

Contents

1	Introduction	1
1.1	What Is Data Science?	1
1.2	Why R?	2
1.3	Goal of This Book	3
1.4	Book Overview	3
	References	4
2	Overview of the R Programming Language	5
2.1	Installing R	5
2.1.1	Development Tools	6
2.2	R Programming Language	7
2.2.1	Operators	7
2.2.2	Printing Values	8
2.2.3	Basic Data Types	9
2.2.4	Control Structures	10
2.2.5	Functions	12
2.3	Packages	13
2.3.1	R Help System	15
2.4	Running R Code	16
	Reference	17
3	Getting Data into R	19
3.1	Reading Data	20
3.1.1	Text Files	20
3.2	Cleaning Up Data	25
3.2.1	Identifying Data Types	26
3.2.2	Data Entry Errors	27
3.2.3	Missing Values	28
3.3	Chapter Summary	30
	References	30

4	Data Visualization	31
4.1	Introduction	31
4.2	Basic Visualizations	32
4.2.1	Scatterplots	33
4.2.2	Visualizing Aggregate Values with Bar plots and Pie charts	39
4.2.3	Common Plotting Tasks	45
4.3	Layered Visualizations Using ggplot2	48
4.3.1	Creating Plots Using qplot()	48
4.3.2	ggplot(): Specifying the Grammar of the Visualization	53
4.3.3	Themes	55
4.4	Interactive Visualizations Using Shiny	55
4.5	Chapter Summary and Further Reading	59
	References	60
5	Exploratory Data Analysis	61
5.1	Summary Statistics	62
5.1.1	Dataset Size	62
5.1.2	Summarizing the Data	63
5.1.3	Ordering Data by a Variable	65
5.1.4	Group and Split Data by a Variable	66
5.1.5	Variable Correlation	68
5.2	Getting a Sense of Data Distribution	71
5.2.1	Box Plots	71
5.2.2	Histograms	75
5.2.3	Measuring Data Symmetry Using Skewness and Kurtosis	80
5.3	Putting It All Together: Outlier Detection	82
5.4	Chapter Summary	84
	References	85
6	Regression	87
6.1	Introduction	87
6.1.1	Regression Models	88
6.2	Parametric Regression Models	89
6.2.1	Simple Linear Regression	90
6.2.2	Multivariate Linear Regression	99
6.2.3	Log-Linear Regression Models	101
6.3	Nonparametric Regression Models	103
6.3.1	Locally Weighted Regression	104
6.3.2	Kernel Regression	107
6.3.3	Regression Trees	109
6.4	Chapter Summary	114
	References	114

7	Classification	115
7.1	Introduction	115
7.1.1	Training and Test Datasets	116
7.2	Parametric Classification Models	117
7.2.1	Naive Bayes	117
7.2.2	Logistic Regression	122
7.2.3	Support Vector Machines	126
7.3	Nonparametric Classification Models	130
7.3.1	Nearest Neighbors	131
7.3.2	Decision Trees	133
7.4	Chapter Summary	135
	References	136
8	Text Mining	137
8.1	Introduction	137
8.2	Dataset	138
8.3	Reading Text Input Data	138
8.4	Common Text Preprocessing Tasks	141
8.4.1	Stop Word Removal	142
8.4.2	Stemming	143
8.5	Term Document Matrix	144
8.5.1	TF-IDF Weighting Function	147
8.6	Text Mining Applications	149
8.6.1	Frequency Analysis	149
8.6.2	Text Classification	151
8.7	Chapter Summary	157

Beginning Data Science with R

Pathak, M.A.

2014, XI, 157 p. 155 illus., 26 illus. in color., Hardcover

ISBN: 978-3-319-12065-2