

# Table of Contents

<b>Preface</b>	<b>v</b>
<b>Chapter 1: First Steps</b>	<b>1</b>
<b>Introducing data science and Python</b>	<b>2</b>
<b>Installing Python</b>	<b>3</b>
Python 2 or Python 3?	3
Step-by-step installation	4
A glance at the essential Python packages	5
NumPy	5
SciPy	6
pandas	6
Scikit-learn	6
IPython	7
Matplotlib	7
Statsmodels	8
Beautiful Soup	8
NetworkX	8
NLTK	9
Gensim	9
PyPy	9
The installation of packages	10
Package upgrades	11
<b>Scientific distributions</b>	<b>12</b>
Anaconda	12
Enthought Canopy	13
PythonXY	13
WinPython	13
<b>Introducing IPython</b>	<b>13</b>
The IPython Notebook	15

<b>Datasets and code used in the book</b>	<b>22</b>
Scikit-learn toy datasets	22
The MLdata.org public repository	26
LIBSVM data examples	26
Loading data directly from CSV or text files	27
Scikit-learn sample generators	30
<b>Summary</b>	<b>31</b>
<b>Chapter 2: Data Munging</b>	<b>33</b>
<b>The data science process</b>	<b>34</b>
<b>Data loading and preprocessing with pandas</b>	<b>35</b>
Fast and easy data loading	35
Dealing with problematic data	38
Dealing with big datasets	41
Accessing other data formats	45
Data preprocessing	47
Data selection	49
<b>Working with categorical and textual data</b>	<b>52</b>
<del>A special type of data — text</del>	54
<b>Data processing with NumPy</b>	<b>60</b>
NumPy's n-dimensional array	61
The basics of NumPy ndarray objects	62
<b>Creating NumPy arrays</b>	<b>63</b>
From lists to unidimensional arrays	63
Controlling the memory size	64
Heterogeneous lists	65
From lists to multidimensional arrays	66
Resizing arrays	68
Arrays derived from NumPy functions	69
Getting an array directly from a file	71
Extracting data from pandas	71
<b>NumPy fast operation and computations</b>	<b>72</b>
Matrix operations	75
Slicing and indexing with NumPy arrays	76
Stacking NumPy arrays	79
<b>Summary</b>	<b>81</b>
<b>Chapter 3: The Data Science Pipeline</b>	<b>83</b>
<b>Introducing EDA</b>	<b>83</b>
<b>Feature creation</b>	<b>87</b>
<b>Dimensionality reduction</b>	<b>90</b>
The covariance matrix	90
Principal Component Analysis (PCA)	91

A variation of PCA for big data—randomized PCA	95
Latent Factor Analysis (LFA)	96
Linear Discriminant Analysis (LDA)	97
Latent Semantical Analysis (LSA)	97
Independent Component Analysis (ICA)	98
Kernel PCA	99
Restricted Boltzmann Machine (RBM)	100
<b>The detection and treatment of outliers</b>	<b>102</b>
Univariate outlier detection	103
EllipticEnvelope	105
OneClassSVM	110
<b>Scoring functions</b>	<b>114</b>
Multilabel classification	114
Binary classification	116
Regression	117
<b>Testing and validating</b>	<b>118</b>
<b>Cross-validation</b>	<b>123</b>
Using cross-validation iterators	125
Sampling and bootstrapping	127
<b>Hyper-parameters' optimization</b>	<b>129</b>
<del>Building custom scoring functions</del>	<del>132</del>
<del>Reducing the grid search runtime</del>	<del>135</del>
<b>Feature selection</b>	<b>136</b>
Univariate selection	137
Recursive elimination	139
Stability and L1-based selection	140
<b>Summary</b>	<b>142</b>
<b>Chapter 4: Machine Learning</b>	<b>143</b>
<b>Linear and logistic regression</b>	<b>143</b>
<b>Naïve Bayes</b>	<b>147</b>
<b>The k-Nearest Neighbors</b>	<b>150</b>
<b>Advanced nonlinear algorithms</b>	<b>152</b>
SVM for classification	152
SVM for regression	155
Tuning SVM	156
<b>Ensemble strategies</b>	<b>158</b>
Pasting by random samples	158
Bagging with weak ensembles	159
Random Subspaces and Random Patches	160
Sequences of models – AdaBoost	162

<b>Gradient tree boosting (GTB)</b>	<b>162</b>
Dealing with big data	163
Creating some big datasets as examples	164
Scalability with volume	165
Keeping up with velocity	167
Dealing with variety	169
A quick overview of Stochastic Gradient Descent (SGD)	171
<b>A peek into Natural Language Processing (NLP)</b>	<b>172</b>
Word tokenization	173
Stemming	174
Word Tagging	174
Named Entity Recognition (NER)	175
Stopwords	176
A complete data science example – text classification	177
<b>An overview of unsupervised learning</b>	<b>179</b>
Summary	184
<b>Chapter 5: Social Network Analysis</b>	<b>187</b>
Introduction to graph theory	187
Graph algorithms	192
Graph loading, dumping, and sampling	199
Summary	203
<b>Chapter 6: Visualization</b>	<b>205</b>
<b>Introducing the basics of matplotlib</b>	<b>205</b>
Curve plotting	206
Using panels	208
Scatterplots	209
Histograms	210
Bar graphs	212
Image visualization	213
<b>Selected graphical examples with pandas</b>	<b>215</b>
Boxplots and histograms	216
Scatterplots	218
Parallel coordinates	221
<b>Advanced data learning representation</b>	<b>221</b>
Learning curves	222
Validation curves	224
<b>Feature importance</b>	<b>225</b>
GBT partial dependence plot	227
Summary	228
<b>Index</b>	<b>231</b>