

Contents

1. Introduction to Data Mining	1
1.1 The Data Explosion	1
1.2 Knowledge Discovery	2
1.3 Applications of Data Mining	3
1.4 Labelled and Unlabelled Data	4
1.5 Supervised Learning: Classification	5
1.6 Supervised Learning: Numerical Prediction	7
1.7 Unsupervised Learning: Association Rules	7
1.8 Unsupervised Learning: Clustering	8
2. Data for Data Mining	9
2.1 Standard Formulation	9
2.2 Types of Variable	10
2.2.1 Categorical and Continuous Attributes	12
2.3 Data Preparation	12
2.3.1 Data Cleaning	13
2.4 Missing Values	15
2.4.1 Discard Instances	15
2.4.2 Replace by Most Frequent/Average Value	15
2.5 Reducing the Number of Attributes	16
2.6 The UCI Repository of Datasets	17
2.7 Chapter Summary	18
2.8 Self-assessment Exercises for Chapter 2	18
Reference	19

3. Introduction to Classification: Naïve Bayes and Nearest Neighbour	21
3.1 What Is Classification?	21
3.2 Naïve Bayes Classifiers	22
3.3 Nearest Neighbour Classification	29
3.3.1 Distance Measures	32
3.3.2 Normalisation	35
3.3.3 Dealing with Categorical Attributes	36
3.4 Eager and Lazy Learning	36
3.5 Chapter Summary	37
3.6 Self-assessment Exercises for Chapter 3	37
4. Using Decision Trees for Classification	39
4.1 Decision Rules and Decision Trees	39
4.1.1 Decision Trees: The Golf Example	40
4.1.2 Terminology	41
4.1.3 The <i>degrees</i> Dataset	42
4.2 The TDIDT Algorithm	45
4.3 Types of Reasoning	47
4.4 Chapter Summary	48
4.5 Self-assessment Exercises for Chapter 4	48
References	48
5. Decision Tree Induction: Using Entropy for Attribute Selection	49
5.1 Attribute Selection: An Experiment	49
5.2 Alternative Decision Trees	50
5.2.1 The Football/Netball Example	51
5.2.2 The <i>anonymous</i> Dataset	53
5.3 Choosing Attributes to Split On: Using Entropy	54
5.3.1 The <i>lens24</i> Dataset	55
5.3.2 Entropy	57
5.3.3 Using Entropy for Attribute Selection	58
5.3.4 Maximising Information Gain	60
5.4 Chapter Summary	61
5.5 Self-assessment Exercises for Chapter 5	61
6. Decision Tree Induction: Using Frequency Tables for Attribute Selection	63
6.1 Calculating Entropy in Practice	63
6.1.1 Proof of Equivalence	64
6.1.2 A Note on Zeros	66

6.2	Other Attribute Selection Criteria: Gini Index of Diversity	66
6.3	The χ^2 Attribute Selection Criterion	68
6.4	Inductive Bias	71
6.5	Using Gain Ratio for Attribute Selection	73
	6.5.1 Properties of Split Information	74
	6.5.2 Summary	75
6.6	Number of Rules Generated by Different Attribute Selection Criteria	75
6.7	Missing Branches	76
6.8	Chapter Summary	77
6.9	Self-assessment Exercises for Chapter 6	77
	References	78
7.	Estimating the Predictive Accuracy of a Classifier	79
7.1	Introduction	79
7.2	Method 1: Separate Training and Test Sets	80
	7.2.1 Standard Error	81
	7.2.2 Repeated Train and Test	82
7.3	Method 2: k -fold Cross-validation	82
7.4	Method 3: N -fold Cross-validation	83
7.5	Experimental Results I	84
7.6	Experimental Results II: Datasets with Missing Values	86
	7.6.1 Strategy 1: Discard Instances	87
	7.6.2 Strategy 2: Replace by Most Frequent/Average Value . .	87
	7.6.3 Missing Classifications	89
7.7	Confusion Matrix	89
	7.7.1 True and False Positives	90
7.8	Chapter Summary	91
7.9	Self-assessment Exercises for Chapter 7	91
	Reference	92
8.	Continuous Attributes	93
8.1	Introduction	93
8.2	Local versus Global Discretisation	95
8.3	Adding Local Discretisation to TDIDT	96
	8.3.1 Calculating the Information Gain of a Set of Pseudo- attributes	97
	8.3.2 Computational Efficiency	102
8.4	Using the ChiMerge Algorithm for Global Discretisation	105
	8.4.1 Calculating the Expected Values and χ^2	108
	8.4.2 Finding the Threshold Value	113
	8.4.3 Setting <i>minIntervals</i> and <i>maxIntervals</i>	113

8.4.4	The ChiMerge Algorithm: Summary	115
8.4.5	The ChiMerge Algorithm: Comments	115
8.5	Comparing Global and Local Discretisation for Tree Induction	116
8.6	Chapter Summary	118
8.7	Self-assessment Exercises for Chapter 8	118
	Reference	119
9.	Avoiding Overfitting of Decision Trees	121
9.1	Dealing with Clashes in a Training Set	122
9.1.1	Adapting TDIDT to Deal with Clashes	122
9.2	More About Overfitting Rules to Data	127
9.3	Pre-pruning Decision Trees	128
9.4	Post-pruning Decision Trees	130
9.5	Chapter Summary	135
9.6	Self-assessment Exercise for Chapter 9	136
	References	136
10.	More About Entropy	137
10.1	Introduction	137
10.2	Coding Information Using Bits	140
10.3	Discriminating Amongst M Values (M Not a Power of 2)	142
10.4	Encoding Values That Are Not Equally Likely	143
10.5	Entropy of a Training Set	146
10.6	Information Gain Must Be Positive or Zero	147
10.7	Using Information Gain for Feature Reduction for Classification Tasks	149
10.7.1	Example 1: The <i>genetics</i> Dataset	150
10.7.2	Example 2: The <i>bcst96</i> Dataset	154
10.8	Chapter Summary	156
10.9	Self-assessment Exercises for Chapter 10	156
	References	156
11.	Inducing Modular Rules for Classification	157
11.1	Rule Post-pruning	157
11.2	Conflict Resolution	159
11.3	Problems with Decision Trees	162
11.4	The Prism Algorithm	164
11.4.1	Changes to the Basic Prism Algorithm	171
11.4.2	Comparing Prism with TDIDT	172
11.5	Chapter Summary	173
11.6	Self-assessment Exercise for Chapter 11	173
	References	174

- 12. Measuring the Performance of a Classifier** 175
 - 12.1 True and False Positives and Negatives 176
 - 12.2 Performance Measures 178
 - 12.3 True and False Positive Rates versus Predictive Accuracy 181
 - 12.4 ROC Graphs 182
 - 12.5 ROC Curves 184
 - 12.6 Finding the Best Classifier 185
 - 12.7 Chapter Summary 186
 - 12.8 Self-assessment Exercise for Chapter 12 187

- 13. Dealing with Large Volumes of Data** 189
 - 13.1 Introduction 189
 - 13.2 Distributing Data onto Multiple Processors 192
 - 13.3 Case Study: PMCRI 194
 - 13.4 Evaluating the Effectiveness of a Distributed System: PMCRI . 197
 - 13.5 Revising a Classifier Incrementally 201
 - 13.6 Chapter Summary 207
 - 13.7 Self-assessment Exercises for Chapter 13 207
 - References 208

- 14. Ensemble Classification** 209
 - 14.1 Introduction 209
 - 14.2 Estimating the Performance of a Classifier 212
 - 14.3 Selecting a Different Training Set for Each Classifier 213
 - 14.4 Selecting a Different Set of Attributes for Each Classifier 214
 - 14.5 Combining Classifications: Alternative Voting Systems 215
 - 14.6 Parallel Ensemble Classifiers 219
 - 14.7 Chapter Summary 219
 - 14.8 Self-assessment Exercises for Chapter 14 220
 - References 220

- 15. Comparing Classifiers** 221
 - 15.1 Introduction 221
 - 15.2 The Paired t-Test 223
 - 15.3 Choosing Datasets for Comparative Evaluation 229
 - 15.3.1 Confidence Intervals 231
 - 15.4 Sampling 231
 - 15.5 How Bad Is a ‘No Significant Difference’ Result? 234
 - 15.6 Chapter Summary 235
 - 15.7 Self-assessment Exercises for Chapter 15 235
 - References 236

16. Association Rule Mining I	237
16.1 Introduction	237
16.2 Measures of Rule Interestingness	239
16.2.1 The Piatetsky-Shapiro Criteria and the RI Measure	241
16.2.2 Rule Interestingness Measures Applied to the <i>chess</i> Dataset	243
16.2.3 Using Rule Interestingness Measures for Conflict Resolution	245
16.3 Association Rule Mining Tasks	245
16.4 Finding the Best N Rules	246
16.4.1 The J -Measure: Measuring the Information Content of a Rule	247
16.4.2 Search Strategy	248
16.5 Chapter Summary	251
16.6 Self-assessment Exercises for Chapter 16	251
References	251
17. Association Rule Mining II	253
17.1 Introduction	253
17.2 Transactions and Itemsets	254
17.3 Support for an Itemset	255
17.4 Association Rules	256
17.5 Generating Association Rules	258
17.6 Apriori	259
17.7 Generating Supported Itemsets: An Example	262
17.8 Generating Rules for a Supported Itemset	264
17.9 Rule Interestingness Measures: Lift and Leverage	266
17.10 Chapter Summary	268
17.11 Self-assessment Exercises for Chapter 17	269
Reference	269
18. Association Rule Mining III: Frequent Pattern Trees	271
18.1 Introduction: FP-Growth	271
18.2 Constructing the FP-tree	274
18.2.1 Pre-processing the Transaction Database	274
18.2.2 Initialisation	276
18.2.3 Processing Transaction 1: f, c, a, m, p	277
18.2.4 Processing Transaction 2: f, c, a, b, m	279
18.2.5 Processing Transaction 3: f, b	283
18.2.6 Processing Transaction 4: c, b, p	285
18.2.7 Processing Transaction 5: f, c, a, m, p	287
18.3 Finding the Frequent Itemsets from the FP-tree	288

- 18.3.1 Itemsets Ending with Item p 291
- 18.3.2 Itemsets Ending with Item m 301
- 18.4 Chapter Summary 308
- 18.5 Self-assessment Exercises for Chapter 18 309
- Reference 309

- 19. Clustering** 311
 - 19.1 Introduction 311
 - 19.2 k -Means Clustering 314
 - 19.2.1 Example 315
 - 19.2.2 Finding the Best Set of Clusters 319
 - 19.3 Agglomerative Hierarchical Clustering 320
 - 19.3.1 Recording the Distance Between Clusters 323
 - 19.3.2 Terminating the Clustering Process 326
 - 19.4 Chapter Summary 327
 - 19.5 Self-assessment Exercises for Chapter 19 327

- 20. Text Mining** 329
 - 20.1 Multiple Classifications 329
 - 20.2 Representing Text Documents for Data Mining 330
 - 20.3 Stop Words and Stemming 332
 - 20.4 Using Information Gain for Feature Reduction 333
 - 20.5 Representing Text Documents: Constructing a Vector Space Model 333
 - 20.6 Normalising the Weights 335
 - 20.7 Measuring the Distance Between Two Vectors 336
 - 20.8 Measuring the Performance of a Text Classifier 337
 - 20.9 Hypertext Categorisation 338
 - 20.9.1 Classifying Web Pages 338
 - 20.9.2 Hypertext Classification versus Text Classification 339
 - 20.10 Chapter Summary 343
 - 20.11 Self-assessment Exercises for Chapter 20 343

- A. Essential Mathematics** 345
 - A.1 Subscript Notation 345
 - A.1.1 Sigma Notation for Summation 346
 - A.1.2 Double Subscript Notation 347
 - A.1.3 Other Uses of Subscripts 348
 - A.2 Trees 348
 - A.2.1 Terminology 349
 - A.2.2 Interpretation 350
 - A.2.3 Subtrees 351

A.3	The Logarithm Function $\log_2 X$	351
A.3.1	The Function $-X \log_2 X$	354
A.4	Introduction to Set Theory	355
A.4.1	Subsets.....	357
A.4.2	Summary of Set Notation.....	359
B.	Datasets	361
	References	381
C.	Sources of Further Information	383
	Websites	383
	Books	383
	Books on Neural Nets	384
	Conferences	385
	Information About Association Rule Mining	385
D.	Glossary and Notation	387
E.	Solutions to Self-assessment Exercises	407
Index	435



<http://www.springer.com/978-1-4471-4883-8>

Principles of Data Mining

Bramer, M.

2013, XIV, 440 p. 101 illus., Softcover

ISBN: 978-1-4471-4883-8