TURING 图灵程序设计丛书

Packt

Hands-On Data Science and Python Machine Learning

Python 数据科学与机器学习 从入门到实践

[美] 弗兰克·凯恩◎著　　陈光欣◎译

寓复杂问题于简单实践，轻松掌握Python数据分析和机器学习技能

中国工信出版集团　人民邮电出版社 POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

Hands-On Data Science and Python Machine Learning

Python 数据科学与机器学习 从入门到实践

[美] 弗兰克·凯恩◎著　　陈光欣◎译

人民邮电出版社
北　京

图书在版编目（CIP）数据

Python数据科学与机器学习 : 从入门到实践 / (美)弗兰克·凯恩 (Frank Kane) 著 ; 陈光欣译. -- 北京 : 人民邮电出版社, 2019.6
(图灵程序设计丛书)
ISBN 978-7-115-51241-3

Ⅰ. ①P… Ⅱ. ①弗… ②陈… Ⅲ. ①软件工具－程序设计②机器学习 Ⅳ. ①TP311.561②TP181

中国版本图书馆CIP数据核字(2019)第087642号

内 容 提 要

本书介绍了使用 Python 进行数据分析和高效的机器学习，首先从一节 Python 速成课开始，然后回顾统计学和概率论的基础知识，接着深入讨论与数据挖掘和机器学习相关的 60 多个主题，包括贝叶斯定理、聚类、决策树、回归分析、实验设计等。

本书适合有一定 Python 编程基础，想要了解数据分析和机器学习的读者阅读。

◆ 著　[美] 弗兰克·凯恩
译　陈光欣
责任编辑　张海艳
责任印制　周昇亮

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路11号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：17.75
字数：420千字　2019年6月第1版
印数：1－3 500册　2019年6月北京第1次印刷
著作权合同登记号　图字：01-2017-7476号

定价：69.00元

读者服务热线：(010)51095183转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

前　言

当今世界，高科技行业的数据科学家是回报最为丰厚的职业之一。我仔细地研究了高科技公司中数据科学家的职位描述，并根据其职位要求确定了本书内容。

本书内容非常全面，从一节 Python 速成课开始，然后回顾统计学和概率论的基础知识，接着深入讨论与数据挖掘和机器学习相关的 60 多个主题。这些主题包括贝叶斯定理、聚类、决策树、回归分析、实验设计等，其中有些内容非常有意思。

我们将使用真实的电影评分数据创建一个电影推荐系统，还会创建一个能实际运行的维基百科数据搜索引擎。此外，我们还将构建一个垃圾邮件分类器，它可以对邮件账户中的垃圾邮件和正常邮件进行正确的分类。本书还会专门用一章介绍如何将这个分类器扩展到使用 Apache Spark 的大数据集群系统上。

如果你是希望转型为数据科学家的软件开发人员或程序员，那么从本书中可以学到当下最热门的技能，掌握这些技能并不需要你具有高深的数学背景，当然你也就不能以数学不好作为托词了。我们会解释相关的概念，并提供一些确实有效的 Python 代码。你可以仔细研究这些代码，也可以任意修改，以此来领悟书中概念。如果你是金融行业的数据分析师，本书也可以帮你转型进入高科技行业。只需一些程序开发和脚本编写经验，你就可以开始学习本书了。

本书通常会先介绍概念，然后用几节文字和图形化示例对其进行解释。我会介绍数据科学家喜欢使用的一些表示方法和时髦术语，让你和他们有共同语言，这些概念本身是非常简单易懂的。之后，我会提供一些可以实际运行的 Python 代码供你执行和修改，这样你就可以知道如何将所学的概念应用到真实的数据上了。这些代码是以 IPython Notebook 文件的形式提供的，这种文件可以将代码和注释集成在一起，其中注释可以对代码做出解释。学完本书之后，你可以将这些文件作为工作中的快速参考。介绍完每个概念之后，我都会鼓励你仔细研究一下 Python 代码，并随意进行修改。通过这些实际的练习和修改，你会对代码更熟悉，并且更加清楚它们的作用。

目标读者

如果你是崭露头角的数据科学家，或是想使用 Python 对数据进行分析并获取实用知识的数据分析师，那么本书正是你需要的。对于那些具有 Python 编程经验并想进入数据科学领域淘金

的程序员来说，本书也会使他们受益匪浅。

排版约定

本书使用了不同的文本样式来区分不同种类的信息。下面给出了几种样式示例，并对其含义进行了解释。

文本中的代码、数据库表名、虚拟 URL 和用户输入都表示为："我们可以使用 `sklearn.metrics` 中的 `r2_score()` 函数来对其进行测量。"

以下是一段代码：

```
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)
```

如果我们想让你特别注意代码段中的某个部分，就会将相关的行或项目用粗体表示：

```
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)
```

所有命令行输入和输出都如下所示：

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

新名词和重点词会以黑体显示。显示器屏幕（比如菜单或对话框）上的词在文本中表示为："在 Windows 10 系统中，你需要打开 Start 菜单，然后使用 Windows System | Control Panel 来打开 Control Panel。"

警告或重要的注意事项。

提示或小技巧。

读者反馈

十分欢迎读者提供反馈意见。请让我们知道你对本书的看法：喜欢哪些部分，不喜欢哪些部

分。读者反馈对我们非常重要，因为这可以帮助我们编写出真正对大家有所裨益的图书。

要想提供反馈，只需发送邮件到 feedback@packtpub.com，并在邮件标题中注明书名即可。

如果你有擅长的领域，想参与图书编写或出版，请参阅我们的作者指南：www.packtpub.com/authors。

客户支持

现在你已经拥有了这本由 Packt 出版的图书，我们将提供一系列服务来使你获得最大收益。

下载示例代码

你可以到图灵社区本书页面下载代码文件，网址是 http://ituring.cn/book/2426。

文件下载结束之后，请确定使用以下软件的最新版本来解压或提取文件。

- Windows 系统：使用 WinRAR 或 7-Zip
- Mac 系统：使用 Zipeg、iZip 或 UnRarX
- Linux 系统：使用 7-Zip 或 PeaZip

下载本书彩色图片

我们还提供了一个 PDF 文件，里面有本书所使用的屏幕截图和图表的彩色图片。彩色图片可以帮助你更好地理解输出的变化。你同样可以从图灵社区本书页面下载，网址是：http://ituring.cn/book/2426。

勘误

尽管我们做了各种努力来保证内容的准确性，但错误终难避免。如果你在我们的任何一本书中发现了错误，不论是正文中的还是代码中的，都请提交给我们，我们将非常感谢。通过提交勘误，不仅可以提升其他读者的阅读体验，还能帮助我们在本书的后续版本中做出改进。不管你发现了什么错误，都可以通过 http://www.packtpub.com/submit-errata 这个网址告诉我们。首先选择相应的图书，然后点击 Errata Submission Form 这个链接，输入勘误的具体细节即可。[①]一旦勘误校验通过，你提供的信息将被接受，勘误信息将上传到我们的网站，或者添加到相应图书的勘误表中。

① 本书中文版勘误请到 http://ituring.cn/book/2426 查看和提交。——编者注

要想查看以前提交的勘误信息，请访问 https://www.packtpub.com/books/content/support，在搜索框中输入书名，相应信息就会出现在 Errata 部分。

举报盗版

所有正版内容在互联网上都面临的一个问题就是侵权。Packt 严格保护版权和授权。如果你在网上发现我社图书的任何形式的盗版，请立即将地址或网站名称提供给我们，以便我们采取进一步的措施。

请将疑似侵权的网站链接发送至 copyright@packtpub.com。

非常感谢你对保护作者知识产权所做的工作，我们将竭诚为读者提供有价值的内容。

问题

对本书有任何疑问，都可以联系 question@packtpub.com，我们会尽最大努力来解决问题。

电子书

扫描如下二维码，即可购买本书电子版。

目　录

第 1 章　入门 ····· 1

1.1　安装 Enthought Canopy ····· 1
1.2　使用并理解 IPython/Jupyter Notebook ····· 6
1.3　Python 基础——第一部分 ····· 9
1.4　理解 Python 代码 ····· 11
1.5　导入模块 ····· 13
1.5.1　数据结构 ····· 13
1.5.2　使用列表 ····· 14
1.5.3　元组 ····· 17
1.5.4　字典 ····· 18
1.6　Python 基础——第二部分 ····· 20
1.6.1　Python 中的函数 ····· 20
1.6.2　循环 ····· 23
1.6.3　探索活动 ····· 24
1.7　运行 Python 脚本 ····· 24
1.7.1　运行 Python 代码的其他方式 ····· 25
1.7.2　在命令行中运行 Python 脚本 ····· 25
1.7.3　使用 Canopy IDE ····· 26
1.8　小结 ····· 28

第 2 章　统计与概率复习以及 Python 实现 ····· 29

2.1　数据类型 ····· 29
2.1.1　数值型数据 ····· 30
2.1.2　分类数据 ····· 30
2.1.3　定序数据 ····· 31
2.2　均值、中位数和众数 ····· 32
2.2.1　均值 ····· 32
2.2.2　中位数 ····· 33
2.2.3　众数 ····· 34
2.3　在 Python 中使用均值、中位数和众数 ····· 35
2.3.1　使用 NumPy 包计算均值 ····· 35
2.3.2　使用 NumPy 包计算中位数 ····· 36
2.3.3　使用 SciPy 包计算众数 ····· 37
2.4　标准差和方差 ····· 40
2.4.1　方差 ····· 40
2.4.2　标准差 ····· 42
2.4.3　总体方差与样本方差 ····· 42
2.4.4　在直方图上分析标准差和方差 ····· 44
2.4.5　使用 Python 计算标准差和方差 ····· 44
2.4.6　自己动手 ····· 45
2.5　概率密度函数和概率质量函数 ····· 45
2.5.1　概率密度函数 ····· 45
2.5.2　概率质量函数 ····· 46
2.6　各种类型的数据分布 ····· 47
2.6.1　均匀分布 ····· 47
2.6.2　正态分布或高斯分布 ····· 48
2.6.3　指数概率分布与指数定律 ····· 50
2.6.4　二项式概率质量函数 ····· 50
2.6.5　泊松概率质量函数 ····· 51
2.7　百分位数和矩 ····· 52
2.7.1　百分位数 ····· 53
2.7.2　矩 ····· 56
2.8　小结 ····· 60

第 3 章 Matplotlib 与概率高级概念……61
3.1 Matplotlib 快速学习……61
3.1.1 在一张图形上进行多次绘图……62
3.1.2 将图形保存为文件……63
3.1.3 调整坐标轴……64
3.1.4 添加网格……65
3.1.5 修改线型和颜色……65
3.1.6 标记坐标轴并添加图例……68
3.1.7 一个有趣的例子……69
3.1.8 生成饼图……70
3.1.9 生成条形图……71
3.1.10 生成散点图……72
3.1.11 生成直方图……72
3.1.12 生成箱线图……73
3.1.13 自己动手……74
3.2 协方差与相关系数……74
3.2.1 概念定义……75
3.2.2 相关系数……76
3.2.3 在 Python 中计算协方差和相关系数……76
3.2.4 相关系数练习……80
3.3 条件概率……80
3.3.1 Python 中的条件概率练习……81
3.3.2 条件概率作业……84
3.3.3 作业答案……85
3.4 贝叶斯定理……86
3.5 小结……88
第 4 章 预测模型……89
4.1 线性回归……89
4.1.1 普通最小二乘法……90
4.1.2 梯度下降法……91
4.1.3 判定系数或 r 方……91
4.1.4 使用 Python 进行线性回归并计算 r 方……92
4.1.5 线性回归练习……94
4.2 多项式回归……95
4.2.1 使用 NumPy 实现多项式回归……96
4.2.2 计算 r 方误差……98
4.2.3 多项式回归练习……98
4.3 多元回归和汽车价格预测……99
4.3.1 使用 Python 进行多元回归……100
4.3.2 多元回归练习……102
4.4 多水平模型……102
4.5 小结……104
第 5 章 使用 Python 进行机器学习……105
5.1 机器学习及训练/测试法……105
5.1.1 非监督式学习……106
5.1.2 监督式学习……107
5.2 使用训练/测试法防止多项式回归中的过拟合……109
5.3 贝叶斯方法——概念……113
5.4 使用朴素贝叶斯实现垃圾邮件分类器……115
5.5 k 均值聚类……118
5.6 基于收入与年龄进行人群聚类……121
5.7 熵的度量……123
5.8 决策树——概念……124
5.8.1 决策树实例……126
5.8.2 生成决策树……127
5.8.3 随机森林……127
5.9 决策树——使用 Python 预测录用决策……128
5.9.1 集成学习——使用随机森林……132
5.9.2 练习……133
5.10 集成学习……133
5.11 支持向量机简介……135
5.12 使用 scikit-learn 通过 SVM 进行人员聚集……137
5.13 小结……140

第 6 章　推荐系统……141
6.1　什么是推荐系统……141
6.2　基于项目的协同过滤……145
6.3　基于项目的协同过滤是如何工作的……146
6.4　找出电影相似度……149
6.5　改善电影相似度结果……155
6.6　向人们推荐电影……159
6.7　改善推荐结果……165
6.8　小结……167
第 7 章　更多数据挖掘和机器学习技术……168
7.1　k 最近邻的概念……168
7.2　使用 KNN 预测电影评分……170
7.3　数据降维与主成分分析……176
7.3.1　数据降维……176
7.3.2　主成分分析……177
7.4　对鸢尾花数据集的 PCA 示例……178
7.5　数据仓库简介……182
7.6　强化学习……184
7.6.1　Q-learning……185
7.6.2　探索问题……186
7.6.3　时髦名词……186
7.7　小结……188
第 8 章　处理真实数据……189
8.1　偏差-方差权衡……189
8.2　使用 k 折交叉验证避免过拟合……192
8.3　数据清理和标准化……196
8.4　清理 Web 日志数据……198
8.4.1　对 Web 日志应用正则表达式……198
8.4.2　修改 1——筛选请求字段……200
8.4.3　修改 2——筛选 post 请求……201
8.4.4　修改 3——检查用户代理……203
8.4.5　筛选爬虫与机器人……204
8.4.6　修改 4——使用网站专用筛选器……205
8.4.7　Web 日志数据练习……206
8.5　数值型数据的标准化……207
8.6　检测异常值……208
8.6.1　处理异常值……209
8.6.2　异常值练习……211
8.7　小结……211
第 9 章　Apache Spark——大数据上的机器学习……212
9.1　安装 Spark……212
9.1.1　在 Windows 系统中安装 Spark……213
9.1.2　在其他操作系统上安装 Spark……214
9.1.3　安装 Java Development Kit……214
9.1.4　安装 Spark……217
9.2　Spark 简介……227
9.2.1　可伸缩……227
9.2.2　速度快……228
9.2.3　充满活力……229
9.2.4　易于使用……229
9.2.5　Spark 组件……229
9.2.6　在 Spark 中使用 Python 还是 Scala……230
9.3　Spark 和弹性分布式数据集……231
9.3.1　SparkContext 对象……231
9.3.2　创建 RDD……232
9.3.3　更多创建 RDD 的方法……233
9.3.4　RDD 操作……233
9.4　MLlib 简介……235
9.4.1　MLlib 功能……235
9.4.2　MLlib 特殊数据类型……236
9.5　在 Spark 中使用 MLlib 实现决策树……236
9.6　在 Spark 中实现 k 均值聚类……245
9.7　TF-IDF……250
9.7.1　TF-IDF 实战……250
9.7.2　使用 TF-IDF……251
9.8　使用 Spark MLlib 搜索维基百科……251
9.8.1　导入语句……252
9.8.2　创建初始 RDD……252

9.8.3 创建并转换 HashingTF 对象 …253
9.8.4 计算 TF-IDF 得分 ……254
9.8.5 使用维基百科搜索引擎算法 ……254
9.8.6 运行算法 ……255
9.9 使用 Spark 2.0 中的 MLlib 数据框 API ……255
9.10 小结 ……259
第 10 章 测试与实验设计 ……260
10.1 A/B 测试的概念 ……260
10.1.1 A/B 测试 ……260
10.1.2 A/B 测试的转化效果测量 ……262
10.1.3 小心方差 ……263
10.2 *t* 检验与 *p* 值 ……263
10.2.1 *t* 统计量或 *t* 检验 ……264
10.2.2 *p* 值 ……264
10.3 使用 Python 计算 *t* 统计量和 *p* 值 ……265
10.3.1 使用实验数据进行 A/B 测试 ……265
10.3.2 样本量有关系吗 ……267
10.4 确定实验持续时间 ……268
10.5 A/B 测试中的陷阱 ……269
10.5.1 新奇性效应 ……270
10.5.2 季节性效应 ……271
10.5.3 选择性偏差 ……271
10.5.4 数据污染 ……272
10.5.5 归因错误 ……272
10.6 小结 ……273

第 1 章 入 门

因为本书中有很多代码和样本数据，所以我首先要告诉你如何获取它们，然后再进行下一步。我们需要做一些准备工作。当务之急就是获取学习本书所需的代码和数据，这样你就可以“愉快地开始玩耍了”。获取代码和数据最简单的方法就是按照本章的指示去做。

在本章中，我们首先要安装并准备好 Python 工作环境：

- 安装 Enthought Canopy；
- 安装 Python 库文件；
- 使用 IPython/Jupyter Notebook；
- 使用、读取和运行本书中的代码文件。

然后，我们将通过一节速成课来学习如何理解 Python 代码：

- Python 基础——第一部分；
- 理解 Python 代码；
- 导入模块；
- 使用列表；
- 元组；
- Python 基础——第二部分；
- 运行 Python 脚本。

通过这一章的学习，你可以搭建 Python 工作环境，熟悉 Python 语言，然后就可以使用 Python 开始数据科学的奇妙之旅了。

1.1 安装 Enthought Canopy

本节介绍如何在你的计算机上安装 Python 环境，以开发数据科学所需的代码。我们将安装一个称为 Enthought Canopy 的软件包，其中既包括开发环境，也包括你需要预先安装的所有 Python 包，它可以使你的工作更加容易。但是，如果你之前使用过 Python，那么你的计算机上可能已经

存在 Python 环境了，如果你想继续使用这个环境，或许也可以。

最重要的是，你的 Python 环境应该是能够支持 Jupyter Notebook 的 Python 3.5 或更高版本（因为本书使用的就是 Jupyter Notebook），而且环境中安装了本书需要的关键软件包。我会通过几个简单的步骤，精确地介绍如何进行完整的安装，这种安装是非常简单的。

首先来看一下关键的软件包，Canopy 会自动为我们安装其中的大部分。它将会安装 Python 3.5 以及我们需要的一些软件包：`scikit_learn`、`xlrd` 和 `statsmodels`。我们还需要使用 `pip` 命令，手动安装一个名为 `pydot2plus` 的包。这就行了，使用 Canopy 就是这么简单！

如果完成了以下安装步骤，那么所需的准备工作就全部就绪，可以打开一个小示例文件，进行真正的数据科学工作了。下面，让我们尽快完成所需的准备工作。

(1) 首先，我们需要一个 Python 集成开发环境，又称 IDE。本书中将使用 Enthought Canopy。这是一种科学计算环境，非常适合本书。

(2) 若要安装 Canopy，请访问 www.enthought.com，点击 DOWNLOADS:Canopy。

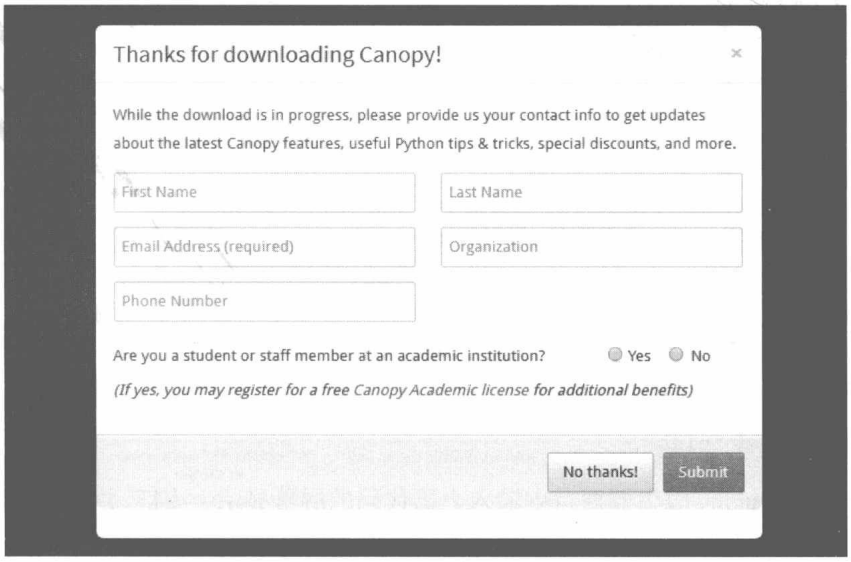
(3) Enthought Canopy 的 Canopy Express 版本是免费的，本书使用的就是这个版本。你必须选择适合自己操作系统的版本，对我来说是 Windows 64 位版。你应该点击对应你的操作系统和 Python 3.5 版的下载按钮。

Standard Installers

v2.1.3 v1.7.4 Documentation

Platform	Python		Released	Size	MD5
Linux [64-bit]	2.7	download	2017-06-16	697.8 MB	57b828e913e15a6ec12f1eb964138c82
Linux [64-bit]	3.5	download	2017-06-16	574.8 MB	7412235d9f72acc603df79bfbe706bee
macOS [64-bit]	2.7	download	2017-06-16	572.1 MB	d0ee780d2e7541e0c11a84ec9f29cbb2
macOS [64-bit]	3.5	download	2017-06-16	464.0 MB	d8c15b4763d8c55202c5dba9dd7f3157
Windows [64-bit]	2.7	download	2017-06-16	513.8 MB	3821c0a63abfe8d13d464ecda58d627c
Windows [32-bit]	2.7	download	2017-06-16	420.9 MB	895bff89399d5f4b59ef101dcb33edfd
Windows [64-bit]	3.5	download	2017-06-16	431.3 MB	82c62c8549a9b02a4fe751484e13bb48
Windows [32-bit]	3.5	download	2017-06-16	350.2 MB	f378349261eeb9d8bc614321d12d0264

(4) 这一步可以不填写任何个人信息。这是一个标准的 Windows 安装程序，开始下载吧。

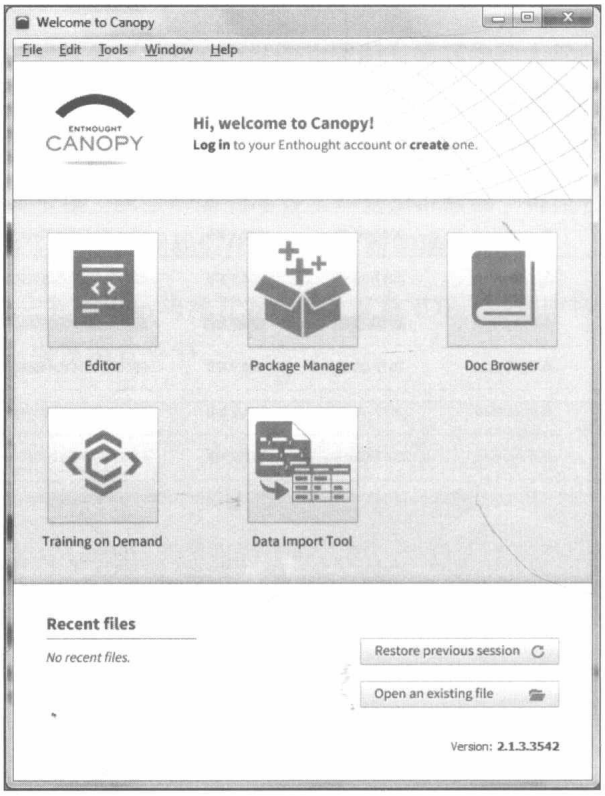


(5) 下载完成之后，找到 Canopy 安装程序并启动，开始安装。在同意许可条款之前，你或许想仔细地读一下，这由你自己决定，然后只需静待安装完成即可。

(6) 安装过程结束后，点击 Finish 按钮，可以自动启动 Canopy。然后你会看到 Canopy 可以

自己搭建 Python 环境，这个功能非常好，不过要持续 1~2 分钟。

(7) 安装程序搭建好 Python 环境之后，你会看到类似下图的屏幕显示。这是一个 Canopy 欢迎页面，含有一些用户友好的大按钮。

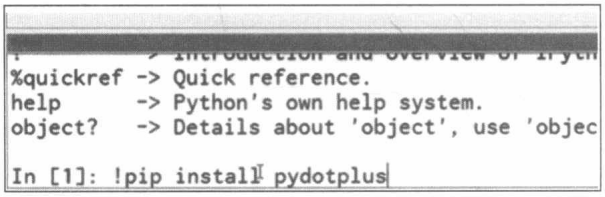


(8) 最棒的是，本书所需的几乎所有软件包都和 Enthought Canopy 一起预先安装好了，这就是我推荐 Enthought Canopy 的原因。

(9) 我们的准备工作还有最后一项。点击 Canopy 欢迎页面的 Editor 按钮，你会看到跳出一个编辑器界面，点击编辑器界面下方的窗口，输入以下代码。

```
!pip install pydotplus
```

(10) 下图是在 Canopy 编辑器窗口中输入上面代码的屏幕显示。别忘了按回车键。



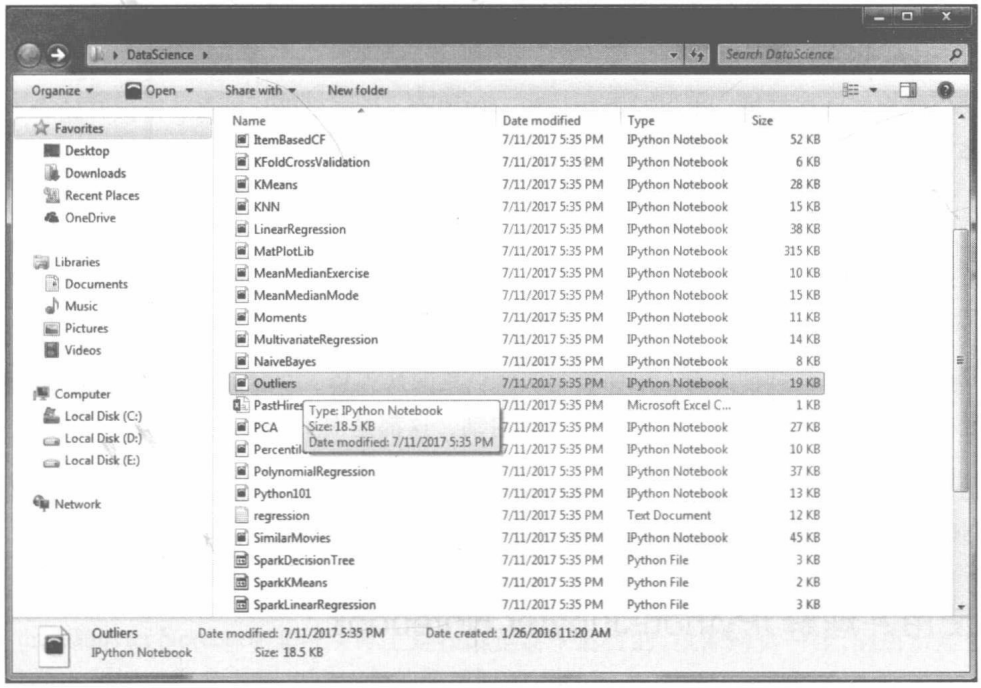
(11) 按下回车键后，就会开始安装一个附加模块，在本书后面讲到决策树和生成决策树时，会用到这个模块。

(12) pydotplus 安装完成后，程序会通知你。恭喜，你已经完成了所有的准备工作！程序安装完成之后，让我们再通过几个步骤确认程序可以顺畅地运行。

程序运行测试

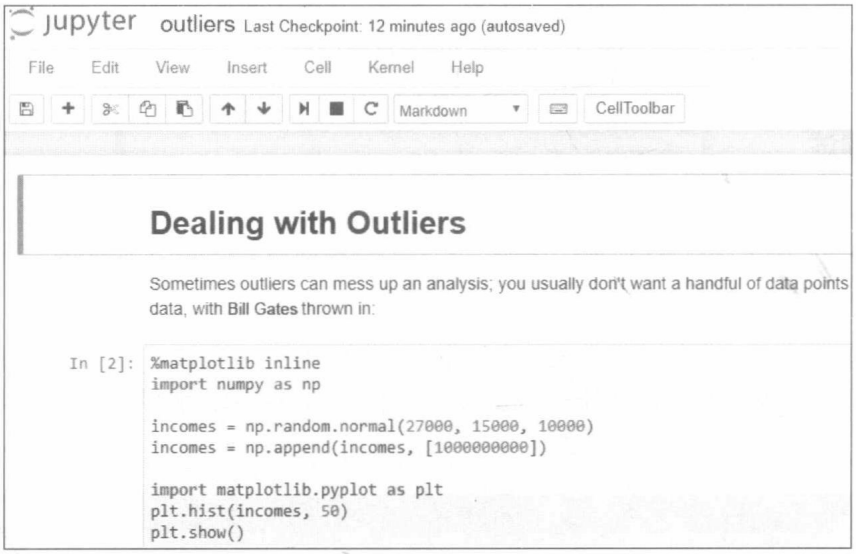
(1) 对安装好的程序进行一下测试。首先，将 Canopy 窗口全部关掉！这是因为我们实际上不会使用 Canopy 编辑器来编写代码，相反，我们要使用的是 IPython Notebook，现在它又称为 Jupyter Notebook。

(2) 让我来告诉你该怎么做。如果你在操作系统中打开了一个窗口，来查看本书附带的文件（文件的下载方法见前言），那么就应该看到下图这个样子，窗口中有一组本书将要用到的.ipynb 代码文件。



现在在列表中找到Outliers文件，也就是Outliers.ipynb文件。双击这个文件，将先启动Canopy，然后会启动你的 Web 浏览器。这是因为 IPython/Jupyter Notebook 实际上是在 Web 浏览器中运行的。开始时，会有一点小小的停顿，这在第一次使用时可能会有点困惑，但你很快就会习惯这种方式。

你很快就能看到，Canopy 和默认 Web 浏览器（我的是 Chrome）相继启动。因为双击了 Outliers.ipynb 文件，所以会看到以下 Jupyter Notebook 页面。



如果你看到了这个页面，就说明安装过程一切正常，你已经做好准备，可以开始学习本书后续的内容了！

如果在打开 IPYNB 文件时出现问题

我发现在双击.ipynb 文件时偶尔会出现一点小问题。别紧张！Canopy 只是偶尔会有一点古怪，有时会要求你输入密码或令牌，有时会显示无法连接。

当出现以上任意一种情况时，不要紧张，这只是偶然现象。有时候，仅仅是因为电脑上的某些程序没有按照正确顺序启动，或者没有及时启动，这都是很正常的。

你能做的就是再次打开文件，有时候需要试两次或三次才能正确加载它。你只要多试几次，Jupyter Notebook 页面总归会出现的，就像前面的 Dealing with Outliers 页面一样。

1.2　使用并理解 IPython/Jupyter Notebook

恭喜你安装完成！下面就开始使用 Jupyter Notebook，也就是 IPython Notebook。眼下，时髦的名称是 Jupyter Noterbook，但很多人还是称其为 IPython Notebook，其实很多开发者是交替使用这两个名称的。IPython Notebook 可以帮助我记住 notebook 文件的后缀.ipynb，本书中会频繁使用这种文件。

下面开始学习如何使用 IPython/Jupyter Notebook。找到本书下载资料中的 DataScience 文件