

# Contents

<b>1</b>	<b>Introduction . . . . .</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.1.1	Data and Knowledge . . . . .	2
1.1.2	Tycho Brahe and Johannes Kepler . . . . .	4
1.1.3	Intelligent Data Analysis . . . . .	6
1.2	The Data Analysis Process . . . . .	7
1.3	Methods, Tasks, and Tools . . . . .	11
1.4	How to Read This Book . . . . .	13
	References . . . . .	14
<b>2</b>	<b>Practical Data Analysis: An Example . . . . .</b>	<b>15</b>
2.1	The Setup . . . . .	15
2.2	Data Understanding and Pattern Finding . . . . .	16
2.3	Explanation Finding . . . . .	20
2.4	Predicting the Future . . . . .	21
2.5	Concluding Remarks . . . . .	23
<b>3</b>	<b>Project Understanding . . . . .</b>	<b>25</b>
3.1	Determine the Project Objective . . . . .	26
3.2	Assess the Situation . . . . .	28
3.3	Determine Analysis Goals . . . . .	30
3.4	Further Reading . . . . .	31
	References . . . . .	32
<b>4</b>	<b>Data Understanding . . . . .</b>	<b>33</b>
4.1	Attribute Understanding . . . . .	34
4.2	Data Quality . . . . .	37
4.3	Data Visualization . . . . .	40
4.3.1	Methods for One and Two Attributes . . . . .	40
4.3.2	Methods for Higher-Dimensional Data . . . . .	48
4.4	Correlation Analysis . . . . .	59

4.5	Outlier Detection . . . . .	62
4.5.1	Outlier Detection for Single Attributes . . . . .	63
4.5.2	Outlier Detection for Multidimensional Data . . . . .	64
4.6	Missing Values . . . . .	65
4.7	A Checklist for Data Understanding . . . . .	68
4.8	Data Understanding in Practice . . . . .	69
4.8.1	Data Understanding in KNIME . . . . .	70
4.8.2	Data Understanding in R . . . . .	73
	References . . . . .	78
<b>5</b>	<b>Principles of Modeling . . . . .</b>	<b>81</b>
5.1	Model Classes . . . . .	82
5.2	Fitting Criteria and Score Functions . . . . .	85
5.2.1	Error Functions for Classification Problems . . . . .	87
5.2.2	Measures of Interestingness . . . . .	89
5.3	Algorithms for Model Fitting . . . . .	89
5.3.1	Closed Form Solutions . . . . .	89
5.3.2	Gradient Method . . . . .	90
5.3.3	Combinatorial Optimization . . . . .	92
5.3.4	Random Search, Greedy Strategies, and Other Heuristics . . . . .	92
5.4	Types of Errors . . . . .	93
5.4.1	Experimental Error . . . . .	94
5.4.2	Sample Error . . . . .	99
5.4.3	Model Error . . . . .	100
5.4.4	Algorithmic Error . . . . .	101
5.4.5	Machine Learning Bias and Variance . . . . .	101
5.4.6	Learning Without Bias? . . . . .	102
5.5	Model Validation . . . . .	102
5.5.1	Training and Test Data . . . . .	102
5.5.2	Cross-Validation . . . . .	103
5.5.3	Bootstrapping . . . . .	104
5.5.4	Measures for Model Complexity . . . . .	105
5.6	Model Errors and Validation in Practice . . . . .	111
5.6.1	Errors and Validation in KNIME . . . . .	111
5.6.2	Validation in R . . . . .	111
5.7	Further Reading . . . . .	113
	References . . . . .	113
<b>6</b>	<b>Data Preparation . . . . .</b>	<b>115</b>
6.1	Select Data . . . . .	115
6.1.1	Feature Selection . . . . .	116
6.1.2	Dimensionality Reduction . . . . .	121
6.1.3	Record Selection . . . . .	121
6.2	Clean Data . . . . .	123
6.2.1	Improve Data Quality . . . . .	123

6.2.2	Missing Values . . . . .	124
6.3	Construct Data . . . . .	127
6.3.1	Provide Operability . . . . .	127
6.3.2	Assure Impartiality . . . . .	129
6.3.3	Maximize Efficiency . . . . .	131
6.4	Complex Data Types . . . . .	134
6.5	Data Integration . . . . .	135
6.5.1	Vertical Data Integration . . . . .	136
6.5.2	Horizontal Data Integration . . . . .	136
6.6	Data Preparation in Practice . . . . .	138
6.6.1	Data Preparation in KNIME . . . . .	139
6.6.2	Data Preparation in R . . . . .	141
	References . . . . .	142
<b>7</b>	<b>Finding Patterns . . . . .</b>	<b>145</b>
7.1	Hierarchical Clustering . . . . .	147
7.1.1	Overview . . . . .	148
7.1.2	Construction . . . . .	150
7.1.3	Variations and Issues . . . . .	152
7.2	Notion of (Dis-)Similarity . . . . .	155
7.3	Prototype- and Model-Based Clustering . . . . .	162
7.3.1	Overview . . . . .	162
7.3.2	Construction . . . . .	164
7.3.3	Variations and Issues . . . . .	167
7.4	Density-Based Clustering . . . . .	169
7.4.1	Overview . . . . .	170
7.4.2	Construction . . . . .	171
7.4.3	Variations and Issues . . . . .	173
7.5	Self-organizing Maps . . . . .	175
7.5.1	Overview . . . . .	175
7.5.2	Construction . . . . .	176
7.6	Frequent Pattern Mining and Association Rules . . . . .	179
7.6.1	Overview . . . . .	179
7.6.2	Construction . . . . .	181
7.6.3	Variations and Issues . . . . .	187
7.7	Deviation Analysis . . . . .	194
7.7.1	Overview . . . . .	194
7.7.2	Construction . . . . .	195
7.7.3	Variations and Issues . . . . .	197
7.8	Finding Patterns in Practice . . . . .	198
7.8.1	Finding Patterns with KNIME . . . . .	199
7.8.2	Finding Patterns in R . . . . .	201
7.9	Further Reading . . . . .	203
	References . . . . .	204

<b>8</b>	<b>Finding Explanations</b>	207
8.1	Decision Trees	208
8.1.1	Overview	209
8.1.2	Construction	210
8.1.3	Variations and Issues	213
8.2	Bayes Classifiers	218
8.2.1	Overview	218
8.2.2	Construction	220
8.2.3	Variations and Issues	224
8.3	Regression	229
8.3.1	Overview	230
8.3.2	Construction	231
8.3.3	Variations and Issues	234
8.3.4	Two Class Problems	242
8.4	Rule learning	244
8.4.1	Propositional Rules	245
8.4.2	Inductive Logic Programming or First-Order Rules	251
8.5	Finding Explanations in Practice	253
8.5.1	Finding Explanations with KNIME	253
8.5.2	Using Explanations with R	255
8.6	Further Reading	257
	References	258
<b>9</b>	<b>Finding Predictors</b>	259
9.1	Nearest-Neighbor Predictors	261
9.1.1	Overview	261
9.1.2	Construction	263
9.1.3	Variations and Issues	265
9.2	Artificial Neural Networks	269
9.2.1	Overview	269
9.2.2	Construction	272
9.2.3	Variations and Issues	276
9.3	Support Vector Machines	277
9.3.1	Overview	278
9.3.2	Construction	282
9.3.3	Variations and Issues	283
9.4	Ensemble Methods	284
9.4.1	Overview	284
9.4.2	Construction	286
9.4.3	Further Reading	289
9.5	Finding Predictors in Practice	290
9.5.1	Finding Predictors with KNIME	290
9.5.2	Using Predictors in R	292
	References	294

<b>10</b>	<b>Evaluation and Deployment</b>	297
10.1	Evaluation	297
10.2	Deployment and Monitoring	299
	References	301
<b>A</b>	<b>Statistics</b>	303
A.1	Terms and Notation	304
A.2	Descriptive Statistics	305
A.2.1	Tabular Representations	305
A.2.2	Graphical Representations	306
A.2.3	Characteristic Measures for One-Dimensional Data	309
A.2.4	Characteristic Measures for Multidimensional Data	316
A.2.5	Principal Component Analysis	318
A.3	Probability Theory	323
A.3.1	Probability	323
A.3.2	Basic Methods and Theorems	327
A.3.3	Random Variables	333
A.3.4	Characteristic Measures of Random Variables	339
A.3.5	Some Special Distributions	343
A.4	Inferential Statistics	349
A.4.1	Random Samples	350
A.4.2	Parameter Estimation	351
A.4.3	Hypothesis Testing	361
<b>B</b>	<b>The R Project</b>	369
B.1	Installation and Overview	369
B.2	Reading Files and R Objects	370
B.3	R Functions and Commands	372
B.4	Libraries/Packages	373
B.5	R Workspace	373
B.6	Finding Help	374
B.7	Further Reading	374
<b>C</b>	<b>KNIME</b>	375
C.1	Installation and Overview	375
C.2	Building Workflows	377
C.3	Example Flow	378
C.4	R Integration	380
	<b>References</b>	383
	Appendix A	383
	Appendix B	383
	<b>Index</b>	385

Guide to Intelligent Data Analysis

How to Intelligently Make Sense of Real Data

Berthold, M.R.; Borgelt, C.; Höppner, F.; Klawonn, F.

2010, XIII, 394 p. 141 illus., 78 illus. in color., Hardcover

ISBN: 978-1-84882-259-7