Contents

Preface xvii

1 Introduction: Becoming a Unicorn 1
  1.1 Aren’t Data Scientists Just Overpaid Statisticians? 2
  1.2 How Is This Book Organized? 3
  1.3 How to Use This Book? 3
  1.4 Why Is It All in Python™, Anyway? 4
  1.5 Example Code and Datasets 4
  1.6 Parting Words 5

Part I The Stuff You’ll Always Use 7

2 The Data Science Road Map 9
  2.1 Frame the Problem 10
  2.2 Understand the Data: Basic Questions 11
  2.3 Understand the Data: Data Wrangling 12
  2.4 Understand the Data: Exploratory Analysis 13
  2.5 Extract Features 14
  2.6 Model 15
  2.7 Present Results 15
  2.8 Deploy Code 16
  2.9 Iterating 16
  2.10 Glossary 17

3 Programming Languages 19
  3.1 Why Use a Programming Language? What Are the Other Options? 19
  3.2 A Survey of Programming Languages for Data Science 20
    3.2.1 Python 20
    3.2.2 R 21
    3.2.3 MATLAB® and Octave 21
    3.2.4 SAS® 21
3.2.5 Scala 22
3.3 Python Crash Course 22
3.3.1 A Note on Versions 22
3.3.2 “Hello World” Script 23
3.3.3 More Complicated Script 23
3.3.4 Atomic Data Types 26
3.4 Strings 27
3.4.1 Comments and Docstrings 28
3.4.2 Complex Data Types 29
3.4.3 Lists 29
3.4.4 Strings and Lists 30
3.4.5 Tuples 31
3.4.6 Dictionaries 31
3.4.7 Sets 32
3.5 Defining Functions 32
3.5.1 For Loops and Control Structures 33
3.5.2 A Few Key Functions 34
3.5.3 Exception Handling 35
3.5.4 Libraries 35
3.5.5 Classes and Objects 35
3.5.6 GOTCHA: Hashable and Unhashable Types 36
3.6 Python’s Technical Libraries 37
3.6.1 Data Frames 38
3.6.2 Series 39
3.6.3 Joining and Grouping 40
3.7 Other Python Resources 42
3.8 Further Reading 42
3.9 Glossary 43

3a Interlude: My Personal Toolkit 45

4 Data Munging: String Manipulation, Regular Expressions, and Data Cleaning 47
4.1 The Worst Dataset in the World 48
4.2 How to Identify Pathologies 48
4.3 Problems with Data Content 49
4.3.1 Duplicate Entries 49
4.3.2 Multiple Entries for a Single Entity 49
4.3.3 Missing Entries 49
4.3.4 NULLs 50
4.3.5 Huge Outliers 50
4.3.6 Out-of-Date Data 50
4.3.7 Artificial Entries 50
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.8</td>
<td>Irregular Spacings</td>
<td>51</td>
</tr>
<tr>
<td>4.4</td>
<td>Formatting Issues</td>
<td>51</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Formatting Is Irregular between Different</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Tables/Columns</td>
<td></td>
</tr>
<tr>
<td>4.4.2</td>
<td>Extra Whitespace</td>
<td>51</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Irregular Capitalization</td>
<td>52</td>
</tr>
<tr>
<td>4.4.4</td>
<td>Inconsistent Delimiters</td>
<td>52</td>
</tr>
<tr>
<td>4.4.5</td>
<td>Irregular NULL Format</td>
<td>52</td>
</tr>
<tr>
<td>4.4.6</td>
<td>Invalid Characters</td>
<td>52</td>
</tr>
<tr>
<td>4.4.7</td>
<td>Weird or Incompatible Datetimes</td>
<td>52</td>
</tr>
<tr>
<td>4.4.8</td>
<td>Operating System Incompatibilities</td>
<td>53</td>
</tr>
<tr>
<td>4.4.9</td>
<td>Wrong Software Versions</td>
<td>53</td>
</tr>
<tr>
<td>4.5</td>
<td>Example Formatting Script</td>
<td>54</td>
</tr>
<tr>
<td>4.6</td>
<td>Regular Expressions</td>
<td>55</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Regular Expression Syntax</td>
<td>56</td>
</tr>
<tr>
<td>4.7</td>
<td>Life in the Trenches</td>
<td>60</td>
</tr>
<tr>
<td>4.8</td>
<td>Glossary</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>Visualizations and Simple Metrics</td>
<td>61</td>
</tr>
<tr>
<td>5.1</td>
<td>A Note on Python’s Visualization Tools</td>
<td>62</td>
</tr>
<tr>
<td>5.2</td>
<td>Example Code</td>
<td>62</td>
</tr>
<tr>
<td>5.3</td>
<td>Pie Charts</td>
<td>63</td>
</tr>
<tr>
<td>5.4</td>
<td>Bar Charts</td>
<td>65</td>
</tr>
<tr>
<td>5.5</td>
<td>Histograms</td>
<td>66</td>
</tr>
<tr>
<td>5.6</td>
<td>Means, Standard Deviations, Medians, and</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Quantiles</td>
<td></td>
</tr>
<tr>
<td>5.7</td>
<td>Boxplots</td>
<td>70</td>
</tr>
<tr>
<td>5.8</td>
<td>Scatterplots</td>
<td>72</td>
</tr>
<tr>
<td>5.9</td>
<td>Scatterplots with Logarithmic Axes</td>
<td>74</td>
</tr>
<tr>
<td>5.10</td>
<td>Scatter Matrices</td>
<td>76</td>
</tr>
<tr>
<td>5.11</td>
<td>Heatmaps</td>
<td>77</td>
</tr>
<tr>
<td>5.12</td>
<td>Correlations</td>
<td>78</td>
</tr>
<tr>
<td>5.13</td>
<td>Anscombe’s Quartet and the Limits of</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Numbers</td>
<td></td>
</tr>
<tr>
<td>5.14</td>
<td>Time Series</td>
<td>81</td>
</tr>
<tr>
<td>5.15</td>
<td>Further Reading</td>
<td>85</td>
</tr>
<tr>
<td>5.16</td>
<td>Glossary</td>
<td>85</td>
</tr>
<tr>
<td>6</td>
<td>Machine Learning Overview</td>
<td>87</td>
</tr>
<tr>
<td>6.1</td>
<td>Historical Context</td>
<td>88</td>
</tr>
<tr>
<td>6.2</td>
<td>Supervised versus Unsupervised</td>
<td>89</td>
</tr>
<tr>
<td>6.3</td>
<td>Training Data, Testing Data, and the Great</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Boogeyman of Overfitting</td>
<td></td>
</tr>
<tr>
<td>6.4</td>
<td>Further Reading</td>
<td>91</td>
</tr>
<tr>
<td>6.5</td>
<td>Glossary</td>
<td>91</td>
</tr>
</tbody>
</table>
7 Interlude: Feature Extraction Ideas 93
  7.1 Standard Features 93
  7.2 Features That Involve Grouping 94
  7.3 Preview of More Sophisticated Features 95
  7.4 Defining the Feature You Want to Predict 95

8 Machine Learning Classification 97
  8.1 What Is a Classifier, and What Can You Do with It? 97
  8.2 A Few Practical Concerns 98
  8.3 Binary versus Multiclass 99
  8.4 Example Script 99
  8.5 Specific Classifiers 101
    8.5.1 Decision Trees 101
    8.5.2 Random Forests 103
    8.5.3 Ensemble Classifiers 104
    8.5.4 Support Vector Machines 105
    8.5.5 Logistic Regression 108
    8.5.6 Lasso Regression 110
    8.5.7 Naive Bayes 110
    8.5.8 Neural Nets 112
  8.6 Evaluating Classifiers 114
    8.6.1 Confusion Matrices 114
    8.6.2 ROC Curves 115
    8.6.3 Area under the ROC Curve 116
  8.7 Selecting Classification Cutoffs 117
    8.7.1 Other Performance Metrics 118
    8.7.2 Lift–Reach Curves 118
  8.8 Further Reading 119
  8.9 Glossary 119

9 Technical Communication and Documentation 121
  9.1 Several Guiding Principles 122
    9.1.1 Know Your Audience 122
    9.1.2 Show Why It Matters 122
    9.1.3 Make It Concrete 123
    9.1.4 A Picture Is Worth a Thousand Words 123
    9.1.5 Don’t Be Arrogant about Your Tech Knowledge 124
    9.1.6 Make It Look Decent 124
  9.2 Slide Decks 124
    9.2.1 C.R.A.P. Design 125
    9.2.2 A Few Tips and Rules of Thumb 127
  9.3 Written Reports 128
  9.4 Speaking: What Has Worked for Me 130
Contents

9.5 Code Documentation 131
9.6 Further Reading 132
9.7 Glossary 132

Part II Stuff You Still Need to Know 133

10 Unsupervised Learning: Clustering and Dimensionality Reduction 135
10.1 The Curse of Dimensionality 136
10.2 Example: Eigenfaces for Dimensionality Reduction 138
10.3 Principal Component Analysis and Factor Analysis 140
10.4 Skree Plots and Understanding Dimensionality 142
10.5 Factor Analysis 143
10.6 Limitations of PCA 143
10.7 Clustering 144
10.7.1 Real-World Assessment of Clusters 144
10.7.2 $k$-Means Clustering 145
10.7.3 Gaussian Mixture Models 146
10.7.4 Agglomerative Clustering 147
10.7.5 Evaluating Cluster Quality 148
10.7.6 Silhouette Score 148
10.7.7 Rand Index and Adjusted Rand Index 149
10.7.8 Mutual Information 150
10.8 Further Reading 151
10.9 Glossary 151

11 Regression 153
11.1 Example: Predicting Diabetes Progression 153
11.2 Least Squares 156
11.3 Fitting Nonlinear Curves 157
11.4 Goodness of Fit: $R^2$ and Correlation 159
11.5 Correlation of Residuals 160
11.6 Linear Regression 161
11.7 LASSO Regression and Feature Selection 162
11.8 Further Reading 164
11.9 Glossary 164

12 Data Encodings and File Formats 165
12.1 Typical File Format Categories 165
12.1.1 Text Files 166
12.1.2 Dense Numerical Arrays 166
12.1.3 Program-Specific Data Formats 166
12.1.4 Compressed or Archived Data 166
12.2 CSV Files 167
12.3 JSON Files 168
12.4 XML Files 170
12.5 HTML Files 172
12.6 Tar Files 174
12.7 GZip Files 175
12.8 Zip Files 175
12.9 Image Files: Rasterized, Vectorized, and/or Compressed 176
12.10 It's All Bytes at the End of the Day 177
12.11 Integers 178
12.12 Floats 179
12.13 Text Data 180
12.14 Further Reading 183
12.15 Glossary 183

13 Big Data 185
13.1 What Is Big Data? 185
13.2 Hadoop: The File System and the Processor 187
13.3 Using HDFS 188
13.4 Example PySpark Script 189
13.5 Spark Overview 190
13.6 Spark Operations 192
13.7 Two Ways to Run PySpark 193
13.8 Configuring Spark 194
13.9 Under the Hood 195
13.10 Spark Tips and Gotchas 196
13.11 The MapReduce Paradigm 197
13.12 Performance Considerations 199
13.13 Further Reading 200
13.14 Glossary 200

14 Databases 203
14.1 Relational Databases and MySQL® 204
14.1.1 Basic Queries and Grouping 204
14.1.2 Joins 207
14.1.3 Nesting Queries 208
14.1.4 Running MySQL and Managing the DB 209
14.2 Key-Value Stores 210
14.3 Wide Column Stores 211
14.4 Document Stores 211
14.4.1 MongoDB® 212
14.5 Further Reading 214
14.6 Glossary 214
17.11 Fourier Analysis: Sometimes a Magic Bullet 256
17.12 Time Series in Context: The Whole Suite of Features 259
17.13 Further Reading 259
17.14 Glossary 260

18 Probability 261
18.1 Flipping Coins: Bernoulli Random Variables 261
18.2 Throwing Darts: Uniform Random Variables 263
18.3 The Uniform Distribution and Pseudorandom Numbers 263
18.4 Nondiscrete, Noncontinuous Random Variables 265
18.5 Notation, Expectations, and Standard Deviation 267
18.6 Dependence, Marginal and Conditional Probability 268
18.7 Understanding the Tails 269
18.8 Binomial Distribution 271
18.9 Poisson Distribution 272
18.10 Normal Distribution 272
18.11 Multivariate Gaussian 273
18.12 Exponential Distribution 274
18.13 Log-Normal Distribution 276
18.14 Entropy 277
18.15 Further Reading 279
18.16 Glossary 279

19 Statistics 281
19.1 Statistics in Perspective 281
19.2 Bayesian versus Frequentist: Practical Tradeoffs and Differing Philosophies 282
19.3 Hypothesis Testing: Key Idea and Example 283
19.4 Multiple Hypothesis Testing 285
19.5 Parameter Estimation 286
19.6 Hypothesis Testing: t-Test 287
19.7 Confidence Intervals 290
19.8 Bayesian Statistics 291
19.9 Naive Bayesian Statistics 293
19.10 Bayesian Networks 293
19.11 Choosing Priors: Maximum Entropy or Domain Knowledge 294
19.12 Further Reading 295
19.13 Glossary 295

20 Programming Language Concepts 297
20.1 Programming Paradigms 297
20.1.1 Imperative 298
20.1.2 Functional 298
Contents

20.1.3 Object-Oriented 301
20.2 Compilation and Interpretation 305
20.3 Type Systems 307
20.3.1 Static versus Dynamic Typing 308
20.3.2 Strong versus Weak Typing 308
20.4 Further Reading 309
20.5 Glossary 309

21 Performance and Computer Memory 311
21.1 Example Script 311
21.2 Algorithm Performance and Big-O Notation 314
21.3 Some Classic Problems: Sorting a List and Binary Search 315
21.4 Amortized Performance and Average Performance 318
21.5 Two Principles: Reducing Overhead and Managing Memory 320
21.6 Performance Tip: Use Numerical Libraries When Applicable 322
21.7 Performance Tip: Delete Large Structures You Don’t Need 323
21.8 Performance Tip: Use Built-In Functions When Possible 324
21.9 Performance Tip: Avoid Superfluous Function Calls 324
21.10 Performance Tip: Avoid Creating Large New Objects 325
21.11 Further Reading 325
21.12 Glossary 325

Part III Specialized or Advanced Topics 327

22 Computer Memory and Data Structures 329
22.1 Virtual Memory, the Stack, and the Heap 329
22.2 Example C Program 330
22.3 Data Types and Arrays in Memory 330
22.4 Structs 332
22.5 Pointers, the Stack, and the Heap 333
22.6 Key Data Structures 337
22.6.1 Strings 337
22.6.2 Adjustable-Size Arrays 338
22.6.3 Hash Tables 339
22.6.4 Linked Lists 340
22.6.5 Binary Search Trees 342
22.7 Further Reading 343
22.8 Glossary 343

23 Maximum Likelihood Estimation and Optimization 345
23.1 Maximum Likelihood Estimation 345
23.2 A Simple Example: Fitting a Line 346