[美] Prateek Joshi 著 陶俊杰 陈小莉 译

Python 机器学习经典实例

Python Machine Learning Cookbook

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

Python 机器学习经典实例

Python Machine Learning Cookbook

[美] Prateek Joshi 著 陶俊杰 陈小莉 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Python机器学习经典实例 / （美）普拉提克·乔西（Prateek Joshi）著 ; 陶俊杰, 陈小莉译. -- 北京 : 人民邮电出版社, 2017.8
（图灵程序设计丛书）
ISBN 978-7-115-46527-6

Ⅰ. ①P… Ⅱ. ①普… ②陶… ③陈… 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(2017)第178314号

内 容 提 要

在如今这个处处以数据驱动的世界中，机器学习正变得越来越大众化。它已经被广泛地应用于不同领域，如搜索引擎、机器人、无人驾驶汽车等。本书首先通过实用的案例介绍机器学习的基础知识，然后介绍一些稍微复杂的机器学习算法，例如支持向量机、极端随机森林、隐马尔可夫模型、条件随机场、深度神经网络，等等。

本书是为想用机器学习算法开发应用程序的 Python 程序员准备的。它适合 Python 初学者阅读，不过熟悉 Python 编程方法对体验示例代码大有裨益。

◆ 著 [美] Prateek Joshi
译 陶俊杰 陈小莉
责任编辑 朱 巍
执行编辑 徐晓娟
责任印制 彭志环

◆ 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：16.5
字数：390千字 2017年8月第1版
印数：1－3500册 2017年8月河北第1次印刷
著作权合同登记号 图字：01-2016-9523号

定价：59.00元

读者服务热线：(010)51095186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

版权声明

译者序

有一天，忽然想到自己整天面对着52个英文字母、9个数字、32个符号[①]和一个空格，经常加班没有双休日，好傻。时间不断被各种噪声碎片化，完全就是毛姆在《月亮和六便士》里写的，“If you look on the ground in search of a sixpence, you don't look up, and so miss the moon”，整天低头刷手机，却不记得举头望明月。生活也愈发无序，感觉渐渐被掏空。薛定谔的《生命是什么》给我提了个醒，他在“以‘负熵’为生”（It Feeds On ‘negative Entropy’）一节指出：“要活着，唯一的办法就是从环境里不断地汲取负熵。”在介绍了熵的概念及其统计学意义之后，他紧接着在“从环境中引出‘有序’以维持组织”（Organization Maintained By Extracting ‘Order’ From The Environment）一节进一步总结：“一个有机体使本身稳定在较高的有序水平上（等于熵的相当低的水平上）的办法，就是从环境中不断地吸取秩序。”这个秩序（负熵、$k\log(1/n)$）可以是食物，也可以是知识，按主流叫法就是“正能量”（有些所谓正能量却碰巧是增加系统无序水平的正熵）。于是，我开始渐渐放弃那些让人沮丧的老梗，远离那些引发混乱的噪声，重新读书，试着翻译，学会去爱。这几年最大的收获就是明白了“隔行如隔山”的道理，试着循序渐进，教学相长，做力所能及之事，让编程变简单。

一般人都不喜欢编程，更不喜欢动手编程（时间消耗：编写 & 测试 40%、重构 40%、风格 & 文档 20%），却喜欢在心里、嘴上编程：“先这样，再那样，如果要XX，就YY，最后就可以ZZ了。”分分钟就可以说完几万行代码的项目，水还剩大半杯。一旦大期将近，即使要亲自动手Copy代码，也会觉得苦堪搬砖，键盘不是红与黑、屏幕不能左右推、小狗总是闹跑追，不断在数不清的理由中增加自己的熵。偶尔看编程书的目的也很明确，就是为了快速上手，找到答案。当然也是在Google、StackOverflow、GitHub网站上找不到答案之后，无可奈何之举。编程书把看着复杂的知识写得更复杂，虽然大多篇幅不输“飞雪连天射白鹿，笑书神侠倚碧鸳”等经典，且纲举目张、图文并茂，甚至有作者爱引经据典，却极少有令人拍案的惊奇之处。为什么同样是文以载道，编程书却不能像武侠小说一样简单具体，反而显得了无生趣，令人望而却步？虽然编程的目的就是用计算机系统解决问题，但是大多数问题的知识都在其他领域中，许多作者在介绍编程技巧时，又试图介绍一些并不熟悉的背景知识，显得生涩难懂，且增加了书的厚度。

---

① 见文末Python示例代码。

有时我们真正需要的，就是能快刀斩乱麻的代码。（Talk is cheap, show me the code.）编程与研究数理化不同，没有任何假设、原命题、思维实验，并非科学；与舞剑、奏乐、炒菜相似，都是手艺，只要基础扎实，便结果立判。编程技巧也可以像剑谱、乐谱、食谱一般立竿见影，这本《Python机器学习经典实例》正是如此，直接上代码，照着做就行，不用纠结为什么。

机器学习是交叉学科，应用广泛，目前主流方法为统计机器学习。既然是以统计学为基础，那么就不只是计算机与数学专业的私房菜了，机器学习在自然科学、农业科学、医药科学、工程与技术科学、人文与社会科学等多种学科中均可应用。如果你遇到了回归、分类、预测、聚类、文本分析、语音识别、图像处理等经典问题，需要快速用Python解决，那么这本菜谱适合你。即使你对机器学习方法还一知半解，也不妨一试。毕竟是Python的机器学习，还能难到哪儿去呢？目前十分流行的Python机器学习库scikit-learn[①]是全书主角之一，功能全面，接口友好，许多经典的数据集和机器学习案例都来自Kaggle[②]。若有时间追根溯源，请研究周志华教授的《机器学习》西瓜书，周教授啃着西瓜把机器学习调侃得淋漓尽致，详细的参考文献尤为珍贵。但是想当作菜谱看，拿来就用，还是需要费一番功夫；若看书不过瘾，还有吴恩达（Andrew Ng）教授在Coursera上的机器学习公开课[③]，机器学习入门最佳视频教程，吴教授用的工具是Matlab的免费开源版本Octave[④]，你也可以用Python版[⑤]演示教学示例。

学而时习之，不亦乐乎。学习编程技巧，解决实际问题，是一件快乐的事情。希望这本Python机器学习经典案例，可以成为你的负熵，帮你轻松化解那些陈年老梗。如果再努努力，也许陆汝铃院士在《机器学习》序言中提出的6个问题[⑥]，你也有答案了。

示例代码：

```
"""打印ASCII字母表、数字、标点符号"""
import string
for item in [string.ascii_letters,
```

① scikit-learn网址：http://scikit-learn.org/stable/。

② Kaggle是一个2010年成立的数据建模和数据分析竞赛平台，全球数据科学家、统计学家、机器学习工程师的聚集地，上面有丰富的数据集，经典的机器学习基础教程，以及让人流口水的竞赛奖金，支持Python、R、Julia、SQLite，同时也支持jupyter notebook在线编程环境，2017年3月8日被谷歌收购。

③ 分免费版和付费版（购买结业证书），学习内容一样，https://zh.coursera.org/learn/machine-learning。

④ Octave下载地址：https://www.gnu.org/software/octave/。

⑤ GitHub项目：https://github.com/mstampfer/Coursera-Stanford-ML-Python。

⑥ 陆院士的6个问题分别是：1. 机器学习早期的符号机器学习，如何在统计机器学习主流中发展；2. 统计机器学习算法中并不现实的“独立同分布”假设如何解决；3. 深度学习得益于硬件革命，是否会取代统计机器学习；4. 机器学习用的都是经典的概率统计、代数逻辑，而目前仅有倒向微分方程用于预测，微分几何的流形用于降维（流形学习，Manifold learning，科普见博文http://blog.pluskid.org/?p=533），只是数学领域的一角，其他现代数学理论是否可以参与其中；5. 机器学习方法仍不够严谨，例如目前流形学习直接将高维数据集假设成微分流形，需要进一步完善；6. 大数据与统计机器学习是如何互动的。

```
            string.digits,
            string.punctuation]:
    print('{}\t{}'.format(len(item), item))
```

输出结果：

```
52 abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ
10 0123456789
32 !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~
```

前　言

在如今这个处处以数据驱动的世界中，机器学习正变得越来越大众化。它已经被广泛地应用于不同领域，如搜索引擎、机器人、无人驾驶汽车等。本书不仅可以帮你了解现实生活中机器学习的应用场景，而且通过有趣的菜谱式教程教你掌握处理具体问题的算法。

本书首先通过实用的案例介绍机器学习的基础知识，然后介绍一些稍微复杂的机器学习算法，例如支持向量机、极端随机森林、隐马尔可夫模型、条件随机场、深度神经网络，等等。本书是为想用机器学习算法开发应用程序的Python程序员准备的。它不仅适合Python初学者（当然，熟悉Python编程方法将有助于体验示例代码），而且也适合想要掌握机器学习技术的Python老手。

通过本书，你不仅可以学会如何做出合理的决策，为自己选择合适的算法类型，而且可以学会如何高效地实现算法以获得最佳学习效果。如果你在图像、文字、语音或其他形式的数据处理中遇到困难，书中处理这些数据的机器学习技术一定会对你有所帮助！

本书内容

第1章介绍各种回归分析的监督学习技术。我们将学习如何分析共享自行车的使用模式，以及如何预测房价。

第2章介绍各种数据分类的监督学习技术。我们将学习如何评估收入层级，以及如何通过特征评估一辆二手汽车的质量。

第3章论述支持向量机的预测建模技术。我们将学习如何使用这些技术预测建筑物里事件发生的概率，以及体育场周边道路的交通情况。

第4章阐述无监督学习算法，包括K-means聚类和均值漂移聚类。我们将学习如何将这些算法应用于股票市场数据和客户细分。

第5章介绍推荐引擎的相关算法。我们将学习如何应用这些算法实现协同滤波和电影推荐。

第6章阐述与文本数据分析相关的技术，包括分词、词干提取、词库模型等。我们将学习如何使用这些技术进行文本情感分析和主题建模。

第7章介绍与语音数据分析相关的算法。我们将学习如何建立语音识别系统。

第8章介绍分析时间序列和有序数据的相关技术，包括隐马尔可夫模型和条件随机场。我们将学习如何将这些技术应用到文本序列分析和股市预测中。

第9章介绍图像内容分析与物体识别方面的算法。我们将学习如何提取图像特征，以及建立物体识别系统。

第10章介绍在图像和视频中检测与识别面部的相关技术。我们将学习使用降维算法建立面部识别器。

第11章介绍建立深度神经网络所需的算法。我们将学习如何使用神经网络建立光学文字识别系统。

第12章介绍机器学习使用的数据可视化技术。我们将学习如何创建不同类型的图形和图表。

阅读背景

Python 2.*x*和Python 3.*x*的版本之争尚未平息[①]。一方面，我们坚信世界会向更好的版本不断进化，另一方面，许多开发者仍然喜欢使用Python 2.*x*的版本。目前许多操作系统仍然内置Python 2.*x*。本书的重点是介绍Python机器学习，而非Python语言本身。另外，考虑到程序的兼容性，书中用到了一些尚未被迁移到Python 3.*x*版本的程序库，因此，本书依然选择Python 2.*x*的版本。我们会尽最大努力保持代码兼容各种Python版本，因为这样可以让你轻松地理解代码，并且很方便地将代码应用到不同场景中。

读者对象

本书是为想用机器学习算法开发应用程序的Python程序员准备的。它适合Python初学者阅读，不过熟悉Python编程方法对体验示例代码大有裨益。

内容组织

在本书中，你会频繁地看到下面这些标题（准备工作、详细步骤、工作原理、更多内容、另请参阅）。

为了更好地呈现内容，本书采用以下组织形式。

① 2020年之前应该不会终结。——译者注

准备工作

这部分首先介绍本节目标，然后介绍软件配置方法以及所需的准备工作。

详细步骤

这部分介绍具体的实践步骤。

工作原理

这部分通常是对前一部分内容的详细解释。

更多内容

这部分会补充介绍一些信息，帮助你更好地理解前面的内容。

另请参阅

这部分提供一些参考资料。

排版约定

在本书中，你会发现一些不同的文本样式。这里举例说明它们的含义。

嵌入代码、命令、选项、参数、函数、字段、属性、语句等，用等宽的代码字体显示：“这里，我们将25%的数据用于测试，可以通过test_size参数进行设置。”

代码块用如下格式：

```
import numpy as np
import matplotlib.pyplot as plt

import utilities

# Load input data
input_file = 'data_multivar.txt'
X, y = utilities.load_data(input_file)
```

命令行输入或输出用如下格式：

```
$ python object_recognizer.py --input-image imagefile.jpg --model-file
erf.pkl --codebook-file codebook.pkl
```

新术语和重要文字将采用黑体字。你在屏幕上看到的内容，包括对话框或菜单里的文本，都将这样显示：“如果你将数组改为`(0, 0.2, 0, 0, 0)`，那么Strawberry部分就会高亮显示。”

读者反馈

我们非常欢迎读者的反馈。如果你对本书有些想法，有什么喜欢或是不喜欢的，请反馈给我们，这将有助于我们出版充分满足读者需求的图书。

一般性反馈请发送电子邮件至feedback@packtpub.com，并在邮件主题中注明书名。

如果你在某个领域有专长，并有意编写一本书或是贡献一份力量，请参考我们的作者指南，地址为http://www.packtpub.com/authors。

客户支持

你现在已经是引以为傲的Packt读者了。为了能让你的购买物超所值，我们还为你准备了以下内容。

下载示例代码

你可以用你的账户从http://www.packtpub.com下载所有已购买Packt图书的示例代码文件。如果你是从其他途径购买的本书，可以访问http://www.packtpub.com/support并注册，我们将通过电子邮件把文件发送给你。

可以通过以下步骤下载示例代码文件：

(1) 用你的电子邮件和密码登录或注册我们的网站；
(2) 将鼠标移到网站上方的**客户支持**（SUPPORT）标签；
(3) 单击**代码下载与勘误**（Code Downloads & Errata）按钮；
(4) 在**搜索框**（Search）中输入书名；
(5) 选择你要下载代码文件的书；
(6) 从下拉菜单中选择你的购书途径；
(7) 单击**代码下载**（Code Download）按钮。

你也可以通过单击Packt网站上本书网页上的**代码文件**（Code Files）按钮来下载示例代码，该网页可以通过在**搜索框**（Search）中输入书名获得。以上操作的前提是你已经登录了Packt网站。

下载文件后，请确保用以下软件的最新版来解压文件：

- WinRAR / 7-Zip for Windows；
- Zipeg / iZip / UnRarX for Mac；
- 7-Zip / PeaZip for Linux。

本书的代码包也可以在GitHub上获得，网址是https://github.com/PacktPublishing/Python-Machine-Learning-Cookbook。另外，我们在https://github.com/PacktPublishing上还有其他书的代码包和视频，请需要的读者自行下载。

下载本书的彩色图片

我们也为你提供了一份PDF文件，里面包含了书中的截屏和图表等彩色图片，彩色图片能帮助你更好地理解输出的变化。下载网址为https://www.packtpub.com/sites/default/files/downloads/PythonMachineLearningCookbook_ColorImages.pdf。

勘误

虽然我们已尽力确保本书内容正确，但出错仍旧在所难免。如果你在书中发现错误，不管是文本还是代码，希望能告知我们，我们将不胜感激。这样做，你可以使其他读者免受挫败，也可以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，选择本书，单击**勘误表提交表单**（Errata Submission Form）的链接，并输入详细说明。[①]勘误一经核实，你提交的内容将被接受，此勘误会上传到本公司网站或添加到现有勘误表。

访问https://www.packtpub.com/books/content/support，在搜索框中输入书名，可以在**勘误**（Errata）部分查看已经提交的勘误信息。

盗版

任何媒体都会面临版权内容在互联网上的盗版问题，Packt也不例外。Packt非常重视版权保护。如果你发现我们的作品在互联网上被非法复制，不管以什么形式，都请立即为我们提供相关网址或网站名称，以便我们寻求补救。

请把可疑盗版材料的链接发到copyright@packtpub.com。

保护我们的作者，就是保护我们继续为你带来价值的能力，我们将不胜感激。

---

① 中文版勘误可以到http://www.ituring.com.cn/book/1894查看和提交。——编者注

问题

如果你对本书内容存有疑问，不管是哪个方面的，都可以通过questions@packtpub.com联系我们，我们会尽最大努力解决。

电子书

扫描如下二维码，即可获得本书电子版。

目　录

第1章 监督学习

在这一章，我们将介绍以下主题：

- 数据预处理技术
- 标记编码方法
- 创建线性回归器（linear regressor）
- 计算回归准确性
- 保存模型数据
- 创建岭回归器（ridge regressor）
- 创建多项式回归器（polynomial regressor）
- 估算房屋价格
- 计算特征的相对重要性
- 评估共享单车的需求分布

1.1 简介

如果你熟悉机器学习的基础知识，那么肯定知道什么是监督学习。监督学习是指在有标记的样本（labeled samples）上建立机器学习的模型。例如，如果用尺寸、位置等不同参数建立一套模型来评估一栋房子的价格，那么首先需要创建一个数据库，然后为参数打上标记。我们需要告诉算法，什么样的参数（尺寸、位置）对应什么样的价格。有了这些带标记的数据，算法就可以学会如何根据输入的参数计算房价了。

无监督学习与刚才说的恰好相反，它面对的是没有标记的数据。假设需要把一些数据分成不同的组别，但是对分组的条件毫不知情，于是，无监督学习算法就会以最合理的方式将数据集分成确定数量的组别。我们将在后面章节介绍无监督学习。

建立书中的各种模型时，将使用许多Python程序包，像NumPy、SciPy、scikit-learn、matplotlib等。如果你使用Windows系统，推荐安装兼容SciPy关联程序包的Python发行版，网址为http://www.scipy.org/install.html，这些Python发行版里已经集成了常用的程序包。如果你使用