PACKT PUBLISHING

异步图书 www.epubit.com.cn

60多个实用的开发技巧，帮你探索Python及其强大的数据科学能力

Python数据科学指南

Python Data Science Cookbook

[印度] Gopi Subramanian 著
方延风 刘丹 译

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

Python 数据科学指南

[印度] Gopi Subramanian 著
方延风 刘丹 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Python数据科学指南 / （印）萨伯拉曼尼安（Gopi Subramanian）著 ; 方延风, 刘丹译. -- 北京 : 人民邮电出版社, 2016.12
ISBN 978-7-115-43510-1

Ⅰ. ①P… Ⅱ. ①萨… ②方… ③刘… Ⅲ. ①软件工具－程序设计－指南 Ⅳ. ①TP311.561-62

中国版本图书馆CIP数据核字(2016)第232258号

版权声明

◆ 著 [印度] Gopi Subramanian
译 方延风 刘 丹
责任编辑 胡俊英
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
三河市海波印务有限公司印刷

◆ 开本：800×1000 1/16
印张：25.25
字数：498 千字 2016 年 12 月第 1 版
印数：1－2 500 册 2016 年 12 月河北第 1 次印刷

著作权合同登记号 图字：01-2016-2851 号

定价：79.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

Python 作为一种高级程序设计语言，凭借其简洁、易读及可扩展性日渐成为程序设计领域备受推崇的语言，并成为数据科学家的首选之一。

本书详细介绍了 Python 在数据科学中的应用，包括数据探索、数据分析与挖掘、机器学习、大规模机器学习等主题。每一章都为读者提供了足够的数学知识和代码示例来理解不同深度的算法功能，帮助读者更好地掌握各个知识点。

本书内容结构清晰，示例完整，无论是数据科学领域的新手，还是经验丰富的数据科学家都将从中获益。

译者简介

方延风，高级工程师，现在福建省科学技术信息研究所任职，毕业于清华大学，获得计算机技术工程硕士学位，美国俄勒冈大学访问学者，曾出版过多本计算机图书，目前的研究方向是文本数据挖掘、自然语言处理（Natural Language Processing，NLP）、信息检索技术等。他主要翻译了第 1 章及第 6～10 章的内容。

刘丹，副教授，现任福州外语外贸学院物流系副主任。她主要翻译了第 2～5 章的内容，并对全书内容进行了校译。

作者简介

Gopi Subramanian 是一名数据科学家，他在数据挖掘与机器学习领域有着超过 15 年的经验。在过去的 10 年中，他设计、构思、开发并领导了数据挖掘、文本挖掘、自然语言处理、信息提取和检索等多个项目，涉及不同领域和商务垂直系统，包括工程基础设施、消费金融、医疗保健和材料等多个领域。在忠诚度分析领域，他构思并建立了创新的消费者忠诚度模型，设计了企业范围的个性化促销系统。他在美国和印度的专利局共计申请了 10 多项专利，并以自己的名义出版了许多书籍。目前，他在印度的班加罗尔生活和工作。

审稿人简介

Bastiaan Sjardin 是一位在人工智能、数学及机器学习方面有着雄厚实力的数据科学家和企业家。他从莱顿大学获得了认知科学及数理统计学的授课型硕士学位。在过去的 5 年中，他参与了大量数据科学方面的项目，经常在密歇根大学社会网络分析 Coursera 课程和约翰·霍普金斯大学的实用机器学习课程上担任助教。他常用的编程语言是 R 和 Python。目前是 Quandbee（www.quandbee.com）公司的联合创始人，该公司专门从事机器学习的应用程序开发。

前言

如今，我们生活在一个万物互联的世界，每天都在产生海量数据，不可能依靠人力去分析产生的所有数据并做出决策。人类的决策越来越多地被计算机辅助决策所取代，这也得益于数据科学的发展。数据科学已经深入到我们互联世界中的每个角落，市场对那些十分了解数据科学算法并且有能力用这些算法进行编程的人才需求是不断增长的。数据科学是多领域交叉的，简单列举几个：数据挖掘、机器学习、统计学等。这对那些渴望成为数据科学家以及已经从事这一领域的人们在各方面都倍感压力。把算法当成黑盒子应用到决策系统里，可能会适得其反。面对着无数的算法和数不清的问题，我们需要充分掌握潜在的算法理论，这样才能给每个指定的问题选择最好的算法。

作为一门编程语言，Python 演变至今，已经成为数据科学家的首选之一。在快速原型构建方面，它能充分发挥了脚本语言的能力，对于成熟软件的开发，它精巧的语言结构也十分适合，再加上它在数值计算方面神奇的库，这些都使得它被众多数据科学家和一般的科学编程群体所推崇。不仅如此，由于 Django 和 Flaskweb 等 Web 框架的出现，Python 在 Web 开发人员中也很受欢迎。

本书通过精心编写的内容和精选的主题来满足读者的需求，无论是新手还是经验丰富的数据科学家都将从中获益。本书的内容涉及数据科学的不同方面，包括数据探索、数据分析与挖掘、机器学习、大规模机器学习等。每一章都经过精心编写，带领读者探索相关领域。本书为读者提供了足够的数学知识来理解不同深度的算法功能。只要你有需求，我们都能为好学的读者提供充分的指导，各个主题都十分便于读者学习和理解。

本书给读者带来了数据科学的艺术力和 Python 编程的力量，并帮助他们掌握数据科学的概念。了解 Python 语言并不是死板地跟随本书学习，非 Python 程序员可以从第 1 章开始阅读，里面涵盖了 Python 数据结构及函数编程等概念。

前几章涵盖了数据科学的基础知识，后面的章节则致力于高级数据科学算法。目前最先进的算法已经引领数据科学家在不同的行业实践中进行探索，这些算法包括集成方法、随机森林、正则化回归等，书中将会详细介绍。一些在学术界流行而仍未广泛引入到主流应用中的算法，例如旋转森林等在文中也有详细介绍。

目前市场上有许多个人撰写的数据科学方面的书籍，但我认为它们在将隐藏在数据科学算法背后的数学原理和一些实施中的细节相结合方面仍存在很大空缺，本书志在填补这一空白。每一个主题，恰如其分的数学知识讲解能引导读者理解算法工作原理。我相信读者可以在他们的应用中充分感受这些方法带来的效益。

这里有一个忠告，虽然我们尽可能用客观的语言给读者解释这些主题，但它们并没有作为成品在极端的条件下进行过严格测试。成品的数据科学代码必须符合严格的工程规范。

本书可以作为学习数据科学方法的指南和快速参考书。这是一本独立的、介绍数据科学给新手和一些有一点算法基础的人的书，帮助他们成为这个行业的专家。

本书的主要内容

第 1 章，Python 在数据科学中的应用，介绍了 Python 内置的数据结构及函数，为学习数据科学编程奠定了基础。

第 2 章，Python 环境，介绍了 Python 的科学编程和绘图库，包括 NumPy、matplotlib 和 scikit-learn 等。

第 3 章，数据分析——探索与争鸣，覆盖了数据预处理、转换方法来探测性执行数据分析任务等内容，以便有效地构建数据科学算法。

第 4 章，数据分析——深入理解，引入降维概念来解决数据科学中的维数问题，详细讨论了从简单方法到最先进的降维技术。

第 5 章，数据挖掘——海底捞针，讨论了无监督数据挖掘技术，先精心探讨了基于距离方法、核方法等内容，接着对聚类与异常点检测技术进行详细讨论。

第 6 章，机器学习 1，涵盖了有监督数据挖掘技术，包括最近邻算法、朴素贝叶斯算法及分类树算法，开始部分就重点强调了监督学习的数据准备工作。

第 7 章，机器学习 2，介绍了回归问题和包括 LASSO 和岭回归在内的正则化主题。最后，讨论了运用交叉检验技术为这些方法选择超参数。

第 8 章，集成方法，介绍了各种集成方法，包括挂袋法、提升法及梯度提升法。本章展现了如何在数据科学领域创建强大的、最先进的方法，不是对给定的问题建立单一的模型，而是在集成中构建大量的模型。

第 9 章，生长树，介绍了更多的基于树的挂袋法，基于其对噪声的健壮性和对不同问题的通用性，它们在数据科学界非常流行。

第 10 章，大规模机器学习——在线学习，涵盖了大规模机器学习及解决如此大规模问题的合适算法，其中的算法使用数据流进行工作，使用的数据无法完全加载到内存中。

读者须知

本书所有主题中的代码都在一台安装了 64 位 Windows 7 操作系统的计算机上进行开发和测试，其配置为 Intel i7 CPU 和 8GB 内存。

本书中使用的开发语言和库版本为：Python 2.7.5、NumPy 1.8.0、SciPy 0.13.2、Matplotlib 1.3.1、NLTK 3.0.2 和 scikit-learn 0.15.2。

这里的代码通过适当的库也能在 Linux 各种发行版和 Macs 上运行。另外一种方式是采用这些版本的库创建一个 Python 虚拟环境，这样就能运行所有主题中的代码。

本书的目标读者

本书适合于各个层次的数据科学专业人士，包括学生、业内人士，从新手到专家，各章节的不同主题契合了不同读者的需求。第 1～5 章，新手级读者可以花一些时间认识数据科学。专家级读者可以阅读后面的章节参考并理解如何用 Pyhton 实施一些先进的技术。本书涉及适当的数学内容以满足希望理解数据科学的程序员，给他们一些必要的参考。没有 Python 基础的人也可以有效地使用本书，本书的第 1 章介绍了基于 Python 编程语言的数据科学。如果你已有编程基础，这将对你很有帮助。本书的编写框架基本自成体系，能给入门级读者讲解数据科学，帮助他们成为这方面的专家。

体例

在本书中，经常按：准备工作、操作方法、工作原理、更多内容、参考资料等主题进行讲解。

为了清楚提示如何能够完成这些主题，我们运用如下所示的各个部分。

准备工作

这部分告诉你本主题要讲述的内容，并介绍如何安装软件及所需的初步设置。

操作方法

本部分包含所需依照的操作步骤。

工作原理

这部分通常是对之前内容的详细解释。

更多内容

为了使读者知道更多相关主题的知识，这部分提供了一些附加信息。

参考资料

这部分为主题提供了其他有用信息的配套链接。

约定

在本书中，你会发现一些文本样式被用来区别不同种类的信息，以下是一些样式例子及其各自的含义。

文本中的代码，如函数名，如下所示。

我们调用 `get_iris_data()` 函数来获得输入数据，利用 `Scikit-learnd` 库的 `cross_validation` 模型的 `train_test_split` 函数将输入数据集一分为二。

代码块的格式设置如下。

```
# Shuffle the dataset
shuff_index = np.random.shuffle(range(len(y)))
x_train = x[shuff_index,:].reshape(x.shape)
y_train = np.ravel(y[shuff_index,:])
```

公式通常以图像形式提供，格式如下。

$$x_i = \{x_{i1}, x_{i2}, \cdots, x_{im}\} where\ i = 1\ to\ n$$

通常数学部分在每一节的开头部分被提出，某些章节中，各个主题通用的数学知识统一在简介部分进行介绍。

外部链接的格式如下。

http://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html。

第三方库中一些算法实现的细节的说明的规范如下。

“输入的样本被预测的分类被用来当作具有最高的平均预测概率，如果基准评估器没有实施 predict_proba 方法，则诉诸于投票。”

任何引用科技期刊或者论文作为参考文献的地方，格式规范如下。

你可以阅读 Leo Breiman 的论文来了解挂袋法的更多信息，请参见：

Leo Breiman 著，《Bagging predictors.Mach. Learn》24, 2 (1996 年 8 月)，第 123～140 页，DOI=10.1023/A:1018054314350 http://dx.doi.org/10.1023/A:1018054314350。

程序的输出及图形通常以图像形式提供，例如。

```
Single Model Accuracy on Dev data

              precision    recall  f1-score   support

           0       0.83      0.84      0.83        51
           1       0.85      0.83      0.84        54

avg / total       0.84      0.84      0.84       105


 Bagging Model Accuracy on Dev data

              precision    recall  f1-score   support

           0       0.85      0.88      0.87        51
           1       0.88      0.85      0.87        54

avg / total       0.87      0.87      0.87       105
```

任意命令行的输入/输出格式如下。

```
Counter({'Peter': 4, 'of': 4, 'Piper': 4, 'pickled': 4, 'picked': 4,'peppers': 4, 'peck': 4, 'a': 2, 'A': 1, 'the': 1, 'Wheres': 1, 'If': 1})
```

在 Python shell 中我们希望读者能够检查一些变量，指定的格式如下所示。

```
>>> print b_tuple[0]
1
>>> print b_tuple[-1]
c
>>>
```

这个格子里出现的是警告或者重要的注意点。

这个格子里出现的是提示和技巧。

读者反馈

我们永远欢迎来自读者的反馈。让我们知道你对于这本书的想法——哪些是你喜欢的或者不喜欢的。读者的反馈对我们来说十分重要，它可以帮助我们拓展书的内容，将会使你更加有效地使用本书。

读者可以用电子邮件发送反馈内容到邮箱 feedback@packtpub.com，并在邮件主题中提及本书的标题/书名。

如果你在某个主题中有专业经验，并有兴趣编写或参与图书的出版，请访问 www.packtpub.com/authors 中的作者指南。

用户支持

现在，你荣幸地成为了 Packt 出版的图书的拥有者，我们将尽我们所能帮助你从产品中获得最完整的服务。

示例代码下载

对于你所购买的任意 Packt 出版的图书，你可以在 http://www.packtpub.com 登录自己的账户下载示例代码文件。如果你是在别处购买的本书，也可以通过浏览 http://www.packtpub.com/support 网页并登记信息，我们将会通过电子邮件将文件发送给你。

彩图下载

我们还为你提供本书所包含的彩色截图/图像的 PDF 文件，彩图能帮你更好地了解输出结果。你可以从 http://www.packtpub.com/sites/default/files/downloads/1234OT_ColorImages.pdf 下载此文件。

勘误表

虽然我们已尽力确保内容的准确性，但是难免会有错误发生。如果你在本书中发现文本或者代码错误，请告知我们，我们将感激不尽。这样一来，你可以帮助其他读者避免困惑，并帮助我们改进本书的后续版本。如果你发现任何错误，请访问 http://www.packtpub.com/submit-errata 进行举报，选择你的书，单击勘误表提交表单链接，然后填写你所发现的错误详情。一旦你的勘误通过验证，你所提交的内容将被接受，然后勘误将被上传到我们的网站并添加到该书的勘误列表中。

要查看之前提交的勘误信息，请访问 https://www.packtpub.com/books/content/support，在检索框里输入书名，所需的信息就会出现在勘误表栏目中。

著作权保护

在互联网上以不同媒介对拥有版权的材料进行盗版是一直存在的问题，Packt 非常重视版权和许可的保护。如果你遇到我们的作品在互联网上被以任何形式进行非法拷贝，请向我们提供网址或网站名称，使我们可以立即采取措施补救。请将涉嫌盗版材料的地址链接发送到 copyright@packtpub.com，并与我们联系。

我们感谢你在保护作者方面提供的帮助，让我们能带给你更有价值的内容。

联系我们

如果你对本书有任何方面的问题，可以通过发送邮件到 questions@packtpub.com 联系我们，我们将竭尽所能来解决问题。

目录