Contents

1 **Introduction to Data Science** 1
1.1 Introduction 1
1.2 History of Data Science 2
1.3 Importance of Data Science in Modern Business 4
1.4 Data Scientists 6
1.5 Data Science Activities in Three Dimensions 8
1.5.1 Managing Data Flow 8
1.5.2 Managing Data Curation 11
1.5.3 Data Analytics 14
1.6 Overlapping of Data Science with Other Fields 16
1.7 Data Analytic Thinking 17
1.8 Domains of Application 18
1.8.1 Sustainable Development for Resources 18
1.8.2 Utilization of Social Network Platform for Various Activities 19
1.8.3 Intelligent Web Applications 20
1.8.4 Google's Automatic Statistician Project 20
1.9 Application of Computational Intelligence to Manage Data Science Activities 21
1.10 Scenarios for Data Science in Business 23
1.11 Tools and Techniques Helpful for Doing Data Science 24
1.11.1 Data Cleaning Tools 25
1.11.2 Data Munging and Modelling Tools 26
1.11.3 Data Visualization Tools 28
1.12 Exercises 29
References 30

2 **Data Analytics** 31
2.1 Introduction 31
2.2 Cross Industry Standard Process 33
2.3 Data Analytics Life Cycle 34

2.4 Data Science Project Life Cycle 36
2.5 Complexity of Analytics 39
2.6 From Data to Insights 41
2.7 Building Analytics Capabilities: Case of Banking 42
2.8 Data Quality 44
2.9 Data Preparation Process 45
2.10 Communicating Analytics Outcomes 47
2.10.1 Strategies for Communicating Analytics 47
2.10.2 Data Visualization 48
2.10.3 Techniques for Visualization 50
2.11 Exercises 51
References 52

3 Basic Learning Algorithms 53
3.1 Learning from Data 53
3.2 Supervised Learning 55
3.2.1 Linear Regression 56
3.2.2 Decision Tree 58
3.2.3 Random Forest 65
3.2.4 *k*-Nearest Neighbour 66
3.2.5 Logistic Regression 69
3.2.6 Model Combiners 70
3.2.7 Naive Bayes 74
3.2.8 Bayesian Belief Networks 76
3.2.9 Support Vector Machine 77
3.3 Unsupervised Learning 80
3.3.1 Apriori Algorithm 80
3.3.2 *k*-Means Algorithm 84
3.3.3 Dimensionality Reduction for Data Compression 86
3.4 Reinforcement Learning 87
3.4.1 Markov Decision Process 90
3.5 Case Study: Using Machine Learning for Marketing Campaign 91
3.6 Exercises 92
References 93

4 Fuzzy Logic 95
4.1 Introduction 95
4.2 Fuzzy Membership Functions 98
4.2.1 Triangular Membership Function 99
4.2.2 Trapezoidal Membership Function 99
4.2.3 Gaussian Membership Function 100
4.2.4 Sigmoidal Membership Function 100
4.3 Methods of Membership Value Assignment 101
4.4 Fuzzification and Defuzzification Methods 102

4.5 Fuzzy Set Operations 102
4.5.1 Union of Fuzzy Sets 102
4.5.2 Intersection of Fuzzy Sets 103
4.5.3 Complement of a Fuzzy Set 103
4.6 Fuzzy Set Properties 105
4.7 Fuzzy Relations 106
4.7.1 Example of Operation on Fuzzy Relationship 108
4.8 Fuzzy Propositions 109
4.8.1 Fuzzy Connectives 110
4.8.2 Disjunction 110
4.8.3 Conjunction 111
4.8.4 Negation 111
4.8.5 Implication 111
4.9 Fuzzy Inference 112
4.10 Fuzzy Rule-Based System 112
4.11 Fuzzy Logic for Data Science 114
4.11.1 Application 1: Web Content Mining 116
4.11.2 Application 2: Web Structure Mining 117
4.11.3 Application 3: Web Usage Mining 118
4.11.4 Application 4: Environmental and Social Data Manipulation 119
4.12 Tools and Techniques for Doing Data Science with Fuzzy Logic 120
4.13 Exercises 122
References 122

5 Artificial Neural Network 125
5.1 Introduction 125
5.2 Symbolic Learning Methods 126
5.3 Artificial Neural Network and Its Characteristics 128
5.4 ANN Models 131
5.4.1 Hopfield Model 132
5.4.2 Perceptron Model 133
5.4.3 Multi-Layer Perceptron 136
5.4.4 Deep Learning in Multi-Layer Perceptron 139
5.4.5 Other Models of ANN 141
5.4.6 Linear Regression and Neural Networks 143
5.5 ANN Tools and Utilities 144
5.6 Emotions Mining on Social Network Platform 145
5.6.1 Related Work on Emotions Mining 146
5.6.2 Broad Architecture 146
5.6.3 Design of Neural Network 148
5.7 Applications and Challenges 149
5.8 Concerns 152
5.9 Exercises 153
References 154

6 Genetic Algorithms and Evolutionary Computing 157
6.1 Introduction 157
6.2 Genetic Algorithms 159
6.3 Basic Principles of Genetic Algorithms 161
6.3.1 Encoding Individuals 161
6.3.2 Mutation 163
6.3.3 Crossover 163
6.3.4 Fitness Function 164
6.3.5 Selection 165
6.3.6 Other Encoding Strategies 166
6.4 Example of Function Optimization using Genetic Algorithm 168
6.5 Schemata and Schema Theorem 170
6.5.1 Instance, Defined Bits, and Order of Schema 170
6.5.2 Importance of Schema 171
6.6 Application Specific Genetic Operators 171
6.6.1 Application of the Recombination Operator: Example 173
6.7 Evolutionary Programming 174
6.8 Applications of GA in Healthcare 175
6.8.1 Case of Healthcare 176
6.8.2 Patients Scheduling System Using Genetic Algorithm 177
6.8.3 Encoding of Candidates 178
6.8.4 Operations on Population 180
6.8.5 Other Applications 182
6.9 Exercises 183
References 184

7 Other Metaheuristics and Classification Approaches 185
7.1 Introduction 185
7.2 Adaptive Memory Procedure 186
7.2.1 Tabu Search 186
7.2.2 Scatter Search 188
7.2.3 Path Relinking 191
7.3 Swarm Intelligence 192
7.3.1 Ant Colony Optimization 193
7.3.2 Artificial Bee Colony Algorithm 193
7.3.3 River Formation Dynamics 195
7.3.4 Particle Swarm Optimization 196
7.3.5 Stochastic Diffusion Search 198
7.3.6 Swarm Intelligence and Big Data 199
7.4 Case-Based Reasoning 201
7.4.1 Learning in Case-Based Reasoning 203
7.4.2 Case-Based Reasoning and Data Science 205
7.4.3 Dealing with Complex Domains 205
7.5 Rough Sets 206
7.6 Exercises 208
References 209

8 Analytics and Big Data 211
8.1 Introduction 211
8.2 Traditional Versus Big Data Analytics 213
8.3 Large-Scale Parallel Processing 215
8.3.1 MapReduce 215
8.3.2 Comparison with RDBMS 218
8.3.3 Shared-Memory Parallel Programming 219
8.3.4 Apache Hadoop Ecosystem 219
8.3.5 Hadoop Distributed File System 222
8.4 NoSQL 223
8.4.1 Key-Value Model 224
8.5 SPARK 226
8.6 Data in Motion 228
8.6.1 Data Stream Processing 228
8.6.2 Real-Time Data Streams 229
8.6.3 Data Streams and DBMS 230
8.7 Scaling Up Machine Learning Algorithms 232
8.8 Privacy, Security and Ethics in Data Science 234
8.9 Exercises 235
References 236

9 Data Science Using R 237
9.1 Getting Started 237
9.2 Running Code 238
9.3 R Basics 239
9.4 Analysing Data 242
9.5 Examples 243
9.5.1 Linear Regression 244
9.5.2 Logistic Regression 244
9.5.3 Prediction 244
9.5.4 *k*-Nearest Neighbour Classification 245
9.5.5 Naive Bayes 245
9.5.6 Decision Trees (CART) 246
9.5.7 *k*-Means Clustering 247
9.5.8 Random Forest 248
9.5.9 Apriori 249
9.5.10 AdaBoost 250
9.5.11 Dimensionality Reduction 250
9.5.12 Support Vector Machine 251
9.5.13 Artificial Neural Nets 251
9.6 Visualization in R 253
9.7 Writing Your Own Functions 256
9.8 Open-Source R on Hadoop 258
References 259

Appendices 261
Appendix I: Tools for Data Science 261
BigML 261
Python 262
Natural Language Toolkit 262
DataWrangler 262
OpenRefine 262
Datawrapper 263
Orange 263
RapidMiner 263
Tanagra 263
Weka 264
KNIME 264
Apache Mahout 264
Hive 264
Scikit-Learn 265
D3.js 265
Pandas 265
Tableau Public 265
Exhibit 266
Gephi 266
NodeXL 266
Leaflet 267
Classias 267
Appendix II: Tools for Computational Intelligence 267
NeuroXL 267
Plug&Score 268
Multiple Back-Propagation (MBP) 268
A.I. Solver Studio 268
The MathWorks – Neural Network Toolbox 268
Visual Numerics Java Numerical Library 269
Stuttgart Neural Network Simulator 269
FANN (Fast Artificial Neural Network Library) 269
NeuroIntelligence – Alyuda Research 269
EasyNN-Plus 270
NeuroDimension – Neural Network Software 270
BrainMaker – California Scientific Software 270
Classification & Prediction Tools in Excel 270
SIMBRAIN 271
DELVE 271
SkyMind 271
Prediction.io 271
Parallel Genetic Algorithm Library (pgapack) 271
Parallel PIKAIA 272
Evolving Objects (EO): An Evolutionary Computation Framework 272