.5.5	Model Ensembles	165
3.5	Evaluation	172
3.5.1	The Holdout and Random Subsampling	174
3.5.2	Cross Validation	177
3.5.3	Bootstrap Estimates	179
3.5.4	Recommended Procedures	181
3.6	Reporting and Deployment	182
3.6.1	Reporting Through Dynamic Documents	183
3.6.2	Deployment through Web Applications	186

II Case Studies

191

4 Predicting Algae Blooms

193

4.1	Problem Description and Objectives	193
4.2	Data Description	194
4.3	Loading the Data into R	194
4.4	Data Visualization and Summarization	196
4.5	Unknown Values	205

4.5.1	Removing the Observations with Unknown Values	205
4.5.2	Filling in the Unknowns with the Most Frequent Values	207
4.5.3	Filling in the Unknown Values by Exploring Correlations	208
4.5.4	Filling in the Unknown Values by Exploring Similarities between Cases	212
4.6	Obtaining Prediction Models	214
4.6.1	Multiple Linear Regression	215
4.6.2	Regression Trees	220
4.7	Model Evaluation and Selection	225
4.8	Predictions for the Seven Algae	237
4.9	Summary	239
5	Predicting Stock Market Returns	241
5.1	Problem Description and Objectives	241
5.2	The Available Data	242
5.2.1	Reading the Data from the CSV File	243
5.2.2	Getting the Data from the Web	243
5.3	Defining the Prediction Tasks	244
5.3.1	What to Predict?	244
5.3.2	Which Predictors?	247
5.3.3	The Prediction Tasks	251
5.3.4	Evaluation Criteria	252
5.4	The Prediction Models	254
5.4.1	How Will the Training Data Be Used?	254
5.4.2	The Modeling Tools	256
5.4.2.1	Artificial Neural Networks	256
5.4.2.2	Support Vector Machines	259
5.4.2.3	Multivariate Adaptive Regression Splines	260
5.5	From Predictions into Actions	263
5.5.1	How Will the Predictions Be Used?	263
5.5.2	Trading-Related Evaluation Criteria	264
5.5.3	Putting Everything Together: A Simulated Trader	265
5.6	Model Evaluation and Selection	271
5.6.1	Monte Carlo Estimates	271
5.6.2	Experimental Comparisons	272
5.6.3	Results Analysis	278
5.7	The Trading System	286
5.7.1	Evaluation of the Final Test Data	286
5.7.2	An Online Trading System	291
5.8	Summary	292
6	Detecting Fraudulent Transactions	295
6.1	Problem Description and Objectives	295
6.2	The Available Data	296
6.2.1	Loading the Data into R	296
6.2.2	Exploring the Dataset	297
6.2.3	Data Problems	304
6.2.3.1	Unknown Values	304
6.2.3.2	Few Transactions of Some Products	309

6.3	Defining the Data Mining Tasks	313
6.3.1	Different Approaches to the Problem	313
6.3.1.1	Unsupervised Techniques	313
6.3.1.2	Supervised Techniques	314
6.3.1.3	Semi-Supervised Techniques	315
6.3.2	Evaluation Criteria	316
6.3.2.1	Precision and Recall	316
6.3.2.2	Lift Charts and Precision/Recall Curves	317
6.3.2.3	Normalized Distance to Typical Price	320
6.3.3	Experimental Methodology	321
6.4	Obtaining Outlier Rankings	323
6.4.1	Unsupervised Approaches	323
6.4.1.1	The Modified Box Plot Rule	323
6.4.1.2	Local Outlier Factors (<i>LOF</i>)	327
6.4.1.3	Clustering-Based Outlier Rankings (<i>OR_h</i>)	330
6.4.2	Supervised Approaches	332
6.4.2.1	The Class Imbalance Problem	333
6.4.2.2	Naive Bayes	335
6.4.2.3	AdaBoost	339
6.4.3	Semi-Supervised Approaches	344
6.5	Summary	350
7	Classifying Microarray Samples	353
7.1	Problem Description and Objectives	353
7.1.1	Brief Background on Microarray Experiments	353
7.1.2	The ALL Dataset	354
7.2	The Available Data	354
7.2.1	Exploring the Dataset	357
7.3	Gene (Feature) Selection	359
7.3.1	Simple Filters Based on Distribution Properties	360
7.3.2	ANOVA Filters	362
7.3.3	Filtering Using Random Forests	364
7.3.4	Filtering Using Feature Clustering Ensembles	367
7.4	Predicting Cytogenetic Abnormalities	368
7.4.1	Defining the Prediction Task	368
7.4.2	The Evaluation Metric	369
7.4.3	The Experimental Procedure	369
7.4.4	The Modeling Techniques	370
7.4.5	Comparing the Models	373
7.5	Summary	381
	Bibliography	383
	Subject Index	395
	Index of Data Mining Topics	399
	Index of R Functions	401