Frank Kane 著

实用数据科学和 Python机器学习

（影印版）

Hands-On Data Science and Python Machine Learning

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

Packt

实用数据科学和 Python 机器学习（影印版）

Hands-On Data Science and Python Machine Learning

Frank Kane 著

南京　东南大学出版社

图书在版编目(CIP)数据

实用数据科学和 Python 机器学习：英文/（美）弗兰克·凯恩(Frank Kane)著. —影印本. —南京：东南大学出版社，2019.5

书名原文：Hands-On Data Science and Python Machine Learning

ISBN 978-7-5641-8320-2

Ⅰ.①实…　Ⅱ.①弗…　Ⅲ.①软件工具-程序设计-英文　Ⅳ.①TP311.561

中国版本图书馆 CIP 数据核字（2019）第 046197 号
图字：10-2018-503 号

© 2017 by PACKT Publishing Ltd.

Reprint of the English Edition, jointly published by PACKT Publishing Ltd and Southeast University Press, 2019. Authorized reprint of the original English edition, 2018 PACKT Publishing Ltd, the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 PACKT Publishing Ltd 出版 2018。

英文影印版由东南大学出版社出版 2019。此影印版的出版和销售得到出版权和销售权的所有者—— PACKT Publishing Ltd 的许可。

版权所有，未得书面许可，本书的任何部分和全部不得以任何形式重制。

实用数据科学和 Python 机器学习（影印版）

出版发行：东南大学出版社
地　　址：南京四牌楼 2 号　　邮编：210096
出 版 人：江建中
网　　址：http://www.seupress.com
电子邮件：press@seupress.com
印　　刷：常州市武进第三印刷有限公司
开　　本：787 毫米×980 毫米　16 开本
印　　张：26.25
字　　数：514 千字
版　　次：2019 年 5 月第 1 版
印　　次：2019 年 5 月第 1 次印刷
书　　号：ISBN 978-7-5641-8320-2
定　　价：99.00 元

本社图书若有印装质量问题，请直接与营销部联系。电话（传真）：025-83791830

Credits

Author
Frank Kane

Acquisition Editor
Ben Renow-Clarke

Content Development Editor
Khushali Bhangde

Technical Editor
Nidhisha Shetty

Copy Editor
Tom Jacob

Proofreader
Safis Editing

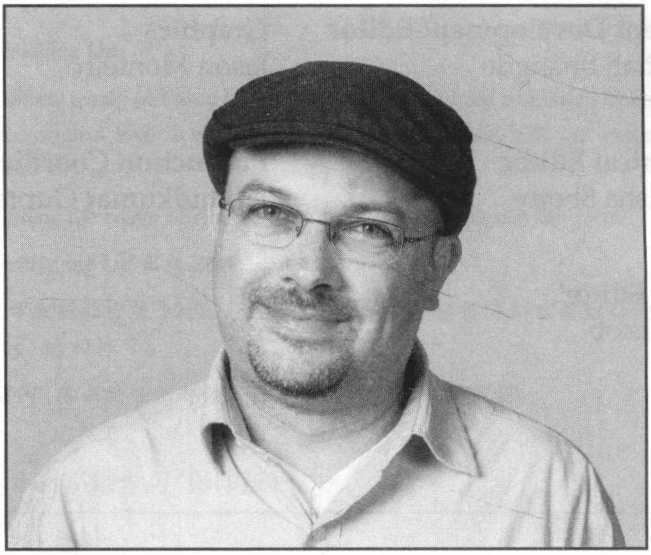
Indexer
Tejal Daruwale Soni

Graphics
Jason Monteiro

Production Coordinator
Arvindkumar Gupta

About the Author

My name is Frank Kane. I spent nine years at `amazon.com` and `imdb.com`, wrangling millions of customer ratings and customer transactions to produce things such as personalized recommendations for movies and products and "people who bought this also bought." I tell you, I wish we had Apache Spark back then, when I spent years trying to solve these problems there. I hold 17 issued patents in the fields of distributed computing, data mining, and machine learning. In 2012, I left to start my own successful company, Sundog Software, which focuses on virtual reality environment technology, and teaching others about big data analysis.



此为试读，需要完整PDF请访问：www.ertongbook.com

www.PacktPub.com

For support files and downloads related to your book, please visit `www.PacktPub.com`. Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com`and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details. At `www.PacktPub.com`, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

Mapt

`https://www.packtpub.com/mapt`

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at `https://www.amazon.com/dp/1787280748`.

If you'd like to join our team of regular reviewers, you can email us at `customerreviews@packtpub.com`. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

Table of Contents

Preface 1

Chapter 1: Getting Started 7

Installing Enthought Canopy 8

Giving the installation a test run 12

If you occasionally get problems opening your IPNYB files 15

Using and understanding IPython (Jupyter) Notebooks 15

Python basics - Part 1 20

Understanding Python code 22

Importing modules 25

Data structures 26

Experimenting with lists 26

Pre colon 27

Post colon 27

Negative syntax 28

Adding list to list 28

The append function 28

Complex data structures 29

Dereferencing a single element 29

The sort function 30

Reverse sort 30

Tuples 30

Dereferencing an element 31

List of tuples 31

Dictionaries 33

Iterating through entries 34

Python basics - Part 2 35

Functions in Python 35

Lambda functions - functional programming 36

Understanding boolean expressions 37

The if statement 37

The if-else loop 38

Looping 38

The while loop 39

Exploring activity 40

Running Python scripts 41

More options than just the IPython/Jupyter Notebook 42

Running Python scripts in command prompt 43

Using the Canopy IDE 44
Summary 47
Chapter 2: Statistics and Probability Refresher, and Python Practice 49
Types of data 50
Numerical data 51
Discrete data 51
Continuous data 51
Categorical data 52
Ordinal data 53
Mean, median, and mode 54
Mean 55
Median 55
The factor of outliers 56
Mode 57
Using mean, median, and mode in Python 58
Calculating mean using the NumPy package 58
Visualizing data using matplotlib 59
Calculating median using the NumPy package 60
Analyzing the effect of outliers 61
Calculating mode using the SciPy package 62
Some exercises 64
Standard deviation and variance 65
Variance 66
Measuring variance 67
Standard deviation 68
Identifying outliers with standard deviation 69
Population variance versus sample variance 70
The Mathematical explanation 70
Analyzing standard deviation and variance on a histogram 72
Using Python to compute standard deviation and variance 73
Try it yourself 74
Probability density function and probability mass function 74
The probability density function and probability mass functions 74
Probability density functions 75
Probability mass functions 76
Types of data distributions 77
Uniform distribution 77
Normal or Gaussian distribution 78
The exponential probability distribution or Power law 80
Binomial probability mass function 82
Poisson probability mass function 83

Percentiles and moments 84
Percentiles 84
Quartiles 86
Computing percentiles in Python 87
Moments 90
Computing moments in Python 93
Summary 95
Chapter 3: Matplotlib and Advanced Probability Concepts 97
A crash course in Matplotlib 98
Generating multiple plots on one graph 99
Saving graphs as images 100
Adjusting the axes 101
Adding a grid 102
Changing line types and colors 103
Labeling axes and adding a legend 107
A fun example 108
Generating pie charts 110
Generating bar charts 111
Generating scatter plots 112
Generating histograms 113
Generating box-and-whisker plots 113
Try it yourself 115
Covariance and correlation 116
Defining the concepts 116
Measuring covariance 117
Correlation 118
Computing covariance and correlation in Python 118
Computing correlation – The hard way 118
Computing correlation – The NumPy way 122
Correlation activity 124
Conditional probability 124
Conditional probability exercises in Python 125
Conditional probability assignment 129
My assignment solution 130
Bayes' theorem 132
Summary 134
Chapter 4: Predictive Models 135
Linear regression 135
The ordinary least squares technique 137

    - The gradient descent technique 138
    - The co-efficient of determination or r-squared 139
        - Computing r-squared 139
        - Interpreting r-squared 139
    - Computing linear regression and r-squared using Python 140
    - Activity for linear regression 143
- **Polynomial regression** 144
    - Implementing polynomial regression using NumPy 145
        - Computing the r-squared error 148
    - Activity for polynomial regression 148
- **Multivariate regression and predicting car prices** 149
    - Multivariate regression using Python 151
    - Activity for multivariate regression 154
- **Multi-level models** 155
- **Summary** 157

Chapter 5: Machine Learning with Python 159

- **Machine learning and train/test** 159
    - Unsupervised learning 160
    - Supervised learning 161
        - Evaluating supervised learning 162
        - K-fold cross validation 164
- **Using train/test to prevent overfitting of a polynomial regression** 164
    - Activity 170
- **Bayesian methods - Concepts** 170
- **Implementing a spam classifier with Naïve Bayes** 172
    - Activity 176
- **K-Means clustering** 177
    - Limitations to k-means clustering 179
- **Clustering people based on income and age** 181
    - Activity 184
- **Measuring entropy** 184
- **Decision trees - Concepts** 186
    - Decision tree example 188
    - Walking through a decision tree 190
    - Random forests technique 190
- **Decision trees - Predicting hiring decisions using Python** 191
    - Ensemble learning – Using a random forest 197
    - Activity 198
- **Ensemble learning** 198
- **Support vector machine overview** 201

此为试读，需要完整PDF请访问：www.ertongbook.com

Using SVM to cluster people by using scikit-learn 203
Activity 207
Summary 208

Chapter 6: Recommender Systems 209

What are recommender systems? 210
User-based collaborative filtering 212
Limitations of user-based collaborative filtering 214
Item-based collaborative filtering 215
Understanding item-based collaborative filtering 215
How item-based collaborative filtering works? 216
Collaborative filtering using Python 220
Finding movie similarities 220
Understanding the code 222
The corrwith function 225
Improving the results of movie similarities 228
Making movie recommendations to people 233
Understanding movie recommendations with an example 238
Using the groupby command to combine rows 240
Removing entries with the drop command 241
Improving the recommendation results 242
Summary 244

Chapter 7: More Data Mining and Machine Learning Techniques 245

K-nearest neighbors - concepts 246
Using KNN to predict a rating for a movie 248
Activity 255
Dimensionality reduction and principal component analysis 256
Dimensionality reduction 256
Principal component analysis 257
A PCA example with the Iris dataset 259
Activity 264
Data warehousing overview 264
ETL versus ELT 266
Reinforcement learning 268
Q-learning 269
The exploration problem 270
The simple approach 270
The better way 271
Fancy words 271
Markov decision process 271
Dynamic programming 272

Summary 274

Chapter 8: Dealing with Real-World Data 275

Bias/variance trade-off 276

K-fold cross-validation to avoid overfitting 279

Example of k-fold cross-validation using scikit-learn 280

Data cleaning and normalisation 284

Cleaning web log data 287

Applying a regular expression on the web log 288

Modification one - filtering the request field 291

Modification two - filtering post requests 293

Modification three - checking the user agents 295

Filtering the activity of spiders/robots 297

Modification four - applying website-specific filters 299

Activity for web log data 301

Normalizing numerical data 301

Detecting outliers 303

Dealing with outliers 304

Activity for outliers 307

Summary 307

Chapter 9: Apache Spark - Machine Learning on Big Data 309

Installing Spark 310

Installing Spark on Windows 310

Installing Spark on other operating systems 311

Installing the Java Development Kit 312

Installing Spark 319

Spark introduction 333

It's scalable 334

It's fast 335

It's young 336

It's not difficult 336

Components of Spark 336

Python versus Scala for Spark 337

Spark and Resilient Distributed Datasets (RDD) 338

The SparkContext object 339

Creating RDDs 340

Creating an RDD using a Python list 340

Loading an RDD from a text file 340

More ways to create RDDs 341

RDD operations 341

Transformations 342
Using map() 343
Actions 343
Introducing MLlib 344
Some MLlib Capabilities 345
Special MLlib data types 345
The vector data type 346
LabeledPoint data type 346
Rating data type 346
Decision Trees in Spark with MLlib 347
Exploring decision trees code 348
Creating the SparkContext 349
Importing and cleaning our data 351
Creating a test candidate and building our decision tree 356
Running the script 357
K-Means Clustering in Spark 359
Within set sum of squared errors (WSSSE) 363
Running the code 364
TF-IDF 365
TF-IDF in practice 366
Using TF- IDF 367
Searching wikipedia with Spark MLlib 367
Import statements 369
Creating the initial RDD 369
Creating and transforming a HashingTF object 370
Computing the TF-IDF score 371
Using the Wikipedia search engine algorithm 371
Running the algorithm 372
Using the Spark 2.0 DataFrame API for MLlib 373
How Spark 2.0 MLlib works 373
Implementing linear regression 374
Summary 378
Chapter 10: Testing and Experimental Design 379
A/B testing concepts 379
A/B tests 379
Measuring conversion for A/B testing 382
How to attribute conversions 383
Variance is your enemy 383
T-test and p-value 384
The t-statistic or t-test 385
The p-value 385

Measuring t-statistics and p-values using Python 387
Running A/B test on some experimental data 387
When there's no real difference between the two groups 388
Does the sample size make a difference? 389
Sample size increased to six-digits 389
Sample size increased seven-digits 390
A/A testing 390
Determining how long to run an experiment for 391
A/B test gotchas 392
Novelty effects 394
Seasonal effects 394
Selection bias 395
Auditing selection bias issues 396
Data pollution 396
Attribution errors 397
Summary 397
Index 399

Preface

Being a data scientist in the tech industry is one of the most rewarding careers on the planet today. I went and studied actual job descriptions for data scientist roles at tech companies and I distilled those requirements down into the topics that you'll see in this course.

Hands-On Data Science and Python Machine Learning is really comprehensive. We'll start with a crash course on Python and do a review of some basic statistics and probability, but then we're going to dive right into over 60 topics in data mining and machine learning. That includes things such as Bayes' theorem, clustering, decision trees, regression analysis, experimental design; we'll look at them all. Some of these topics are really fun.

We're going to develop an actual movie recommendation system using actual user movie rating data. We're going to create a search engine that actually works for Wikipedia data. We're going to build a spam classifier that can correctly classify spam and nonspam emails in your email account, and we also have a whole section on scaling this work up to a cluster that runs on big data using Apache Spark.

If you're a software developer or programmer looking to transition into a career in data science, this course will teach you the hottest skills without all the mathematical notation and pretense that comes along with these topics. We're just going to explain these concepts and show you some Python code that actually works that you can dive in and mess around with to make those concepts sink home, and if you're working as a data analyst in the finance industry, this course can also teach you to make the transition into the tech industry. All you need is some prior experience in programming or scripting and you should be good to go.

The general format of this book is I'll start with each concept, explaining it in a bunch of sections and graphical examples. I will introduce you to some of the notations and fancy terminologies that data scientists like to use so you can talk the same language, but the concepts themselves are generally pretty simple. After that, I'll throw you into some actual Python code that actually works that we can run and mess around with, and that will show you how to actually apply these ideas to actual data. These are going to be presented as IPython Notebook files, and that's a format where I can intermix code and notes surrounding the code that explain what's going on in the concepts. You can take these notebook files with you after going through this book and use that as a handy-quick reference later on in your career, and at the end of each concept, I'll encourage you to actually dive into that Python code, make some modifications, mess around with it, and just gain more familiarity by getting hands-on and actually making some modifications, and seeing the effects they have.

Who this book is for

If you are a budding data scientist or a data analyst who wants to analyze and gain actionable insights from data using Python, this book is for you. Programmers with some experience in Python who want to enter the lucrative world of Data Science will also find this book to be very useful.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "We can measure that using the `r2_score()` function from `sklearn.metrics`."

A block of code is set as follows:

```
import numpy as np
import pandas as pd
from sklearn import tree

input_file = "c:/spark/DataScience/PastHires.csv"
df = pd.read_csv(input_file, header = 0)
```

When we wish to draw your attention to a particular part of a code block, the relevant lines or items are set in bold:

```
import numpy as np
import pandas as pd
from sklearn import tree

input_file = "c:/spark/DataScience/PastHires.csv"
df = pd.read_csv(input_file, header = 0)
```

Any command-line input or output is written as follows:

```
spark-submit SparkKMeans.py
```

此为试读，需要完整PDF请访问：www.ertongbook.com