Practical Data Science with R

NINA ZUMEL
JOHN MOUNT

MANNING
SHELTER ISLAND

To our parents
Olive and Paul Zumel
Peggy and David Mount

brief contents

PART 1 INTRODUCTION TO DATA SCIENCE 1

1 ■ The data science process 3
2 ■ Loading data into R 18
3 ■ Exploring data 35
4 ■ Managing data 64

PART 2 MODELING METHODS 81

5 ■ Choosing and evaluating models 83
6 ■ Memorization methods 115
7 ■ Linear and logistic regression 140
8 ■ Unsupervised methods 175
9 ■ Exploring advanced methods 211

PART 3 DELIVERING RESULTS 253

10 ■ Documentation and deployment 255
11 ■ Producing effective presentations 287

contents

foreword *xv*
preface *xvii*
acknowledgments *xviii*
about this book *xix*
about the cover illustration *xxv*

PART 1 INTRODUCTION TO DATA SCIENCE....................1

1 ***The data science process* 3**

1.1 The roles in a data science project 3

Project roles 4

1.2 Stages of a data science project 6

Defining the goal 7 ▪ Data collection and management 8 Modeling 10 ▪ Model evaluation and critique 11 Presentation and documentation 13 ▪ Model deployment and maintenance 14

1.3 Setting expectations 14

Determining lower and upper bounds on model performance 15

1.4 Summary 17

2 *Loading data into R 18*

2.1 Working with data from files 19

Working with well-structured data from files or URLs 19
Using R on less-structured data 22

2.2 Working with relational databases 24

A production-size example 25 ▪ Loading data from a database into R 30 ▪ Working with the PUMS data 31

2.3 Summary 34

3 *Exploring data 35*

3.1 Using summary statistics to spot problems 36

Typical problems revealed by data summaries 38

3.2 Spotting problems using graphics and visualization 41

Visually checking distributions for a single variable 43
Visually checking relationships between two variables 51

3.3 Summary 62

4 *Managing data 64*

4.1 Cleaning data 64

Treating missing values (NAs) 65 ▪ Data transformations 69

4.2 Sampling for modeling and validation 76

Test and training splits 76 ▪ Creating a sample group column 77 ▪ Record grouping 78 ▪ Data provenance 78

4.3 Summary 79

PART 2 MODELING METHODS 81

5 *Choosing and evaluating models 83*

5.1 Mapping problems to machine learning tasks 84

Solving classification problems 85 ▪ Solving scoring problems 87 ▪ Working without known targets 88
Problem-to-method mapping 90

5.2 Evaluating models 92

Evaluating classification models 93 ▪ Evaluating scoring models 98 ▪ Evaluating probability models 101 ▪ Evaluating ranking models 105 ▪ Evaluating clustering models 105

5.3 Validating models 108

Identifying common model problems 108 ▪ Quantifying model soundness 110 ▪ Ensuring model quality 111

5.4 Summary 113

6 *Memorization methods 115*

6.1 KDD and KDD Cup 2009 116

Getting started with KDD Cup 2009 data 117

6.2 Building single-variable models 118

Using categorical features 119 ▪ Using numeric features 121 Using cross-validation to estimate effects of overfitting 123

6.3 Building models using many variables 125

Variable selection 125 ▪ Using decision trees 127 ▪ Using nearest neighbor methods 130 ▪ Using Naive Bayes 134

6.4 Summary 138

7 *Linear and logistic regression 140*

7.1 Using linear regression 141

Understanding linear regression 141 ▪ Building a linear regression model 144 ▪ Making predictions 145 ▪ Finding relations and extracting advice 149 ▪ Reading the model summary and characterizing coefficient quality 151 ▪ Linear regression takeaways 156

7.2 Using logistic regression 157

Understanding logistic regression 157 ▪ Building a logistic regression model 159 ▪ Making predictions 160 ▪ Finding relations and extracting advice from logistic models 164 Reading the model summary and characterizing coefficients 166 Logistic regression takeaways 173

7.3 Summary 174

8 *Unsupervised methods 175*

8.1 Cluster analysis 176

Distances 176 ▪ Preparing the data 178 ▪ Hierarchical clustering with hclust() 180 ▪ The k-means algorithm 190 Assigning new points to clusters 195 ▪ Clustering takeaways 198

8.2 Association rules 198

Overview of association rules 199 ▪ The example problem 200 Mining association rules with the arules package 201 Association rule takeaways 209

8.3 Summary 209

9 *Exploring advanced methods 211*

9.1 Using bagging and random forests to reduce training variance 212

Using bagging to improve prediction 213 ▪ Using random forests to further improve prediction 216 ▪ Bagging and random forest takeaways 220

9.2 Using generalized additive models (GAMs) to learn non-monotone relationships 221

Understanding GAMs 221 ▪ A one-dimensional regression example 222 ▪ Extracting the nonlinear relationships 226 Using GAM on actual data 228 ▪ Using GAM for logistic regression 231 ▪ GAM takeaways 233

9.3 Using kernel methods to increase data separation 233

Understanding kernel functions 234 ▪ Using an explicit kernel on a problem 238 ▪ Kernel takeaways 241

9.4 Using SVMs to model complicated decision boundaries 242

Understanding support vector machines 242 ▪ Trying an SVM on artificial example data 245 ▪ Using SVMs on real data 248 Support vector machine takeaways 251

9.5 Summary 251

PART 3 DELIVERING RESULTS 253

10 *Documentation and deployment 255*

10.1 The buzz dataset 256

10.2 Using knitr to produce milestone documentation 258

What is knitr? 258 ▪ knitr technical details 261 ▪ Using knitr to document the buzz data 262

10.3 Using comments and version control for running documentation 266

Writing effective comments 266 ▪ Using version control to record history 267 ▪ Using version control to explore your project 272 Using version control to share work 276

10.4 Deploying models 280

Deploying models as R HTTP services 280 ▪ Deploying models by export 283 ▪ What to take away 284

10.5 Summary 286

11 *Producing effective presentations 287*

11.1 Presenting your results to the project sponsor 288

Summarizing the project's goals 289 ▪ Stating the project's results 290 ▪ Filling in the details 292 ▪ Making recommendations and discussing future work 294 Project sponsor presentation takeaways 295

11.2 Presenting your model to end users 295

Summarizing the project's goals 296 ▪ Showing how the model fits the users' workflow 296 ▪ Showing how to use the model 299 End user presentation takeaways 300

11.3 Presenting your work to other data scientists 301

Introducing the problem 301 ▪ Discussing related work 302 Discussing your approach 302 ▪ Discussing results and future work 303 ▪ Peer presentation takeaways 304

11.4 Summary 304

appendix A Working with R and other tools 307

appendix B Important statistical concepts 333

appendix C More tools and ideas worth exploring 369

bibliography 375

index 377