Shuai Huang & Houtao Deng

Data Analytics

A Small Data Approach

Contents

Preface xi

Acknowledgments xiii

Chapter 1: Introduction 1

Who will benefit from this book? 1

Overview of a data analytics pipeline 2

Topics in a nutshell 3

Chapter 2: Abstraction
Regression & Tree Models 5

Overview 5

Regression models 8

Tree models 22

Remarks 29

Exercises 34

Chapter 3: Recognition
Logistic Regression & Ranking 37

Overview 37

Logistic regression model 38

Ranking problem by pairwise comparison 53

Statistical process control using decision tree 55

Remarks 63

Exercises 67

Chapter 4: Resonance
Bootstrap & Random Forests 69

Overview 69

How bootstrap works 70

Random forests 81

Remarks 92

Exercises 95

Chapter 5: Learning (I)
Cross-validation & OOB 97

Overview 97

Cross-validation 98

Out-of-bag error in random forests 110

Remarks 114

Exercises 121

Chapter 6: Diagnosis
Residuals & Heterogeneity 123

Overview 123

Diagnosis in regression 124

Diagnosis in random forests 130

Clustering 131

Remarks 137

Exercises 143

Chapter 7: Learning (II)
SVM & Ensemble Learning 147

Overview 147
Support vector machine 147
Ensemble learning 161
Remarks 170
Exercises 173

Chapter 8: Scalability
LASSO & PCA 175

Overview 175
LASSO 175
Principal component analysis 183
Remarks 191
Exercises 198

Chapter 9: Pragmatism
Experience & Experimental 201

Overview 201
Kernel regression model 201
Conditional variance regression model 210
Remarks 216
Exercises 217

Chapter 10: Synthesis
Architecture & Pipeline 219

Overview 219
Deep learning 219
inTrees 235
Remarks 243
Exercises 245

Conclusion 247

Appendix: A Brief Review of Background Knowledge 249

The Normal Distribution 249

Matrix Operations 251

Optimization 253

Index 255