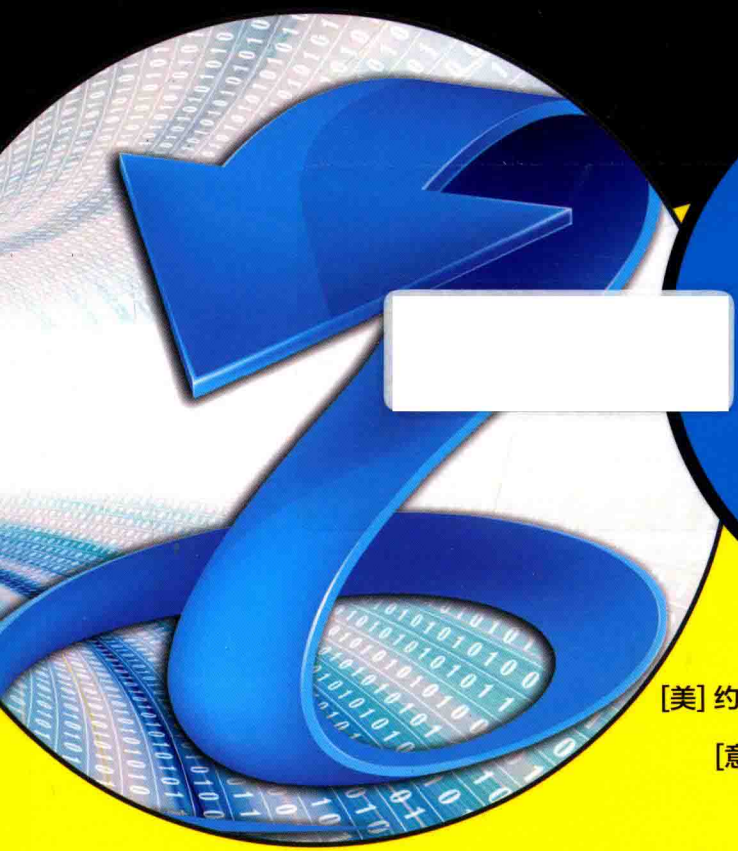
达人速

Python 数据科学入门

Drone for dummies
A Wiley Brand



本书将教会你

- 掌握Python数据分析来编程
- 使用Python对象、函数、模块和库
- 应用概率和随机分布等统计学概念
- 使用NumPy、SciPy、Scikit-learn和Pandas库

[美] 约翰·保罗·穆勒（John Paul Mueller）
[意] 卢卡·马萨罗（Luca Massaron）著
徐旭彬 译

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

达人速

Python 数据科学入门

Drone for dummies

A Wiley Brand

[美] 约翰·保罗·穆勒（John Paul Mueller）

[意] 卢卡·马萨罗（Luca Massaron）著

徐旭彬 译

人 民 邮 电 出 版 社

北 京

图书在版编目（CIP）数据

Python数据科学入门 / （美）约翰·保罗·穆勒
(John Paul Mueller)，（意）卢卡·马萨罗
(Luca Massaron) 著；徐旭彬 译. -- 北京：人民邮
电出版社，2018.5
（达人迷）
ISBN 978-7-115-47962-4

Ⅰ. ①P… Ⅱ. ①约… ②卢… ③徐… Ⅲ. ①软件工
具－程序设计 Ⅳ. ①TP311.561

中国版本图书馆CIP数据核字(2018)第073392号

版权声明

◆ 著　[美] 约翰·保罗·穆勒（John Paul Mueller）
[意] 卢卡·马萨罗（Luca Massaron）
译　徐旭彬
责任编辑　陈冀康
责任印制　焦志炜

◆ 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
三河市君旺印务有限公司印刷

◆ 开本：800×1000　1/16
印张：24.5
字数：556 千字　2018 年 5 月第 1 版
印数：1－3 000 册　2018 年 5 月河北第 1 次印刷
著作权合同登记号　图字：01-2016-8616 号

定价：69.00 元

读者服务热线：(010)81055410　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

内容提要

本书的目标是介绍如何使用 Python 语言及其工具，解决和数据科学所关联的复杂任务。

全书共 6 个部分，分 22 章，涵盖了 Python 数据科学基础知识，数据的采集、整理、整形、应用，数据的可视化，数据分析和处理，数据学习，以及和数据科学相关的 10 个话题等。本书将重点放在使用正确的工具上，教读者如何使用 Anaconda、MatPlotLib、NumPy、pandas、Scikit-learn 等常用的工具来解决数据科学的相关问题。

本书适合对数据科学的知识和应用方法感兴趣的读者阅读，特别适合有志于学习 Python 数据分析和处理的读者学习参考。

作者简介

Luca Massaron 是一名数据科学家和市场研究主管，他专攻多变量统计分析、机器学习和客户洞察，在解决真实世界的问题以及应用推理、统计学、数学挖掘和算法来为利益相关者创造价值方面具有超过 10 年的经验。他始终热衷于数据和分析相关的所有一切以及向专家和非专家演示由数据驱动的知识探索的潜力。相比不必要的复杂性，他更倾向于简单，他相信通过理解并实践不可或缺的关键部分就可以取得数据科学的很多成就。

John Mueller 是一位自由职业作家和技术编辑。他的写作天赋与生俱来，到现在为止已经创作了 97 本书和 600 多篇文章。主题范围从网络到人工智能，从数据库管理到编程实践。他作为技术编辑已经帮助了超过 63 位作者来改善他们的手稿。John 也为《Data Based Advisor》和《Coast Compute》杂志提供技术编辑服务。在 John 为《Data Based Advisor》提供服务期间，他第一次接触到 MATLAB，从此以后他继续跟踪 MATLAB 的发展。John 在 Cubic 公司期间，他接触到了可靠性工程，并保持了在概率方面的兴趣。一定要在 http://blog.johnmuellerbooks.com/ 读一读 John 的博客。

当 John 不在计算机旁工作时，你可以发现他在花园中、伐木，或者说就是欣赏大自然。John 也喜欢酿造葡萄酒、烘焙饼干和编织，以及制作甘油肥皂和蜡烛。你可以通过 John@JohnMuellerBook.com 联系他，其个人网站为 http://www.johnmuellerbooks.com。

Luca的献辞

我想要把本书献给我的父母，Renzo 和 Licia，他们都喜欢简单而且便于解释的想法，以及那些阅读我们所写的书的人。现在他们更多地理解了我在数据科学中的日常工作，并且理解了这个新领域是将要如何改变我们理解世界的方式以及其中的运转方式。

John的献辞

本书献给世上的科学家、工程师、梦想家和哲学家——那些名不见经传的群体对这个星球上的每个人的生活做出了如此之大的改变。

Luca的致谢

向我的家庭、Yukiko 和 Amelia 表达我的最大谢意，为了他们的支持和充满爱意的耐心。

我也要感谢我的全部 Kaggler 搭档，为了他们的帮助和不停歇地交换想法和观点。特别要致谢 Alberto Boschetti、Giuliano Janson、Bastiaan Sjardin 和 Zacharias Voulgaris。

John的致谢

感谢我的妻子 Rebecca。虽然她现在已经不在人世了，但她的精神在我所写的每本书中，在书页上的每个单词中出现。当没有一个人相信我时，她信任我。

应该感谢 Russ Mullen，为了他对本书的技术编辑。他大大增加了你在这里所见到的资料的准确性和深度。Russ 查找了难以发现的 URL，并也提出了许多建议，极其努力地工作来帮助本书的出版。

应该感谢我的代理 Matt Wagner，他帮助我拿到出版合同并始终会料理大多数作者无暇顾及的所有细节。我感谢他的协助。

很多人阅读了本书的全部或一部分来帮助我改善方法、测试脚本并提供了所有读者可能普遍会想到的内容。这些无偿的志愿者的帮助方式太多了，这里无法一一提及。我特别要感谢 Eva Beattie、Glenn A. Russell、Osvaldo Téllez Almirall 和 Thomas Zinckgraf 的努力，他们提供了广泛的内容，阅读了整本书并无私地致力于这个项目。

最后，我想要感谢 Kyle Looper、Susan Christophersen 和余下的编辑和产品的同事们。

出版者的致谢

策划编辑：Katie Mohr

高级编辑助理：Cherie Case

项目及文字编辑：Susan Christophersen

项目协调者：Vinitha Vikraman

技术编辑：Russ Mullen

封面图片：© iStock.com/Magnilion; © iStock.com/nadla

编辑助理：Claire Brock

前言

你每天完全依靠着数据科学来执行众多的任务或者从他人那里获得服务。事实上，你可能以自己未曾预料过的方式使用了数据科学。例如，当你在今天早晨使用你喜欢的搜索引擎来查找某些事物时，它会对可选事项做出建议。这些事项是由数据科学所提供的。当你去看医生并查明了你所发现的肉块并不是癌症时，医生可能凭借数据科学的帮助来做出诊断。事实上，你可能每天使用数据科学来工作，即使你并不知情。本书不仅让你开始着手使用数据科学来执行众多现实的任务，而且也帮助你了解有多少地方用到了数据科学。通过知晓如何来回答数据科学的问题以及从何使用数据科学，你就比其他人获得了显著的优势，增加了你升职或者获得你真心想要的新工作的机会。

本书简介

本书的主要目标是通过向你展示数据科学不但真的有趣而且借助 Python 相当具有可操作性，从而驱走你对数据科学的恐惧。你可能会希望成为一名计算机科学天才来执行通常和数据科学所关联的复杂任务，但这远远与事实不符。Python 具有许多有用的库来为你在后台挑起所有的重担。你甚至没有意识到有多少工作正在进行中，你不必在意。你真正需要知道的所有一切就是你想要执行的特定的任务，并且 Python 让这些任务变得唾手可得。

本书的部分重点就是在使用正确的工具上面。你从使用 Anaconda 开始，这是一个包含了 IPython 和 IPython Notebook 的产品——两个让人容易使用 Python 来工作的工具。你在一个完全交互式的环境下用 IPython 来做实验。你在 IPython Notebook 中存放的代码具有可供演示的质量，你可以在文档中恰当地混合一些演示元素。它不像完全真正地使用一个开发环境。

你也会在本书中发现一些有趣的技术。例如，你可以使用 MatPlotLib 来为你的全部数据科学实验制作绘图，为此本书给你提供了所有的细节。本书也花费了相当的时间仅仅向你展示可供利用的东西，以及你该如何使用它来执行一些真正有趣的计算。很多人想要知道如何执行手写识别——如果你是他们中的一员，你能够使用本书在此过程中领先一步。

当然，你可能还是会对完整的编程环境问题感到焦虑，本书也不会把你抛

弃在黑暗中。一开始，你就可找到完整的 Anaconda 安装教程以及一个针对你所需运行的基础 Python 编程的快速入门书（和参考）。重点就是让你上手并尽快运行，让例子变得直截了当和简单，这样代码就不会成为学习的绊脚石。

甚至为了更快地吸收概念，本书使用了下面的约定。

- 你应该键入的文字在本书中用粗体显示。除非当你正在遍历一个步骤列表的时候：因为每个步骤是粗体的，所以要键入的文字就不是粗体的。
- 当你看到以斜体显示的作为键入序列的部分词组时，你需要把它们的值替换成一些对你有用的文字。例如，如果你看到“键入**你的名字**并按回车键”，你就需要把你的名字替换成你实际的名字。
- Web 地址和编程源码以 `monofont` 字体显示。如果你正在一个与因特网连接的设备上阅读本书的电子版时，注意你可以点击 web 地址来访问那个网站，就如这样：`http://www.dummies.com`。
- 当你需要键入命令序列时，你会看到它们被一个特殊的箭头分割开了，就如这样：文件→新文件。在这种情况下，你首先要去文件菜单，接着在那个菜单上选择“新文件”项。结果你就会看到一个新文件被创建出来了。

对读者的要求

你可能发现我们对你所做的一切假设都是难以置信的——毕竟，我们甚至还没见过你！尽管大多数假设可能显得多余，我们还是做出了这些假设来为本书提供一个起点。

你要对想要使用的平台有所了解，这是重要的，因为本书在这方面没有提供任何指导（第 3 章确实提供了 Anaconda 的安装教程）。为了提供给你 Python 在数据科学方面应用的最大信息量，本书没有讨论任何平台相关的问题。在你开始使用本书工作前，你真的需要了解如何安装应用程序，如何使用应用程序，以及如何用你所选择的平台做一般性的工作。

本书不是一本数学入门读物。是的，你会看见许多复杂数学的例子，但是重点是在帮助你使用 Python 和数据科学来执行分析任务，而不是学习数学理

论。第 1 章和第 2 章让你更精确地理解你所需知道的内容，从而得以成功使用本书。

本书也假设你能够在因特网上获取资源。全书散布着许多可以提升你的学习经验的在线参考资料。无论如何，只有当你实际找到并使用了这些附加资源，它们才是有用的。

本书所使用的图标

当你阅读本书时，你会看见在页边上的图标，它们用来表示资料的有趣程度（或者在有些情况下没有）。本小节简单地描述了本书中的每个图标。

提示是很有用的，因为它们帮你节约时间或者运行一些无需大量额外工作的任务。在本书中的提示是节约时间的技术或者指向了你应该尝试的资源，从而可以从 Python 得到最大的收益，或者可以运行数据科学相关的任务。

你应该避免做出任何以警告标志标记的事情。否则，你可能发现你的应用程序会如预期的那样失效，你会从看起来无懈可击的方程式中得到不正确的答案，或者（在最糟糕的场景下）你会丢失数据。

无论何时当你看到这个图标时，想一想高级的提示或技术。你可能会发现这些花絮信息是有用的，只是用语言表达出来太繁琐，或者它们可能包含了让程序得以运行起来所需的解决方案。无论何时只要你喜欢，就可以略过这些信息。

如果你没有从一个特定章节或小节中吸取到任何东西，记住以这个图标所标记的资料。这段文字通常包含了一个必不可少的过程或者一些信息，它们是你用 Python 工作或者成功运行数据科学相关任务所必须要知道的。

本书之外的内容

本书不是你的 Python 或者数据科学经历的终点——它其实只是一个开端。我们提供了在线内容使得本书更加灵活而且更好地满足你的需求。那样，当我们接收到你的邮件时，我们能够解决问题并告诉你 Python 的升级或者它所关联的附属物的升级如何对本书的内容产生影响。事实上，你可以参阅这些很棒的附件。

- **作弊单：**你记不记得在学校使用作弊条就为了在测试中取得一个更好的成绩？你这样做过吗？嗯，作弊单就是这样一种类型。它提供给你一些特殊的便条，就是关于你能够使用 Python、IPython、IPython Notebook 和数据科学来做的任务，这不是每个人都知道的。你能够在 http://www.dummies.com/how-to/content/python-for-datascience-for-dummies-cheat-sheet.html 上找到针对本书的作弊单。它包含了真正简洁的信息，例如在使用 Python 时使人们遭受痛苦的最常见的编程错误。
- **Dummies.com 的在线文章：**许多读者把“For Dummies”系列书籍的部件页省略过去了，所以出版商决定补救。现在你有了一个真正充分的理由来阅读部件页——在线的内容。每个部件页都有一篇和它关联的文章来提供不适合放在书中的额外的有趣信息。你能够在 http://www.dummies.com/extras/pythonfordatascience 上找到针对本书的文章。
- **更新：**有时候会发生变更。例如，在写本书期间，当我们预测未来时，我们可能没有看到即将浮现出来的变化。在过去，这种可能性简直意味着书籍变得过时和可用性的降低，但是现在你能够在 http://www.dummies.com/extras/pythonfordatascience 上找到本书的更新。
- **配套文件：**嗨！有没有谁真的想键入书中所有的代码并手工重建所有的绘图？绝大多数读者宁愿花时间实际地用 Python 来工作，执行数据科学任务，看看它们能够去实践的有趣的事情，而不是去键入代码。对你来说，幸运的是，在书中所用的例子可供下载，所以你所需做的一切就是阅读本书来学习 Python 的数据科学应用技术。你可以在 http://www.dummies.com/store/product/Python-for-Data-Science-For-Dummies.productCd-1118844181,descCd-DOWNLOAD.html 上找到这些文件。

如何阅读本书

是时候开启你的 Python 数据科学冒险之旅了！如果你对 Python 和它在数据科学方面的应用完全是一个新手，你应该从第 1 章开始并且一步一个脚印按照书本循序渐进来吸收尽可能多的资料。

如果你是一个新手，完全迫不及待地想要尽可能快地上手 Python 数据科学，你

可以跳到第 3 章，但要理解以后你可能会对一些主题产生一点困扰。如果你已经安装了 Anaconda（在本书中所使用的编程产品），跳到第 4 章是有可能的，但是要确保略读第 3 章，这样你就知道当我们在写此书时所做的假设。确保在 Python 2.7.9 版本已经安装的前提下安装 Anaconda，从而从书中的源码中获得最佳效果。

对 Python 有一些认识以及安装过 Anaconda 的读者可以直接转到第 5 章来节省阅读时间。当你有疑问时，你总是能够根据需要回到更早的章节。无论如何，在你转移到下一章之前，你要理解每一项技术是如何工作的，这点是重要的。每一个技术点、代码例子及步骤对你都有着重要的教训，如果你开始略过太多的信息，你就可能会错失关键内容。

目录