

Python[®] Machine Learning

Wei-Meng Lee

WILEY

Contents

Introduction	xxiii
Chapter 1 Introduction to Machine Learning	1
What Is Machine Learning?	2
What Problems Will Machine Learning Be Solving in This Book?	3
Classification	4
Regression	4
Clustering	5
Types of Machine Learning Algorithms	5
Supervised Learning	5
Unsupervised Learning	7
Getting the Tools	8
Obtaining Anaconda	8
Installing Anaconda	9
Running Jupyter Notebook for Mac	9
Running Jupyter Notebook for Windows	10
Creating a New Notebook	11
Naming the Notebook	12
Adding and Removing Cells	13
Running a Cell	14
Restarting the Kernel	16
Exporting Your Notebook	16
Getting Help	17
Chapter 2 Extending Python Using NumPy	19
What Is NumPy?	19
Creating NumPy Arrays	20
Array Indexing	22

	Boolean Indexing	22
	Slicing Arrays	23
	NumPy Slice Is a Reference	25
	Reshaping Arrays	26
	Array Math	27
	Dot Product	29
	Matrix	30
	Cumulative Sum	31
	NumPy Sorting	32
	Array Assignment	34
	Copying by Reference	34
	Copying by View (Shallow Copy)	36
	Copying by Value (Deep Copy)	37
Chapter 3	Manipulating Tabular Data Using Pandas	39
	What Is Pandas?	39
	Pandas Series	40
	Creating a Series Using a Specified Index	41
	Accessing Elements in a Series	41
	Specifying a Datetime Range as the Index of a Series	42
	Date Ranges	43
	Pandas DataFrame	45
	Creating a DataFrame	45
	Specifying the Index in a DataFrame	46
	Generating Descriptive Statistics on the DataFrame	47
	Extracting from DataFrames	49
	Selecting the First and Last Five Rows	49
	Selecting a Specific Column in a DataFrame	50
	Slicing Based on Row Number	50
	Slicing Based on Row and Column Numbers	51
	Slicing Based on Labels	52
	Selecting a Single Cell in a DataFrame	54
	Selecting Based on Cell Value	54
	Transforming DataFrames	54
	Checking to See If a Result Is a DataFrame or Series	55
	Sorting Data in a DataFrame	55
	Sorting by Index	55
	Sorting by Value	56
	Applying Functions to a DataFrame	57
	Adding and Removing Rows and Columns in a DataFrame	60
	Adding a Column	61
	Removing Rows	61
	Removing Columns	62
	Generating a Crosstab	63
Chapter 4	Data Visualization Using matplotlib	67
	What Is matplotlib?	67
	Plotting Line Charts	68

Adding Title and Labels	69
Styling	69
Plotting Multiple Lines in the Same Chart	71
Adding a Legend	72
Plotting Bar Charts	73
Adding Another Bar to the Chart	74
Changing the Tick Marks	75
Plotting Pie Charts	77
Exploding the Slices	78
Displaying Custom Colors	79
Rotating the Pie Chart	80
Displaying a Legend	81
Saving the Chart	82
Plotting Scatter Plots	83
Combining Plots	83
Subplots	84
Plotting Using Seaborn	85
Displaying Categorical Plots	86
Displaying Lmplots	88
Displaying Swarmplots	90
Chapter 5 Getting Started with Scikit-learn for Machine Learning	93
Introduction to Scikit-learn	93
Getting Datasets	94
Using the Scikit-learn Dataset	94
Using the Kaggle Dataset	97
Using the UCI (University of California, Irvine)	
Machine Learning Repository	97
Generating Your Own Dataset	98
Linearly Distributed Dataset	98
Clustered Dataset	98
Clustered Dataset Distributed in Circular Fashion	100
Getting Started with Scikit-learn	100
Using the LinearRegression Class for Fitting the Model	101
Making Predictions	102
Plotting the Linear Regression Line	102
Getting the Gradient and Intercept of the Linear	
Regression Line	103
Examining the Performance of the Model by Calculating the	
Residual Sum of Squares	104
Evaluating the Model Using a Test Dataset	105
Persisting the Model	106
Data Cleansing	107
Cleaning Rows with NaNs	108
Replacing NaN with the Mean of the Column	109
Removing Rows	109
Removing Duplicate Rows	110
Normalizing Columns	112

	Removing Outliers	113
	Tukey Fences	113
	Z-Score	116
Chapter 6	Supervised Learning—Linear Regression	119
	Types of Linear Regression	119
	Linear Regression	120
	Using the Boston Dataset	120
	Data Cleansing	125
	Feature Selection	126
	Multiple Regression	128
	Training the Model	131
	Getting the Intercept and Coefficients	133
	Plotting the 3D Hyperplane	133
	Polynomial Regression	135
	Formula for Polynomial Regression	138
	Polynomial Regression in Scikit-learn	138
	Understanding Bias and Variance	141
	Using Polynomial Multiple Regression on the Boston Dataset	144
	Plotting the 3D Hyperplane	146
Chapter 7	Supervised Learning—Classification Using Logistic Regression	151
	What Is Logistic Regression?	151
	Understanding Odds	153
	Logit Function	153
	Sigmoid Curve	154
	Using the Breast Cancer Wisconsin (Diagnostic) Data Set	156
	Examining the Relationship Between Features	156
	Plotting the Features in 2D	157
	Plotting in 3D	158
	Training Using One Feature	161
	Finding the Intercept and Coefficient	162
	Plotting the Sigmoid Curve	162
	Making Predictions	163
	Training the Model Using All Features	164
	Testing the Model	166
	Getting the Confusion Matrix	166
	Computing Accuracy, Recall, Precision, and Other Metrics	168
	Receiver Operating Characteristic (ROC) Curve	171
	Plotting the ROC and Finding the Area Under the Curve (AUC)	174
Chapter 8	Supervised Learning—Classification Using Support Vector Machines	177
	What Is a Support Vector Machine?	177
	Maximum Separability	178
	Support Vectors	179

Formula for the Hyperplane	180
Using Scikit-learn for SVM	181
Plotting the Hyperplane and the Margins	184
Making Predictions	185
Kernel Trick	186
Adding a Third Dimension	187
Plotting the 3D Hyperplane	189
Types of Kernels	191
C	194
Radial Basis Function (RBF) Kernel	196
Gamma	197
Polynomial Kernel	199
Using SVM for Real-Life Problems	200
Chapter 9 Supervised Learning—Classification Using K-Nearest Neighbors (KNN)	205
What Is K-Nearest Neighbors?	205
Implementing KNN in Python	206
Plotting the Points	206
Calculating the Distance Between the Points	207
Implementing KNN	208
Making Predictions	209
Visualizing Different Values of K	209
Using Scikit-Learn's KNeighborsClassifier Class for KNN	211
Exploring Different Values of K	213
Cross-Validation	216
Parameter-Tuning K	217
Finding the Optimal K	218
Chapter 10 Unsupervised Learning—Clustering Using K-Means	221
What Is Unsupervised Learning?	221
Unsupervised Learning Using K-Means	222
How Clustering in K-Means Works	222
Implementing K-Means in Python	225
Using K-Means in Scikit-learn	230
Evaluating Cluster Size Using the Silhouette Coefficient	232
Calculating the Silhouette Coefficient	233
Finding the Optimal K	234
Using K-Means to Solve Real-Life Problems	236
Importing the Data	237
Cleaning the Data	237
Plotting the Scatter Plot	238
Clustering Using K-Means	239
Finding the Optimal Size Classes	240
Chapter 11 Using Azure Machine Learning Studio	243
What Is Microsoft Azure Machine Learning Studio?	243
An Example Using the Titanic Experiment	244
Using Microsoft Azure Machine Learning Studio	246

Uploading Your Dataset	247
Creating an Experiment	248
Filtering the Data and Making Fields Categorical	252
Removing the Missing Data	254
Splitting the Data for Training and Testing	254
Training a Model	256
Comparing Against Other Algorithms	258
Evaluating Machine Learning Algorithms	260
Publishing the Learning Model as a Web Service	261
Publishing the Experiment	261
Testing the Web Service	263
Programmatically Accessing the Web Service	263
Chapter 12 Deploying Machine Learning Models	269
Deploying ML	269
Case Study	270
Loading the Data	271
Cleaning the Data	271
Examining the Correlation Between the Features	273
Plotting the Correlation Between Features	274
Evaluating the Algorithms	277
Logistic Regression	277
K-Nearest Neighbors	277
Support Vector Machines	278
Selecting the Best Performing Algorithm	279
Training and Saving the Model	279
Deploying the Model	280
Testing the Model	282
Creating the Client Application to Use the Model	283
Index	285