
Contents

CHAPTER 1 ■ Introduction	1
1.1 DATA SCIENCE	1
1.2 BIG DATA	4
1.3 JULIA	5
1.4 JULIA AND R PACKAGES	6
1.5 DATASETS	6
1.5.1 Overview	6
1.5.2 Beer Data	6
1.5.3 Coffee Data	7
1.5.4 Leptograpsus Crabs Data	8
1.5.5 Food Preferences Data	9
1.5.6 x2 Data	9
1.5.7 Iris Data	11
1.6 OUTLINE OF THE CONTENTS OF THIS MONOGRAPH	11
CHAPTER 2 ■ Core Julia	13
2.1 VARIABLE NAMES	13
2.2 OPERATORS	14
2.3 TYPES	15
2.3.1 Numeric	15
2.3.2 Floats	17
2.3.3 Strings	19
2.3.4 Tuples	22
2.4 DATA STRUCTURES	23
2.4.1 Arrays	23

viii ■ Contents

2.4.2	Dictionaries	26
2.5	CONTROL FLOW	28
2.5.1	Compound Expressions	28
2.5.2	Conditional Evaluation	29
2.5.3	Loops	30
2.5.3.1	Basics	30
2.5.3.2	Loop termination	32
2.5.3.3	Exception handling	33
2.6	FUNCTIONS	36
CHAPTER 3	Working with Data	43
3.1	DATAFRAMES	43
3.2	CATEGORICAL DATA	47
3.3	INPUT/OUTPUT	48
3.4	USEFUL DATAFRAME FUNCTIONS	54
3.5	SPLIT-APPLY-COMBINE STRATEGY	56
3.6	QUERY.JL	59
CHAPTER 4	Visualizing Data	67
4.1	GADFLY.JL	67
4.2	VISUALIZING UNIVARIATE DATA	69
4.3	DISTRIBUTIONS	72
4.4	VISUALIZING BIVARIATE DATA	83
4.5	ERROR BARS	90
4.6	FACETS	91
4.7	SAVING PLOTS	91
CHAPTER 5	Supervised Learning	93
5.1	INTRODUCTION	93
5.2	CROSS-VALIDATION	96
5.2.1	Overview	96
5.2.2	<i>K</i> -Fold Cross-Validation	97
5.3	<i>K</i> -NEAREST NEIGHBOURS CLASSIFICATION	99
5.4	CLASSIFICATION AND REGRESSION TREES	102

5.4.1	Overview	102
5.4.2	Classification Trees	103
5.4.3	Regression Trees	106
5.4.4	Comments	108
5.5	BOOTSTRAP	108
5.6	RANDOM FORESTS	111
5.7	GRADIENT BOOSTING	113
5.7.1	Overview	113
5.7.2	Beer Data	116
5.7.3	Food Data	121
5.8	COMMENTS	126
CHAPTER 6	■ Unsupervised Learning	129
6.1	INTRODUCTION	129
6.2	PRINCIPAL COMPONENTS ANALYSIS	132
6.3	PROBABILISTIC PRINCIPAL COMPONENTS ANALYSIS	135
6.4	EM ALGORITHM FOR PPCA	137
6.4.1	Background: EM Algorithm	137
6.4.2	E-step	138
6.4.3	M-step	139
6.4.4	Woodbury Identity	140
6.4.5	Initialization	141
6.4.6	Stopping Rule	141
6.4.7	Implementing the EM Algorithm for PPCA	142
6.4.8	Comments	146
6.5	K-MEANS CLUSTERING	148
6.6	MIXTURE OF PROBABILISTIC PRINCIPAL COMPONENTS ANALYZERS	151
6.6.1	Model	151
6.6.2	Parameter Estimation	152
6.6.3	Illustrative Example: Coffee Data	161
6.7	COMMENTS	162

CHAPTER 7 ■ R Interoperability	165
7.1 ACCESSING R DATASETS	165
7.2 INTERACTING WITH R	166
7.3 EXAMPLE: CLUSTERING AND DATA REDUC- TION FOR THE COFFEE DATA	171
7.3.1 Coffee Data	171
7.3.2 PGMM Analysis	172
7.3.3 VSCC Analysis	175
7.4 EXAMPLE: FOOD DATA	176
7.4.1 Overview	176
7.4.2 Random Forests	176
APPENDIX A ■ Julia and R Packages Used Herein	185
APPENDIX B ■ Variables for Food Data	187
APPENDIX C ■ Useful Mathematical Results	193
C.1 BRIEF OVERVIEW OF EIGENVALUES	193
C.2 SELECTED LINEAR ALGEBRA RESULTS	193
C.3 MATRIX CALCULUS RESULTS	194
APPENDIX D ■ Performance Tips	197
D.1 FLOATING POINT NUMBERS	197
D.1.1 Do Not Test for Equality	197
D.1.2 Use Logarithms for Division	198
D.1.3 Subtracting Two Nearly Equal Numbers	198
D.2 JULIA PERFORMANCE	199
D.2.1 General Tips	199
D.2.2 Array Processing	199
D.2.3 Separate Core Computations	201
APPENDIX E ■ Linear Algebra Functions	203
E.1 VECTOR OPERATIONS	203
E.2 MATRIX OPERATIONS	204

E.3 MATRIX DECOMPOSITIONS	205
References	208
<hr/>	
Index	217
<hr/>	