Business Intelligence: Data Mining and Optimization for Decision Making

Carlo Vercellis

Politecnico di Milano, Italy.

WILEY

A John Wiley and Sons, Ltd., Publication

Contents

Preface **xiii**

I Components of the decision-making process 1

1 Business intelligence 3

1.1 Effective and timely decisions 3

1.2 Data, information and knowledge 6

1.3 The role of mathematical models 8

1.4 Business intelligence architectures 9

1.4.1 Cycle of a business intelligence analysis 11

1.4.2 Enabling factors in business intelligence projects 13

1.4.3 Development of a business intelligence system 14

1.5 Ethics and business intelligence 17

1.6 Notes and readings 18

2 Decision support systems **21**

2.1 Definition of system 21

2.2 Representation of the decision-maldng process 23

2.2.1 Rationality and problem solving 24

2.2.2 The decision-making process 25

2.2.3 Types of decisions 29

2.2.4 Approaches to the decision-making process 33

2.3 Evolution of information systems 35

2.4 Definition of decision support system 36

2.5 Development of a decision support system 40

2.6 Notes and readings 43

3 Data warehousing **45**

3.1 Definition of data warehouse 45

3.1.1 Data marts 49

3.1.2 Data quality 50

3.2 Data warehouse architecture . . . 51
3.2.1 ETL tools . . . 53
3.2.2 Metadata . . . 54
3.3 Cubes and multidimensional analysis . . . 55
3.3.1 Hierarchies of concepts and OLAP operations . . . 60
3.3.2 Materialization of cubes of data . . . 61
3.4 Notes and readings . . . 62

II Mathematical models and methods 63

4 Mathematical models for decision making 65
4.1 Structure of mathematical models . . . 65
4.2 Development of a model . . . 67
4.3 Classes of models . . . ., • . . . 70
4.4 Notes and readings . . . 75

5 Data mining 77
5.1 Definition of data mining . . . 77
5.1.1 Models and methods for data mining . . . 79
5.1.2 Data mining, classical statistics and OLAP . . . 80
5.1.3 Applications of data mining . . . 81
5.2 Representation of input data . . . 82
5.3 Data mining process . . . 84
5.4 Analysis methodologies . . . 90
5.5 Notes and readings . . . 94

6 Data preparation 95
6.1 Data validation . . . 95
6.1.1 Incomplete data . . . 96
6.1.2 Data affected by noise . . . 97
6.2 Data transformation . . . 99
6.2.1 Standardization . . . 99
6.2.2 Feature extraction . . . 100
6.3 Data reduction . . . 100
6.3.1 Sampling . . . 101
6.3.2 Feature selection . . . 102
6.3.3 Principal component analysis . . . 104
6.3.4 Data discretization . . . 109

7 Data exploration 113
7.1 Univariate analysis . . . 113

7.1.1 Graphical analysis of categorical attributes 114
7.1.2 Graphical analysis of numerical attributes 116
7.1.3 Measures of central tendency for numerical attributes 118
7.1.4 Measures of dispersion for numerical attributes 121
7.1.5 Measures of relative location for numerical attributes 126
7.1.6 Identification of outliers for numerical attributes 127
7.1.7 Measures of heterogeneity for categorical attributes 129
7.1.8 Analysis of the empirical density 130
7.1.9 Summary statistics 135
7.2 Bivariate analysis- 136
7.2.1 Graphical analysis 136
7.2.2 Measures of correlation for numerical attributes 142
7.2.3 Contingency tables for categorical attributes 145
7.3 Multivariate analysis ....;.> 147
7.3.1 Graphical analysis 147
7.3.2 Measures of correlation for numerical attributes 149
7.4 Notes and readings 152

Regression **153**
8.1 Structure of regression models 153
8.2 Simple linear regression 156
8.2.1 Calculating the regression line 158
8.3 Multiple linear regression 161
8.3.1 Calculating the regression coefficients 162
8.3.2 Assumptions on the residuals 163
8.3.3 Treatment of categorical predictive attributes 166
8.3.4 Ridge regression 167
8.3.5 Generalized linear regression 168
8.4 Validation of regression models 168
8.4.1 Normality and independence of the residuals 169
8.4.2 Significance of the coefficients 172
8.4.3 Analysis of variance 174
8.4.4 Coefficient of determination 175
8.4.5 Coefficient of linear correlation 176
8.4.6 Multicollinearity of the independent variables 177
8.4.7 Confidence and prediction limits 178
8.5 Selection of predictive variables 179
8.5.1 Example of development of a regression model 180
8.6 Notes and readings 185

9 Time series 187
9.1 Definition of time series . . . 187
9.1.1 Index numbers . . . 190
9.2 Evaluating time series models . . . 192
9.2.1 Distortion measures . . . 192
9.2.2 Dispersion measures . . . 193
9.2.3 Tracking signal . . . 194
9.3 Analysis of the components of time series . . . 195
9.3.1 Moving average . . . 196
9.3.2 Decomposition of a time series . . . 198
9.4 Exponential smoothing models . . . 203
9.4.1 Simple exponential smoothing . . . 203
9.4.2 Exponential smoothing with trend adjustment . . . 204
9.4.3 Exponential smoothing with trend and seasonally . . . 206
9.4.4 Simple adaptive exponential smoothing . . . 207
9.4.5 Exponential smoothing with damped trend . . . 208
9.4.6 Initial values for exponential smoothing models 209
9.4.7 Removal of trend and seasonally . . . 209
9.5 Autoregressive models . . . 210
9.5.1 Moving average models . . . 212
9.5.2 Autoregressive moving average models . . . 212
9.5.3 Autoregressive integrated moving average models . . . 212
9.5.4 Identification of autoregressive models . . . 213
9.6 Combination of predictive models . . . 216
9.7 The forecasting process . . . 217
9.7.1 Characteristics of the forecasting process . . . 217
9.7.2 Selection of a forecasting method . . . 219
9.8 Notes and readings . . . 219

10 Classification 221
10.1 Classification problems . . . 221
10.1.1 Taxonomy of classification models . . . 224
10.2 Evaluation of classification models . . . 226
10.2.1 Holdout method . . . 228
10.2.2 Repeated random sampling . . . 228
10.2.3 Cross-validation . . . 229
10.2.4 Confusion matrices . . . 230
10.2.5 ROC curve charts . . . 233
10.2.6 Cumulative gain and lift charts . . . 234
10.3 Classification trees . . . 236
10.3.1 Splitting rules . . . 240

10.3.2 Univariate splitting criteria . . . 243
10.3.3 Example of development of a classification tree . . . 246
10.3.4 Stopping criteria and pruning rules . . . 250
10.4 Bayesian methods . . . 251
10.4.1 Naive Bayesian classifiers . . . 252
10.4.2 Example of naive Bayes classifier . . . 253
10.4.3 Bayesian networks . . . 256
10.5 Logistic regression . . . 257
10.6 Neural networks . . . 259
10.6.1 The Rosenblatt perceptron . . . 259
10.6.2 Multi-level feed-forward networks . . . 260
10.7 Support vector machines . . . 262
10.7.1 Structural risk minimization . . . 262
10.7.2 Maximal margin hyperplane for linear separation . . . 266
10.7.3 Nonlinear separation . . . 270
10.8 Notes and readings . . . 275

11 Association rules 277
11.1 Motivation and structure of association rules . . . 277
11.2 Single-dimension association rules . . . 281
11.3 Apriori algorithm . . . 284
11.3.1 Generation of frequent itemsets . . . 284
11.3.2 Generation of strong rules . . . 285
11.4 General association rules . . . 288
11.5 Notes and readings . . . 290

12 Clustering 293
12.1 Clustering methods . . . 293
12.1.1 Taxonomy of clustering methods . . . 294
12.1.2 Affinity measures . . . 296
12.2 Partition methods . . . 302
12.2.1 AT-means algorithm . . . 302
12.2.2 A'-medoids algorithm . . . 305
12.3 Hierarchical methods . . . 307
12.3.1 Agglomerative hierarchical methods . . . 308
12.3.2 Divisive hierarchical methods . . . 310
12.4 Evaluation of clustering models . . . 312
12.5 Notes and readings . . . 315

III Business intelligence applications 317

13 Marketing models 319
13.1 Relational marketing . . . 320
13.1.1 Motivations and objectives . . . 320
13.1.2 An environment for relational marketing analysis . . . 327
13.1.3 Lifetime value . . . 329
13.1.4 The effect of latency in predictive models . . . 332
13.1.5 Acquisition . . . 333
13.1.6 Retention . . . 334
13.1.7 Cross-selling and up-selling . . . 335
13.1.8 Market basket analysis . . . 335
13.1.9 Web mining . . . 336
13.2 Salesforce management . . . 338
13.2.1 Decision processes in salesforce management . . . 339
13.2.2 Models for salesforce management . . . 342
13.2.3 Response functions . . . 343
13.2.4 Sales territory design . . . 346
13.2.5 Calls and product presentations planning . . . 347
13.3 Business case studies . . . 352
13.3.1 Retention in telecommunications . . . 352
13.3.2 Acquisition in the automotive industry . . . 354
13.3.3 Cross-selling in the retail industry . . . 358
13.4 Notes and readings . . . 360

14 Logistic and production models 361
14.1. Supply chain optimization . . . 362
14.2 Optimization models for logistics planning . . . 364
14.2.1 Tactical planning . . . 364
14.2.2 Extra capacity . . . 365
14.2.3 Multiple resources . . . 366
14.2.4 Backlogging . . . 366
14.2.5 Minimum lots and fixed costs . . . 369
14.2.6 Bill of materials . . . 370
14.2.7 Multiple plants . . . 371
14.3 Revenue management systems . . . 372
14.3.1 Decision processes in revenue management . . . 373
14.4 Business case studies . . . 376
14.4.1 Logistics planning in the food industry . . . 376
14.4.2 Logistics planning in the packaging industry . . . 383
14.5 Notes and readings . . . 384

15 Data envelopment analysis **385**

15.1 Efficiency measures 386

15.2 Efficient frontier 386

15.3 The CCR model 390

15.3.1 Definition of target objectives 392

15.3.2 Peer groups 393

15.4 Identification of good operating practices 394

15.4.1 Cross-efficiency analysis 394

15.4.2 Virtual inputs and virtual outputs 395

15.4.3 Weight restrictions 396

15.5 Other models 396

15.6 Notes and readings 397

Appendix A Software tools **399**

Appendix B Dataset repositories **401**

References **403**

Index **413**