Table of Contents

Preface 1

Section 1: Fundamentals and Core Algorithms

Chapter 1: Overview of Algorithms 11
What is an algorithm? 12
The phases of an algorithm 12
Specifying the logic of an algorithm 14
Understanding pseudocode 14
A practical example of pseudocode 15
Using snippets 16
Creating an execution plan 17
Introducing Python packages 18
Python packages 18
The SciPy ecosystem 19
Implementing Python via the Jupyter Notebook 20
Algorithm design techniques 21
The data dimension 22
Compute dimension 23
A practical example 23
Performance analysis 24
Space complexity analysis 25
Time complexity analysis 25
Estimating the performance 26
The best case 26
The worst case 27
The average case 27
Selecting an algorithm 27
Big O notation 28
Constant time (O(1)) complexity 28
Linear time (O(n)) complexity 29
Quadratic time (O(n2)) complexity 29
Logarithmic time (O(logn)) complexity 30
Validating an algorithm 32
Exact, approximate, and randomized algorithms 32
Explainability 33
Summary 34

Chapter 2: Data Structures Used in Algorithms 35
Exploring data structures in Python 36
List 36

Using lists 36
Lambda functions 39
The range function 40
The time complexity of lists 41
Tuples 41
The time complexity of tuples 42
Dictionary 42
The time complexity of a dictionary 44
Sets 44
Time complexity analysis for sets 45
DataFrames 46
Terminologies of DataFrames 46
Creating a subset of a DataFrame 47
Column selection 47
Row selection 48
Matrix 48
Matrix operations 49
Exploring abstract data types 49
Vector 50
Stacks 50
The time complexity of stacks 52
Practical example 53
Queues 53
The basic idea behind the use of stacks and queues 55
Tree 55
Terminology 56
Types of trees 56
Practical examples 58
Summary 58

Chapter 3: Sorting and Searching Algorithms 59
Introducing Sorting Algorithms 60
Swapping Variables in Python 60
Bubble Sort 61
Understanding the Logic Behind Bubble Sort 61
A Performance Analysis of Bubble Sort 63
Insertion Sort 63
Merge Sort 65
Shell Sort 67
A Performance Analysis of Shell Sort 68
Selection Sort 69
The performance of the selection sort algorithm 70
Choosing a sorting algorithm 70
Introduction to Searching Algorithms 70
Linear Search 71
The Performance of Linear Search 71
Binary Search 72
The Performance of Binary Search 72

Interpolation Search 73
The Performance of Interpolation Search 73
Practical Applications 74
Summary 76

Chapter 4: Designing Algorithms 77
Introducing the basic concepts of designing an algorithm 78
Concern 1 – Will the designed algorithm produce the result we expect? 79
Concern 2 – Is this the optimal way to get these results? 79
Characterizing the complexity of the problem 80
Concern 3 – How is the algorithm going to perform on larger datasets? 83
Understanding algorithmic strategies 83
Understanding the divide-and-conquer strategy 84
Practical example – divide-and-conquer applied to Apache Spark 84
Understanding the dynamic programming strategy 86
Understanding greedy algorithms 87
Practical application – solving the TSP 88
Using a brute-force strategy 89
Using a greedy algorithm 93
Presenting the PageRank algorithm 95
Problem definition 95
Implementing the PageRank algorithm 95
Understanding linear programming 98
Formulating a linear programming problem 98
Defining the objective function 98
Specifying constraints 99
Practical application – capacity planning with linear programming 99
Summary 101

Chapter 5: Graph Algorithms 103
Representations of graphs 104
Types of graphs 105
Undirected graphs 106
Directed graphs 106
Undirected multigraphs 107
Directed multigraphs 107
Special types of edges 108
Ego-centered networks 108
Social network analysis 109
Introducing network analysis theory 110
Understanding the shortest path 111
Creating a neighborhood 111
Triangles 112
Density 112
Understanding centrality measures 112
Degree 113
Betweenness 114

Fairness and closeness 114
Eigenvector centrality 115
Calculating centrality metrics using Python 115
Understanding graph traversals 117
Breadth-first search 117
Initialization 118
The main loop 118
Depth-first search 121
Case study – fraud analytics 124
Conducting simple fraud analytics 126
Presenting the watchtower fraud analytics methodology 127
Scoring negative outcomes 128
Degree of suspicion 128
Summary 130

Section 2: Machine Learning Algorithms

Chapter 6: Unsupervised Machine Learning Algorithms 133
Introducing unsupervised learning 134
Unsupervised learning in the data-mining life cycle 134
Current research trends in unsupervised learning 137
Practical examples 137
Voice categorization 138
Document categorization 138
Understanding clustering algorithms 139
Quantifying similarities 139
Euclidean distance 140
Manhattan distance 141
Cosine distance 142
K-means clustering algorithm 143
The logic of k-means clustering 143
Initialization 143
The steps of the k-means algorithm 144
Stop condition 145
Coding the k-means algorithm 145
Limitation of k-means clustering 147
Hierarchical clustering 148
Steps of hierarchical clustering 148
Coding a hierarchical clustering algorithm 149
Evaluating the clusters 150
Application of clustering 150
Dimensionality reduction 151
Principal component analysis 152
Limitations of PCA 155
Association rules mining 155
Examples of use 155
Market basket analysis 156
Association rules 157

Types of rule 157
Trivial rules 157
Inexplicable rules 158
Actionable rules 158
Ranking rules 159
Support 159
Confidence 160
Lift 160
Algorithms for association analysis 161
Apriori Algorithm 161
Limitations of the apriori algorithm 161
FP-growth algorithm 162
Populating the FP-tree 162
Mining Frequent Patterns 164
Code for using FP-growth 165
Practical application– clustering similar tweets together 167
Topic modeling 168
Clustering 168
Anomaly-detection algorithms 168
Using clustering 169
Using density-based anomaly detection 169
Using support vector machines 169
Summary 170

Chapter 7: Traditional Supervised Learning Algorithms 171
Understanding supervised machine learning 172
Formulating supervised machine learning 173
Understanding enabling conditions 175
Differentiating between classifiers and regressors 176
Understanding classification algorithms 176
Presenting the classifiers challenge 177
The problem statement 177
Feature engineering using a data processing pipeline 178
Importing data 178
Feature selection 179
One-hot encoding 179
Specifying the features and label 180
Dividing the dataset into testing and training portions 180
Scaling the features 181
Evaluating the classifiers 181
Confusion matrix 182
Performance metrics 182
Understanding overfitting 184
Bias 184
Variance 184
Bias-variance trade-off 184
Specifying the phases of classifiers 185
Decision tree classification algorithm 186
Understanding the decision tree classification algorithm 187
Using the decision tree classification algorithm for the classifiers challenge 188

The strengths and weaknesses of decision tree classifiers 189
Strengths 189
Weaknesses 189
Use cases 190
Classifying records 190
Feature selection 190
Understanding the ensemble methods 190
Implementing gradient boosting with the XGBoost algorithm 191
Using the random forest algorithm 192
Training a random forest algorithm 192
Using random forest for predictions 192
Differentiating the random forest algorithm from ensemble boosting 193
Using the random forest algorithm for the classifiers challenge 193
Logistic regression 195
Assumptions 195
Establishing the relationship 195
The loss and cost functions 196
When to use logistic regression 196
Using the logistic regression algorithm for the classifiers challenge 197
The SVM algorithm 198
Using the SVM algorithm for the classifiers challenge 199
Understanding the naive Bayes algorithm 199
Bayes, theorem 200
Calculating probabilities 200
Multiplication rules for AND events 201
The general multiplication rule 201
Addition rules for OR events 201
Using the naive Bayes algorithm for the classifiers challenge 202
For classification algorithms, the winner is... 203
Understanding regression algorithms 203
Presenting the regressors challenge 204
The problem statement of the regressors challenge 204
Exploring the historical dataset 204
Feature engineering using a data processing pipeline 205
Linear regression 206
Simple linear regression 206
Evaluating the regressors 208
Multiple regression 209
Using the linear regression algorithm for the regressors challenge 209
When is linear regression used? 210
The weaknesses of linear regression 210
The regression tree algorithm 211
Using the regression tree algorithm for the regressors challenge 211
The gradient boost regression algorithm 212
Using gradient boost regression algorithm for the regressors challenge 212
For regression algorithms, the winner is... 213
Practical example – how to predict the weather 213
Summary 216
Chapter 8: Neural Network Algorithms 217

Understanding ANNs 218
The Evolution of ANNs 220
Training a Neural Network 222
Understanding the Anatomy of a Neural Network 222
Defining Gradient Descent 223
Activation Functions 226
Threshold Function 226
Sigmoid 227
Rectified linear unit (ReLU) 228
Leaky ReLU 229
Hyperbolic tangent (tanh) 230
Softmax 231
Tools and Frameworks 232
Keras 232
Backend Engines of Keras 232
Low-level layers of the deep learning stack 233
Defining hyperparameters 233
Defining a Keras model 234
Choosing sequential or functional model 236
Understanding TensorFlow 236
Presenting TensorFlow's Basic Concepts 236
Understanding Tensor Mathematics 237
Understanding the Types of Neural Networks 238
Convolutional Neural Networks 239
Convolution 239
Pooling 239
Recurrent Neural Networks 240
Generative Adversarial Networks 240
Transfer Learning 240
Case study – using deep learning for fraud detection 241
Methodology 242
Summary 246

Chapter 9: Algorithms for Natural Language Processing 247
Introducing NLP 248
Understanding NLP terminology 248
Normalization 248
Corpus 249
Tokenization 249
Named entity recognition 249
Stopwords 249
Sentiment analysis 250
Stemming and lemmatization 250
NLTK 251
BoW-based NLP 251
Introduction to word embedding 254
The neighborhood of a word 255
Properties of word embeddings 255

Using RNNs for NLP 256
Using NLP for sentiment analysis 257
Case study: movie review sentiment analysis 259
Summary 262

Chapter 10: Recommendation Engines 263
Introducing recommendation systems 264
Types of recommendation engines 264
Content-based recommendation engines 264
Finding similarities between unstructured documents 265
Using a co-occurrence matrix 266
Collaborative filtering recommendation engines 267
Hybrid recommendation engines 269
Generating a similarity matrix of the items 269
Generating reference vectors of the users 270
Generating recommendations 270
Understanding the limitations of recommender systems 271
The cold start problem 271
Metadata requirements 272
The data sparsity problem 272
Bias due to social influence 272
Limited data 272
Areas of practical applications 272
Practical example – creating a recommendation engine 273
Summary 276

Section 3: Advanced Topics

Chapter 11: Data Algorithms 279
Introduction to data algorithms 279
Data classification 280
Presenting data storage algorithms 281
Understanding data storage strategies 281
Presenting the CAP theorem 281
CA systems 282
AP systems 283
CP systems 283
Presenting streaming data algorithms 284
Applications of streaming 284
Presenting data compression algorithms 284
Lossless compression algorithms 285
Understanding the basic techniques of lossless compression 285
Huffman coding 286
A practical example – Twitter real-time sentiment analysis 287
Summary 291

Chapter 12: Cryptography 293

Introduction to Cryptography 294
Understanding the Importance of the Weakest Link 294
The Basic Terminology 295
Understanding the Security Requirements 295
Identifying the Entities 296
Establishing the Security Goals 296
Understanding the Sensitivity of the Data 297
Understanding the Basic Design of Ciphers 297
Presenting Substitution Ciphers 298
Understanding Transposition Ciphers 300
Understanding the Types of Cryptographic Techniques 301
Using the Cryptographic Hash Function 301
Implementing cryptographic hash functions 302
Understanding MD5-tolerated 302
Understanding SHA 303
An Application of the Cryptographic Hash Function 304
Using Symmetric Encryption 304
Coding Symmetric Encryption 305
The Advantages of Symmetric Encryption 306
The Problems with Symmetric Encryption 306
Asymmetric Encryption 306
The SSL/TLS Handshaking Algorithm 307
Public Key Infrastructure 309
Example – Security Concerns When Deploying a Machine Learning Model 310
MITM attacks 311
How to prevent MITM attacks 312
Avoiding Masquerading 313
Data and Model Encrpytion 313
Summary 316

Chapter 13: Large-Scale Algorithms 317
Introduction to large-scale algorithms 318
Defining a well-designed, large-scale algorithm 318
Terminology 318
Latency 318
Throughput 319
Network bisection bandwidth 319
Elasticity 319
The design of parallel algorithms 319
Amdahl's law 320
Conducting sequential process analysis 320
Conducting parallel execution analysis 321
Understanding task granularity 323
Load balancing 323
Locality issues 324
Enabling concurrent processing in Python 324
Strategizing multi-resource processing 324

Introducing CUDA 325
Designing parallel algorithms on CUDA 326
Using GPUs for data processing in Python 327
Cluster computing 328
Implementing data processing in Apache Spark 329
The hybrid strategy 331
Summary 331

Chapter 14: Practical Considerations 333
Introducing practical considerations 334
The sad story of an AI Twitter Bot 335
The explainability of an algorithm 335
Machine learning algorithms and explainability 336
Presenting strategies for explainability 336
Implementing explainability 337
Understanding ethics and algorithms 340
Problems with learning algorithms 340
Understanding ethical considerations 341
Inconclusive evidence 341
Traceability 342
Misguided evidence 342
Unfair outcomes 342
Reducing bias in models 342
Tackling NP-hard problems 343
Simplifying the problem 344
Example 344
Customizing a well-known solution to a similar problem 344
Example 344
Using a probabilistic method 345
Example 345
When to use algorithms 345
A practical example – black swan events 346
Four criteria to classify an event as a black swan event 346
Applying algorithms to black swan events 347
Summary 348

Other Books You May Enjoy 349

Index 353