

Contents

Preface page xi

1 Introduction 1

- 1.1 Background 1
- 1.2 Book structure 2
- 1.3 Classification 2
- 1.4 Clustering 3
- 1.5 Structures in data 3
- 1.6 Glossary 5
- 1.7 Recommended reading and other resources 10

2 Exploratory data analysis 12

- 2.1 Background 12
- 2.2 Dimensionality 13
- 2.3 Goodness of fit testing 14
- 2.4 Graphical methods 15
- 2.5 Variance-based data projections 16
- 2.6 Distance-based data projections 29
- 2.7 Other projection methods 32
- 2.8 Other methods 36
- 2.9 Data dredging 38
- 2.10 Example EDA analysis 38

3 Cluster analysis 46

- 3.1 Background 46
- 3.2 Distance and similarity measures 48
- 3.3 Partitioning methods 55
- 3.4 Agglomerative hierarchical methods 58
- 3.5 How many groups are there? 62
- 3.6 Divisive hierarchical methods 65

3.7	Two-way clustering and gene shaving	66
3.8	Recommended reading	67
3.9	Example analyses	68
4	Introduction to classification	78
4.1	Background	78
4.2	Black-box classifiers	81
4.3	Nature of a classifier	82
4.4	No-free-lunch	85
4.5	Bias and variance	86
4.6	Variable (feature) selection	87
4.7	Multiple classifiers	92
4.8	Why do classifiers fail?	94
4.9	Generalisation	95
4.10	Types of classifier	96
5	Classification algorithms 1	97
5.1	Background	97
5.2	Naïve Bayes	99
5.3	Discriminant analysis	100
5.4	Logistic regression	117
5.5	Discriminant analysis or logistic regression?	128
5.6	Generalised additive models	130
5.7	Summary	136
6	Other classification methods	137
6.1	Background	137
6.2	Decision trees	137
6.3	Support vector machines	154
6.4	Artificial neural networks	156
6.5	Genetic algorithms	170
6.6	Others	175
6.7	Where next?	177
7	Classification accuracy	179
7.1	Background	179
7.2	Appropriate metrics	180
7.3	Binary accuracy measures	180
7.4	Appropriate testing data	183
7.5	Decision thresholds	186
7.6	Example	187
7.7	ROC plots	190
7.8	Incorporating costs	194
7.9	Comparing classifiers	196
7.10	Recommended reading	199

Appendix A	200
Appendix B	203
Appendix C	207
Appendix D	208
Appendix E	210
Appendix F	217
Appendix G	220

References	224
------------	-----

Index	241
-------	-----