

Table of contents :

Principles of Data Mining.....	Page 4
About This Book.....	Page 6
Contents.....	Page 8
1.1 The Data Explosion.....	Page 18
1.2 Knowledge Discovery.....	Page 19
1.3 Applications of Data Mining.....	Page 20
1.4 Labelled and Unlabelled Data.....	Page 21
1.5 Supervised Learning: Classification.....	Page 22
1.7 Unsupervised Learning: Association Rules.....	Page 24
1.8 Unsupervised Learning: Clustering.....	Page 25
2.1 Standard Formulation.....	Page 26
2.2 Types of Variable.....	Page 27
2.3 Data Preparation.....	Page 29
2.3.1 Data Cleaning.....	Page 30
2.4.2 Replace by Most Frequent/Average Value.....	Page 32
2.5 Reducing the Number of Attributes.....	Page 33
2.6 The UCI Repository of Datasets.....	Page 34
2.8 Self-assessment Exercises for Chapter 2.....	Page 35
Reference.....	Page 36
3.1 What Is Classification?.....	Page 37
3.2 Naïve Bayes Classifiers.....	Page 38
3.3 Nearest Neighbour Classification.....	Page 45
3.3.1 Distance Measures.....	Page 48
3.3.2 Normalisation.....	Page 51
3.4 Eager and Lazy Learning.....	Page 52
3.6 Self-assessment Exercises for Chapter 3.....	Page 53
4.1 Decision Rules and Decision Trees.....	Page 54
4.1.1 Decision Trees: The Golf Example.....	Page 55
4.1.2 Terminology.....	Page 56
4.1.3 The degrees Dataset.....	Page 57
4.2 The TDIDT Algorithm.....	Page 60
4.3 Types of Reasoning.....	Page 62
References.....	Page 63
5.1 Attribute Selection: An Experiment.....	Page 64
5.2 Alternative Decision Trees.....	Page 65
5.2.1 The Football/Netball Example.....	Page 66
5.2.2 The anonymous Dataset.....	Page 68
5.3 Choosing Attributes to Split On: Using Entropy.....	Page 69
5.3.1 The lens24 Dataset.....	Page 70
5.3.2 Entropy.....	Page 72
5.3.3 Using Entropy for Attribute Selection.....	Page 73
5.3.4 Maximising Information Gain.....	Page 75
5.5 Self-assessment Exercises for Chapter 5.....	Page 76
6.1 Calculating Entropy in Practice.....	Page 78
6.1.1 Proof of Equivalence.....	Page 79
6.2 Other Attribute Selection Criteria: Gini Index of Diversity.....	Page 81
6.3 The chi <sup>2</sup> Attribute Selection Criterion.....	Page 83

6.4 Inductive Bias.....	Page 86
6.5 Using Gain Ratio for Attribute Selection.....	Page 88
6.5.1 Properties of Split Information.....	Page 89
6.6 Number of Rules Generated by Different Attribute Selection Criteria.....	Page 90
6.7 Missing Branches.....	Page 91
6.9 Self-assessment Exercises for Chapter 6.....	Page 92
References.....	Page 93
7.1 Introduction.....	Page 94
7.2 Method 1: Separate Training and Test Sets.....	Page 95
7.2.1 Standard Error.....	Page 96
7.3 Method 2: k-fold Cross-validation.....	Page 97
7.4 Method 3: N-fold Cross-validation.....	Page 98
7.5 Experimental Results I.....	Page 99
7.6 Experimental Results II: Datasets with Missing Values.....	Page 101
7.6.2 Strategy 2: Replace by Most Frequent/Average Value.....	Page 102
7.7 Confusion Matrix.....	Page 104
7.7.1 True and False Positives.....	Page 105
7.9 Self-assessment Exercises for Chapter 7.....	Page 106
Reference.....	Page 107
8.1 Introduction.....	Page 108
8.2 Local versus Global Discretisation.....	Page 110
8.3 Adding Local Discretisation to TDIDT.....	Page 111
8.3.1 Calculating the Information Gain of a Set of Pseudo-attributes.....	Page 112
8.3.2 Computational Efficiency.....	Page 117
8.4 Using the ChiMerge Algorithm for Global Discretisation.....	Page 120
8.4.1 Calculating the Expected Values and $\chi^2$ .....	Page 123
8.4.3 Setting minIntervals and maxIntervals.....	Page 128
8.4.5 The ChiMerge Algorithm: Comments.....	Page 130
8.5 Comparing Global and Local Discretisation for Tree Induction.....	Page 131
8.7 Self-assessment Exercises for Chapter 8.....	Page 133
Reference.....	Page 134
9. Avoiding Overfitting of Decision Trees.....	Page 135
9.1.1 Adapting TDIDT to Deal with Clashes.....	Page 136
9.2 More About Overfitting Rules to Data.....	Page 141
9.3 Pre-pruning Decision Trees.....	Page 142
9.4 Post-pruning Decision Trees.....	Page 144
9.5 Chapter Summary.....	Page 149
References.....	Page 150
10.1 Introduction.....	Page 151
10.2 Coding Information Using Bits.....	Page 154
10.3 Discriminating Amongst M Values (M Not a Power of 2).....	Page 156
10.4 Encoding Values That Are Not Equally Likely.....	Page 157
10.5 Entropy of a Training Set.....	Page 160
10.6 Information Gain Must Be Positive or Zero.....	Page 161
10.7 Using Information Gain for Feature Reduction for Classification Tasks.....	Page 163
10.7.1 Example 1: The genetics Dataset.....	Page 164
10.7.2 Example 2: The bcst96 Dataset.....	Page 168
References.....	Page 170

11.1 Rule Post-pruning.....	Page 171
11.2 Conflict Resolution.....	Page 173
11.3 Problems with Decision Trees.....	Page 176
11.4 The Prism Algorithm.....	Page 178
11.4.1 Changes to the Basic Prism Algorithm.....	Page 185
11.4.2 Comparing Prism with TDIDT.....	Page 186
11.6 Self-assessment Exercise for Chapter 11.....	Page 187
References.....	Page 188
12. Measuring the Performance of a Classifier.....	Page 189
12.1 True and False Positives and Negatives.....	Page 190
12.2 Performance Measures.....	Page 192
12.3 True and False Positive Rates versus Predictive Accuracy.....	Page 195
12.4 ROC Graphs.....	Page 196
12.5 ROC Curves.....	Page 198
12.6 Finding the Best Classifier.....	Page 199
12.7 Chapter Summary.....	Page 200
12.8 Self-assessment Exercise for Chapter 12.....	Page 201
13.1 Introduction.....	Page 202
13.2 Distributing Data onto Multiple Processors.....	Page 205
13.3 Case Study: PMCRI.....	Page 207
13.4 Evaluating the Effectiveness of a Distributed System: PMCRI.....	Page 210
13.5 Revising a Classifier Incrementally.....	Page 214
13.7 Self-assessment Exercises for Chapter 13.....	Page 220
References.....	Page 221
14.1 Introduction.....	Page 222
14.2 Estimating the Performance of a Classifier.....	Page 225
14.3 Selecting a Different Training Set for Each Classifier.....	Page 226
14.4 Selecting a Different Set of Attributes for Each Classifier.....	Page 227
14.5 Combining Classifications: Alternative Voting Systems.....	Page 228
14.7 Chapter Summary.....	Page 232
References.....	Page 233
15.1 Introduction.....	Page 234
15.2 The Paired t-Test.....	Page 236
15.3 Choosing Datasets for Comparative Evaluation.....	Page 242
15.4 Sampling.....	Page 244
15.5 How Bad Is a 'No Significant Difference' Result?.....	Page 247
15.7 Self-assessment Exercises for Chapter 15.....	Page 248
References.....	Page 249
16.1 Introduction.....	Page 250
16.2 Measures of Rule Interestingness.....	Page 252
16.2.1 The Piatetsky-Shapiro Criteria and the RI Measure.....	Page 254
16.2.2 Rule Interestingness Measures Applied to the chess Dataset.....	Page 256
16.3 Association Rule Mining Tasks.....	Page 258
16.4 Finding the Best N Rules.....	Page 259
16.4.1 The J-Measure: Measuring the Information Content of a Rule.....	Page 260
16.4.2 Search Strategy.....	Page 261
References.....	Page 264
17.1 Introduction.....	Page 265

17.2 Transactions and Itemsets.....	Page 266
17.3 Support for an Itemset.....	Page 267
17.4 Association Rules.....	Page 268
17.5 Generating Association Rules.....	Page 270
17.6 Apriori.....	Page 271
17.7 Generating Supported Itemsets: An Example.....	Page 274
17.8 Generating Rules for a Supported Itemset.....	Page 276
17.9 Rule Interestingness Measures: Lift and Leverage.....	Page 278
17.10 Chapter Summary.....	Page 280
Reference.....	Page 281
18.1 Introduction: FP-Growth.....	Page 282
18.2.1 Pre-processing the Transaction Database.....	Page 285
18.2.2 Initialisation.....	Page 287
18.2.3 Processing Transaction 1: f, c, a, m, p.....	Page 288
18.2.4 Processing Transaction 2: f, c, a, b, m.....	Page 290
18.2.5 Processing Transaction 3: f, b.....	Page 294
18.2.6 Processing Transaction 4: c, b, p.....	Page 296
18.2.7 Processing Transaction 5: f, c, a, m, p.....	Page 298
18.3 Finding the Frequent Itemsets from the FP-tree.....	Page 299
18.3.1 Itemsets Ending with Item p.....	Page 302
18.3.2 Itemsets Ending with Item m.....	Page 312
18.4 Chapter Summary.....	Page 319
Reference.....	Page 320
19.1 Introduction.....	Page 321
19.2 k-Means Clustering.....	Page 324
19.2.1 Example.....	Page 325
19.2.2 Finding the Best Set of Clusters.....	Page 329
19.3 Agglomerative Hierarchical Clustering.....	Page 330
19.3.1 Recording the Distance Between Clusters.....	Page 333
19.3.2 Terminating the Clustering Process.....	Page 336
19.5 Self-assessment Exercises for Chapter 19.....	Page 337
20.1 Multiple Classifications.....	Page 339
20.2 Representing Text Documents for Data Mining.....	Page 340
20.3 Stop Words and Stemming.....	Page 342
20.5 Representing Text Documents: Constructing a Vector Space Model.....	Page 343
20.6 Normalising the Weights.....	Page 345
20.7 Measuring the Distance Between Two Vectors.....	Page 346
20.8 Measuring the Performance of a Text Classifier.....	Page 347
20.9.1 Classifying Web Pages.....	Page 348
20.9.2 Hypertext Classification versus Text Classification.....	Page 349
20.11 Self-assessment Exercises for Chapter 20.....	Page 353
21.1 Introduction.....	Page 354
21.2 Building an H-Tree: Updating Arrays.....	Page 356
21.2.1 Array currentAtts.....	Page 357
21.2.3 Sorting a record to the appropriate leaf node.....	Page 358
21.2.6 Array acvCounts.....	Page 359
21.3.1 Step (a): Initialise Root Node 0.....	Page 361
21.3.2 Step (b): Begin Reading Records.....	Page 362

21.3.3 Step (c): Consider Splitting at Node 0.....	Page 363
21.3.4 Step (d): Split on Root Node and Initialise New Leaf Nodes.....	Page 364
21.3.5 Step (e): Process the Next Set of Records.....	Page 366
21.3.6 Step (f): Consider Splitting at Node 2.....	Page 367
21.3.7 Step (g): Process the Next Set of Records.....	Page 368
21.3.8 Outline of the H-Tree Algorithm.....	Page 369
21.4 Splitting on an Attribute: Using Information Gain.....	Page 372
21.5 Splitting on An Attribute: Using a Hoeffding Bound.....	Page 374
21.6 H-Tree Algorithm: Final Version.....	Page 379
21.7 Using an Evolving H-Tree to Make Predictions.....	Page 381
21.7.1 Evaluating the Performance of an H-Tree.....	Page 382
21.8.1 The lens24 Dataset.....	Page 383
21.8.2 The vote Dataset.....	Page 385
21.10 Self-assessment Exercises for Chapter 21.....	Page 386
References.....	Page 387
22.1 Stationary versus Time-dependent Data.....	Page 388
22.2 Summary of the H-Tree Algorithm.....	Page 390
22.2.1 Array currentAtts.....	Page 391
22.2.4 Array classtotals.....	Page 392
22.2.7 Pseudocode for the H-Tree Algorithm.....	Page 393
22.4 From H-Tree to CDH-Tree: Incrementing Counts.....	Page 396
22.5 The Sliding Window Method.....	Page 397
22.6 Resplitting at Nodes.....	Page 402
22.7 Identifying Suspect Nodes.....	Page 403
22.8 Creating Alternate Nodes.....	Page 405
22.9 Growing/Forgetting an Alternate Node and its Descendants.....	Page 409
22.10 Replacing an Internal Node by One of its Alternate Nodes.....	Page 411
22.11 Experiment: Tracking Concept Drift.....	Page 419
22.11.1 lens24 Data: Alternative Mode.....	Page 421
22.11.2 Introducing Concept Drift.....	Page 423
22.11.3 An Experiment with Alternating lens24 Data.....	Page 424
22.11.4 Comments on Experiment.....	Page 432
22.13 Self-assessment Exercises for Chapter 22.....	Page 433
Reference.....	Page 434
23.1 Introduction.....	Page 435
23.2 Neural Nets Example 1.....	Page 438
23.3 Neural Nets Example 2.....	Page 442
23.3.1 Forward Propagating the Values of the Input Nodes.....	Page 445
23.3.2 Forward Propagation: Summary of Formulae.....	Page 450
23.4.1 Stochastic Gradient Descent.....	Page 451
23.4.2 Finding the Gradients.....	Page 453
23.4.3 Working backwards from the output layer to the hidden layer.....	Page 455
23.4.4 Working backwards from the hidden layer to the input layer.....	Page 457
23.4.5 Updating the Weights.....	Page 460
23.5 Processing a Multi-instance Training Set.....	Page 463
23.6 Using a Neural Net for Classification: the iris Dataset.....	Page 464
23.7 Using a Neural Net for Classification: the seeds Dataset.....	Page 469
23.8 Neural Nets: A Note of Caution.....	Page 471

23.9 Chapter Summary.....	Page 473
23.10 Self-assessment Exercises for Chapter 23.....	Page 474
A.1 Subscript Notation.....	Page 475
A.1.1 Sigma Notation for Summation.....	Page 476
A.1.2 Double Subscript Notation.....	Page 477
A.2 Trees.....	Page 478
A.2.1 Terminology.....	Page 479
A.2.2 Interpretation.....	Page 480
A.3 The Logarithm Function $\log_2 X$ .....	Page 481
A.3.1 The Function $-X \log_2 X$ .....	Page 484
A.4 Introduction to Set Theory.....	Page 485
A.4.1 Subsets.....	Page 487
A.4.2 Summary of Set Notation.....	Page 489
B. Datasets.....	Page 490
References.....	Page 511
Books.....	Page 512
Conferences.....	Page 513
Information About Association Rule Mining.....	Page 514
D. Glossary and Notation.....	Page 515
E. Solutions to Self-assessment Exercises.....	Page 540
Index.....	Page 570