"Data Mining: Practical Machine Learning Tools and Techniques"
13 assignments: from chapter 1 to 5

LIST OF FIGURES xv
LIST OF TABLES xix
PREFACE xxi
Updated and Revised Content xxv
Second Edition xxv
Third Edition xxvi
ACKNOWLEDGMENTS xxix
ABOUT THE AUTHORS xxxiii

PART I INTRODUCTION TO DATA MINING

CHAPTER 1 What's It All About? 3
1.1 Data Mining and Machine Learning 3
Describing Structural Patterns 5
Machine Learning 7
Data Mining 8
assign 1 **1.2** Simple Examples: The Weather Problem and Others 9
The Weather Problem 9
Contact Lenses: An Idealized Problem 12
Irises: A Classic Numeric Dataset 13
CPU Performance: Introducing Numeric Prediction 15
Labor Negotiations: A More Realistic Example 15
Soybean Classification: A Classic Machine Learning Success 19
1.3 Fielded Applications 21
Web Mining 21
Decisions Involving Judgment 22
Screening Images 23
Load Forecasting 24
assign 2 Diagnosis 25
Marketing and Sales 26
Other Applications 27
1.4 Machine Learning and Statistics 28
1.5 Generalization as Search 29
1.6 Data Mining and Ethics 33
Reidentification 33
Using Personal Information 34
Wider Issues 35
1.7 Further Reading 36

assign 3

CHAPTER 2 Input: Concepts, Instances, and Attributes 39
2.1 What's a Concept? 40
2.2 What's in an Example? 42
Relations 43
Other Example Types 46
2.3 What's in an Attribute? 49

assign 4

2.4 Preparing the Input 51
Gathering the Data Together 51
ARFF Format 52
Sparse Data 56
Attribute Types 56
Missing Values 58
Inaccurate Values 59
Getting to Know Your Data 60
2.5 Further Reading 60

CHAPTER 3 Output: Knowledge Representation 61
3.1 Tables 61

assign 5

3.2 Linear Models 62
3.3 Trees 64
3.4 Rules 67
Classification Rules 69
Association Rules 72

assign 6

Rules with Exceptions 73
More Expressive Rules 75
3.5 Instance-Based Representation 78
3.6 Clusters 81
3.7 Further Reading 83

CHAPTER 4 Algorithms: The Basic Methods 85
4.1 Inferring Rudimentary Rules 86

assign 7

Missing Values and Numeric Attributes 87
Discussion 89
4.2 Statistical Modeling 90
Missing Values and Numeric Attributes 94
Naïve Bayes for Document Classification 97
Discussion 99
4.3 Divide-and-Conquer: Constructing Decision Trees 99
Calculating Information 103
Highly Branching Attributes 105
Discussion 107

assign 8

(cont. to section 4.4)

assign 8

4.4 Covering Algorithms: Constructing Rules 108
Rules versus Trees 109
A Simple Covering Algorithm 110
Rules versus Decision Lists 115

assign 9

4.5 Mining Association Rules 116
Item Sets 116
Association Rules 119
Generating Rules Efficiently 122
Discussion 123
4.6 Linear Models 124
Numeric Prediction: Linear Regression 124
Linear Classification: Logistic Regression 125
Linear Classification Using the Perceptron 127
Linear Classification Using Winnow 129

assign 10

4.7 Instance-Based Learning 131
Distance Function 131
Finding Nearest Neighbors Efficiently 132
Discussion 137
4.8 Clustering 138
Iterative Distance-Based Clustering 139
Faster Distance Calculations 139
Discussion 141
4.9 Multi-Instance Learning 141
Aggregating the Input 142
Aggregating the Output 142
Discussion 142
4.10 Further Reading 143
4.11 Weka Implementations 145

assign 11

CHAPTER 5 Credibility: Evaluating What's Been Learned 147
5.1 Training and Testing 148
5.2 Predicting Performance 150
5.3 Cross-Validation 152
5.4 Other Estimates 154
Leave-One-Out Cross-Validation 154
The Bootstrap 155
5.5 Comparing Data Mining Schemes 156
5.6 Predicting Probabilities 159
Quadratic Loss Function 160
Informational Loss Function 161
Discussion 162

assign 12

5.7 Counting the Cost 163
Cost-Sensitive Classification 166
Cost-Sensitive Learning 167
Lift Charts 168
ROC Curves 172
Recall–Precision Curves 174
Discussion 175
Cost Curves 177

assign 13

5.8 Evaluating Numeric Prediction 180
5.9 Minimum Description Length Principle 183
5.10 Applying the MDL Principle to Clustering 186
5.11 Further Reading 187

PART II ADVANCED DATA MINING

CHAPTER 6 Implementations: Real Machine Learning Schemes 191
6.1 Decision Trees 192
Numeric Attributes 193
Missing Values 194
Pruning 195
Estimating Error Rates 197
Complexity of Decision Tree Induction 199
From Trees to Rules 200
C4.5: Choices and Options 201
Cost-Complexity Pruning 202
Discussion 202
6.2 Classification Rules 203
Criteria for Choosing Tests 203
Missing Values, Numeric Attributes 204
Generating Good Rules 205
Using Global Optimization 208
Obtaining Rules from Partial Decision Trees 208
Rules with Exceptions 212
Discussion 215
6.3 Association Rules 216
Building a Frequent-Pattern Tree 216
Finding Large Item Sets 219
Discussion 222
6.4 Extending Linear Models 223
Maximum-Margin Hyperplane 224
Nonlinear Class Boundaries 226