WILEY

Applied Data Mining

Statistical Methods for Business and Industry

Paolo Giudici

Contents

Preface **xi**

1 Introduction **1**
- 1.1 What is data mining? 1
  - 1.1.1 Data mining and computing 3
  - 1.1.2 Data mining and statistics 5
- 1.2 The data mining process 6
- 1.3 Software for data mining 11
- 1.4 Organisation of the book 12
  - 1.4.1 Chapters 2 to 6: methodology 13
  - 1.4.2 Chapters 7 to 12: business cases 13
- 1.5 Further reading 14

Part I Methodology **17**

2 Organisation of the data 19
- 2.1 From the data warehouse to the data marts 20
  - 2.1.1 The data warehouse 20
  - 2.1.2 The data webhouse 21
  - 2.1.3 Data marts 22
- 2.2 Classification of the data 22
- 2.3 The data matrix 23
  - 2.3.1 Binarisation of the data matrix 25
- 2.4 Frequency distributions 25
  - 2.4.1 Univariate distributions 26
  - 2.4.2 Multivariate distributions 27
- 2.5 Transformation of the data 29
- 2.6 Other data structures 30
- 2.7 Further reading 31

3 Exploratory data analysis **33**
- 3.1 Univariate exploratory analysis 34
  - 3.1.1 Measures of location 35
  - 3.1.2 Measures of variability 37
  - 3.1.3 Measures of heterogeneity 37

3.1.4 Measures of concentration 39
3.1.5 Measures of asymmetry 41
3.1.6 Measures of kurtosis 43
3.2 Bivariate exploratory analysis 45
3.3 Multivariate exploratory analysis of quantitative data 49
3.4 Multivariate exploratory analysis of qualitative data 51
3.4.1 Independence and association 53
3.4.2 Distance measures 54
3.4.3 Dependency measures 56
3.4.4 Model-based measures 58
3.5 Reduction of dimensionality 61
3.5.1 Interpretation of the principal components 63
3.5.2 Application of the principal components 65
3.6 Further reading 66

4 Computational data mining 69
4.1 Measures of distance 70
4.1.1 Euclidean distance 71
4.1.2 Similarity measures 72
4.1.3 Multidimensional scaling 74
4.2 Cluster analysis 75
4.2.1 Hierarchical methods 77
4.2.2 Evaluation of hierarchical methods 81
4.2.3 Non-hierarchical methods 83
4.3 Linear regression 85
4.3.1 Bivariate linear regression 85
4.3.2 Properties of the residuals 88
4.3.3 Goodness of fit 90
4.3.4 Multiple linear regression 91
4.4 Logistic regression 96
4.4.1 Interpretation of logistic regression 97
4.4.2 Discriminant analysis 98
4.5 Tree models 100
4.5.1 Division criteria 103
4.5.2 Pruning 105
4.6 Neural networks 107
4.6.1 Architecture of a neural network 109
4.6.2 The multilayer perceptron 111
4.6.3 Kohonen networks 117
4.7 Nearest-neighbour models 119
4.8 Local models 121
4.8.1 Association rules 121
4.8.2 Retrieval by content 126
4.9 Further reading 127

5 Statistical data mining **129**
5.1 Uncertainty measures and inference 129
5.1.1 Probability 130
5.1.2 Statistical models 132
5.1.3 Statistical inference 137
5.2 Non-parametric modelling 143
5.3 The normal linear model 146
5.3.1 Main inferential results 147
5.3.2 Application 150
5.4 Generalised linear models 154
5.4.1 The exponential family 155
5.4.2 Definition of generalised linear models 157
5.4.3 The logistic regression model 163
5.4.4 Application 164
5.5 Log-linear models 167
5.5.1 Construction of a log-linear model 167
5.5.2 Interpretation of a log-linear model 169
5.5.3 Graphical log-linear models 171
5.5.4 Log-linear model comparison 174
5.5.5 Application 175
5.6 Graphical models 177
5.6.1 Symmetric graphical models 178
5.6.2 Recursive graphical models 182
5.6.3 Graphical models versus neural networks 184
5.7 Further reading 185

6 Evaluation of data mining methods **187**
6.1 Criteria based on statistical tests 188
6.1.1 Distance between statistical models 188
6.1.2 Discrepancy of a statistical model 190
6.1.3 The Kullback–Leibler discrepancy 192
6.2 Criteria based on scoring functions 193
6.3 Bayesian criteria 195
6.4 Computational criteria 197
6.5 Criteria based on loss functions 200
6.6 Further reading 204

Part II Business cases **207**

7 Market basket analysis **209**
7.1 Objectives of the analysis 209
7.2 Description of the data 210
7.3 Exploratory data analysis 212
7.4 Model building 215
7.4.1 Log-linear models 215
7.4.2 Association rules 218

7.5 Model comparison 224
7.6 Summary report 226

8 Web clickstream analysis 229
8.1 Objectives of the analysis 229
8.2 Description of the data 229
8.3 Exploratory data analysis 232
8.4 Model building 238
8.4.1 Sequence rules 238
8.4.2 Link analysis 242
8.4.3 Probabilistic expert systems 244
8.4.4 Markov chains 245
8.5 Model comparison 250
8.6 Summary report 252

9 Profiling website visitors 255
9.1 Objectives of the analysis 255
9.2 Description of the data 255
9.3 Exploratory analysis 258
9.4 Model building 258
9.4.1 Cluster analysis 258
9.4.2 Kohonen maps 262
9.5 Model comparison 264
9.6 Summary report 271

10 Customer relationship management 273
10.1 Objectives of the analysis 273
10.2 Description of the data 273
10.3 Exploratory data analysis 275
10.4 Model building 278
10.4.1 Logistic regression models 278
10.4.2 Radial basis function networks 280
10.4.3 Classification tree models 281
10.4.4 Nearest-neighbour models 285
10.5 Model comparison 286
10.6 Summary report 290

11 Credit scoring 293
11.1 Objectives of the analysis 293
11.2 Description of the data 294
11.3 Exploratory data analysis 296
11.4 Model building 299
11.4.1 Logistic regression models 299
11.4.2 Classification tree models 303
11.4.3 Multilayer perceptron models 314

11.5 Model comparison 314
11.6 Summary report 319

12 Forecasting television audience 323
12.1 Objectives of the analysis 323
12.2 Description of the data 324
12.3 Exploratory data analysis 327
12.4 Model building 337
12.5 Model comparison 347
12.6 Summary report 350

Bibliography 353
Index 357