Statistical Data Analytics

Foundations for Data Mining, Informatics, and Knowledge Discovery

Walter W. Piegorsch

University of Arizona, USA

WILEY

Contents

Preface xiii

Part I Background: Introductory Statistical Analytics 1

1 Data analytics and data mining 3
- 1.1 Knowledge discovery: finding structure in data 3
- 1.2 Data quality versus data quantity 5
- 1.3 Statistical modeling versus statistical description 7

2 Basic probability and statistical distributions 10
- 2.1 Concepts in probability 10
  - 2.1.1 Probability rules 11
  - 2.1.2 Random variables and probability functions 12
  - 2.1.3 Means, variances, and expected values 17
  - 2.1.4 Median, quartiles, and quantiles 18
  - 2.1.5 Bivariate expected values, covariance, and correlation 20
- 2.2 Multiple random variables* 21
- 2.3 Univariate families of distributions 23
  - 2.3.1 Binomial distribution 23
  - 2.3.2 Poisson distribution 26
  - 2.3.3 Geometric distribution 27
  - 2.3.4 Negative binomial distribution 27
  - 2.3.5 Discrete uniform distribution 28
  - 2.3.6 Continuous uniform distribution 29
  - 2.3.7 Exponential distribution 29
  - 2.3.8 Gamma and chi-square distributions 30
  - 2.3.9 Normal (Gaussian) distribution 32
  - 2.3.10 Distributions derived from normal 37
  - 2.3.11 The exponential family 41

3 Data manipulation 49
3.1 Random sampling 49
3.2 Data types 51
3.3 Data summarization 52
3.3.1 Means, medians, and central tendency 52
3.3.2 Summarizing variation 56
3.3.3 Summarizing (bivariate) correlation 59
3.4 Data diagnostics and data transformation 60
3.4.1 Outlier analysis 60
3.4.2 Entropy* 62
3.4.3 Data transformation 64
3.5 Simple smoothing techniques 65
3.5.1 Binning 66
3.5.2 Moving averages* 67
3.5.3 Exponential smoothing* 69

4 Data visualization and statistical graphics 76
4.1 Univariate visualization 77
4.1.1 Strip charts and dot plots 77
4.1.2 Boxplots 79
4.1.3 Stem-and-leaf plots 81
4.1.4 Histograms and density estimators 83
4.1.5 Quantile plots 87
4.2 Bivariate and multivariate visualization 89
4.2.1 Pie charts and bar charts 90
4.2.2 Multiple boxplots and QQ plots 95
4.2.3 Scatterplots and bubble plots 98
4.2.4 Heatmaps 102
4.2.5 Time series plots* 105

5 Statistical inference 115
5.1 Parameters and likelihood 115
5.2 Point estimation 117
5.2.1 Bias 118
5.2.2 The method of moments 118
5.2.3 Least squares/weighted least squares 119
5.2.4 Maximum likelihood* 120
5.3 Interval estimation 123
5.3.1 Confidence intervals 123
5.3.2 Single-sample intervals for normal (Gaussian) parameters 124
5.3.3 Two-sample intervals for normal (Gaussian) parameters 128
5.3.4 Wald intervals and likelihood intervals* 131
5.3.5 Delta method intervals* 135
5.3.6 Bootstrap intervals* 137
5.4 Testing hypotheses 138
5.4.1 Single-sample tests for normal (Gaussian) parameters 140
5.4.2 Two-sample tests for normal (Gaussian) parameters 142

5.4.3 Walds tests, likelihood ratio tests, and 'exact' tests* 145
5.5 Multiple inferences* 148
5.5.1 Bonferroni multiplicity adjustment 149
5.5.2 False discovery rate 151

Part II Statistical Learning and Data Analytics 161

6 Techniques for supervised learning: simple linear regression 163
6.1 What is "supervised learning?" 163
6.2 Simple linear regression 164
6.2.1 The simple linear model 164
6.2.2 Multiple inferences and simultaneous confidence bands 171
6.3 Regression diagnostics 175
6.4 Weighted least squares (WLS) regression 184
6.5 Correlation analysis 187
6.5.1 The correlation coefficient 187
6.5.2 Rank correlation 190

7 Techniques for supervised learning: multiple linear regression 198
7.1 Multiple linear regression 198
7.1.1 Matrix formulation 199
7.1.2 Weighted least squares for the MLR model 200
7.1.3 Inferences under the MLR model 201
7.1.4 Multicollinearity 208
7.2 Polynomial regression 210
7.3 Feature selection 211
7.3.1 R_p^2 plots 212
7.3.2 Information criteria: AIC and BIC 215
7.3.3 Automated variable selection 216
7.4 Alternative regression methods* 223
7.4.1 Loess 224
7.4.2 Regularization: ridge regression 230
7.4.3 Regularization and variable selection: the Lasso 238
7.5 Qualitative predictors: ANOVA models 242

8 Supervised learning: generalized linear models 258
8.1 Extending the linear regression model 258
8.1.1 Nonnormal data and the exponential family 258
8.1.2 Link functions 259
8.2 Technical details for GLiMs* 259
8.2.1 Estimation 260
8.2.2 The deviance function 261
8.2.3 Residuals 262
8.2.4 Inference and model assessment 264
8.3 Selected forms of GLiMs 265
8.3.1 Logistic regression and binary-data GLiMs 265

8.3.2 Trend testing with proportion data 271
8.3.3 Contingency tables and log-linear models 273
8.3.4 Gamma regression models 281

9 Supervised learning: classification 291
9.1 Binary classification via logistic regression 292
9.1.1 Logistic discriminants 292
9.1.2 Discriminant rule accuracy 296
9.1.3 ROC curves 297
9.2 Linear discriminant analysis (LDA) 297
9.2.1 Linear discriminant functions 297
9.2.2 Bayes discriminant/classification rules 302
9.2.3 Bayesian classification with normal data 303
9.2.4 Naïve Bayes classifiers 308
9.3 *k*-Nearest neighbor classifiers 308
9.4 Tree-based methods 312
9.4.1 Classification trees 312
9.4.2 Pruning 314
9.4.3 Boosting 321
9.4.4 Regression trees 321
9.5 Support vector machines* 322
9.5.1 Separable data 322
9.5.2 Nonseparable data 325
9.5.3 Kernel transformations 326

10 Techniques for unsupervised learning: dimension reduction 341
10.1 Unsupervised versus supervised learning 341
10.2 Principal component analysis 342
10.2.1 Principal components 342
10.2.2 Implementing a PCA 344
10.3 Exploratory factor analysis 351
10.3.1 The factor analytic model 351
10.3.2 Principal factor estimation 353
10.3.3 Maximum likelihood estimation 354
10.3.4 Selecting the number of factors 355
10.3.5 Factor rotation 356
10.3.6 Implementing an EFA 357
10.4 Canonical correlation analysis* 361

11 Techniques for unsupervised learning: clustering and association 373
11.1 Cluster analysis 373
11.1.1 Hierarchical clustering 376
11.1.2 Partitioned clustering 384
11.2 Association rules/market basket analysis 395
11.2.1 Association rules for binary observations 396
11.2.2 Measures of rule quality 397

11.2.3 The Apriori algorithm 398
11.2.4 Statistical measures of association quality 402

A Matrix manipulation 411
A.1 Vectors and matrices 411
A.2 Matrix algebra 412
A.3 Matrix inversion 414
A.4 Quadratic forms 415
A.5 Eigenvalues and eigenvectors 415
A.6 Matrix factorizations 416
A.6.1 QR decomposition 417
A.6.2 Spectral decomposition 417
A.6.3 Matrix square root 417
A.6.4 Singular value decomposition 418
A.7 Statistics via matrix operations 419

B Brief introduction to **R** 421
B.1 Data entry and manipulation 422
B.2 A turbo-charged calculator 426
B.3 **R** functions 427
B.3.1 Inbuilt **R** functions 427
B.3.2 Flow control 429
B.3.3 User-defined functions 429
B.4 **R** packages 430

References **432**

Index **453**