Business Intelligence:
Data Mining and Optimization
for Decision Making

Carlo Vercellis

Politecnico di Milano, Italy.
Business Intelligence
Business Intelligence: Data Mining and Optimization for Decision Making

Carlo Vercellis

Politecnico di Milano, Italy.
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
CONTENTS

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

8 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

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 seasonality .. 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 seasonality .................. 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 K-means algorithm .................................. 302
    12.2.2 K-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

## 13 Marketing models

- **13.1** Relational marketing
  - 13.1.1 Motivations and objectives
  - 13.1.2 An environment for relational marketing analysis
  - 13.1.3 Lifetime value
  - 13.1.4 The effect of latency in predictive models
  - 13.1.5 Acquisition
  - 13.1.6 Retention
  - 13.1.7 Cross-selling and up-selling
  - 13.1.8 Market basket analysis
  - 13.1.9 Web mining

- **13.2** Salesforce management
  - 13.2.1 Decision processes in salesforce management
  - 13.2.2 Models for salesforce management
  - 13.2.3 Response functions
  - 13.2.4 Sales territory design
  - 13.2.5 Calls and product presentations planning

- **13.3** Business case studies
  - 13.3.1 Retention in telecommunications
  - 13.3.2 Acquisition in the automotive industry
  - 13.3.3 Cross-selling in the retail industry

- **13.4** Notes and readings

## 14 Logistic and production models

- **14.1** Supply chain optimization

- **14.2** Optimization models for logistics planning
  - 14.2.1 Tactical planning
  - 14.2.2 Extra capacity
  - 14.2.3 Multiple resources
  - 14.2.4 Backlogging
  - 14.2.5 Minimum lots and fixed costs
  - 14.2.6 Bill of materials
  - 14.2.7 Multiple plants

- **14.3** Revenue management systems
  - 14.3.1 Decision processes in revenue management

- **14.4** Business case studies
  - 14.4.1 Logistics planning in the food industry
  - 14.4.2 Logistics planning in the packaging industry

- **14.5** Notes and readings
15 Data envelopment analysis

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
Preface

Since the 1990s, the socio-economic context within which economic activities are carried out has generally been referred to as the *information and knowledge society*. The profound changes that have occurred in methods of production and in economic relations have led to a growth in the importance of the exchange of intangible goods, consisting for the most part of transfers of information. The acceleration in the pace of current transformation processes is due to two factors. The first is *globalization*, understood as the ever-increasing interdependence between the economies of the various countries, which has led to the growth of a single *global economy* characterized by a high level of integration. The second is the new *information technologies*, marked by the massive spread of the Internet and of wireless devices, which have enabled high-speed transfers of large amounts of data and the widespread use of sophisticated means of communication.

In this rapidly evolving scenario, the wealth of development opportunities is unprecedented. The easy access to information and knowledge offers several advantages to various actors in the socio-economic environment: *individuals*, who can obtain news more rapidly, access services more easily and carry out on-line commercial and banking transactions; *enterprises*, which can develop innovative products and services that can better meet the needs of the users, achieving competitive advantages from a more effective use of the knowledge gained; and, finally, the *public administration*, which can improve the services provided to citizens through the use of e-government applications, such as on-line payments of tax contributions, and e-health tools, by taking into account each patient’s medical history, thus improving the quality of healthcare services.

In this framework of radical transformation, methods of governance within complex organizations also reflect the changes occurring in the socio-economic environment, and appear increasingly more influenced by the immediate access to information for the development of effective action plans. The term *complex organizations* will be used throughout the book to collectively refer to a diversified set of entities operating in the socio-economic context, including enterprises, government agencies, banking and financial institutions, and non-profit organizations.