

Contents

Foreword	i
Preface	v
Why Data Science?	vii
1. Basics in a Nutshell	1
1.1 Data Preparation	1
1.2 Algorithm Selection	5
1.3 Parameter Tuning	9
1.4 Evaluating Results	11
1.5 Summary	14
2. <i>k</i>-Means Clustering	15
2.1 Finding Customer Clusters	15
2.2 Example: Personality Profiles of Movie Fans	15
2.3 Defining Clusters	18
2.4 Limitations	21
2.5 Summary	22
3. Principal Component Analysis	23
3.1 Exploring Nutritional Content of Food	23
3.2 Principal Components	24
3.3 Example: Analyzing Food Groups	26
3.4 Limitations	31
3.5 Summary	33

CONTENTS

4. Association Rules	35
4.1 Discovering Purchasing Patterns	35
4.2 Support, Confidence and Lift	36
4.3 Example: Transacting Grocery Sales	38
4.4 Apriori Principle	40
4.5 Limitations	42
4.6 Summary	42
5. Social Network Analysis	43
5.1 Mapping out Relationships	43
5.2 Example: Geopolitics in Weapons Trade	44
5.3 Louvain Method	48
5.4 PageRank Algorithm	49
5.5 Limitations	52
5.6 Summary	53
6. Regression Analysis	55
6.1 Deriving a Trend Line	55
6.2 Example: Predicting House Prices	55
6.3 Gradient Descent	58
6.4 Regression Coefficients	60
6.5 Correlation Coefficients	61
6.6 Limitations	63
6.7 Summary	64
7. k-Nearest Neighbors and Anomaly Detection	65
7.1 Food Forensics	65
7.2 Birds of a Feather Flock Together	66
7.3 Example: Distilling Differences in Wine	67
7.4 Anomaly Detection	68
7.5 Limitations	70
7.6 Summary	70
8. Support Vector Machine	71
8.1 “No” or “Oh No”?	71

8.2	Example: Predicting Heart Disease	71
8.3	Delineating an Optimal Boundary	73
8.4	Limitations	76
8.5	Summary	76
9.	Decision Tree	77
9.1	Predicting Survival in a Disaster	77
9.2	Example: Escaping from the Titanic	78
9.3	Generating a Decision Tree	80
9.4	Limitations	81
9.5	Summary	82
10.	Random Forests	83
10.1	Wisdom of the Crowd	83
10.2	Example: Forecasting Crime	83
10.3	Ensembles	87
10.4	Bootstrap Aggregating (Bagging)	88
10.5	Limitations	89
10.6	Summary	90
11.	Neural Networks	91
11.1	Building a Brain	91
11.2	Example: Recognizing Handwritten Digits	92
11.3	Components of a Neural Network	96
11.4	Activation Rules	98
11.5	Limitations	99
11.6	Summary	102
12.	A/B Testing and Multi-Armed Bandits	103
12.1	Basics of A/B testing	103
12.2	Limitations of A/B testing	104
12.3	Epsilon-Decreasing Strategy	104
12.4	Example: Multi-Arm Bandits	105
12.5	Fun Fact: Sticking to the Winner	108
12.6	Limitations of an Epsilon-Decreasing Strategy	108

CONTENTS

12.7 Summary	110
Appendix	111
A. Overview of Unsupervised Learning Algorithms . . .	111
B. Overview of Supervised Learning Algorithms	112
C. List of Tuning Parameters	113
D. More Evaluation Metrics	114
Glossary	119
Data Sources and References	127
About the Authors	131