直观的数据项目，应用高级机器学习方法解决日常问题

Python 机器学习实践指南

Python Machine Learning Blueprints

[美] Alexander T. Combs 著
黄申 译

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

Python机器学习实践指南

[美] Alexander T. Combs 著
黄申 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Python机器学习实践指南 / (美) 库姆斯(Alexander T. Combs) 著 ; 黄申译. -- 北京 : 人民邮电出版社, 2017.5
书名原文: Python Machine Learning Blueprints
ISBN 978-7-115-44906-1

Ⅰ. ①P… Ⅱ. ①库… ②黄… 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(2017)第050069号

版权声明

◆ 著　[美] Alexander T. Combs
译　黄　申
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
北京市艺辉印刷有限公司印刷
◆ 开本：800×1000　1/16
印张：17
字数：330 千字　2017 年 5 月第 1 版
印数：1－3 000 册　2017 年 5 月北京第 1 次印刷
著作权合同登记号　图字：01-2016-7608 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：京东工商广字第 8052 号

内容提要

机器学习是近年来渐趋热门的一个领域，同时 Python 语言经过一段时间的发展也已逐渐成为主流的编程语言之一。本书结合了机器学习和 Python 语言两个热门的领域，通过易于理解的项目详细讲述了如何构建真实的机器学习应用程序。

全书共有 10 章。第 1 章讲解了 Python 机器学习的生态系统，剩余 9 章介绍了众多与机器学习相关的算法，包括聚类算法、推荐引擎等，主要包括机器学习在公寓、机票、IPO 市场、新闻源、内容推广、股票市场、图像、聊天机器人和推荐引擎等方面的应用。

本书适合 Python 程序员、数据分析人员、对算法感兴趣的读者、机器学习领域的从业人员及科研人员阅读。

作者简介

Alexander T. Combs 是一位经验丰富的数据科学家、策略师和开发人员。他有金融数据抽取、自然语言处理和生成，以及定量和统计建模的背景。他目前是纽约沉浸式数据科学项目的一名全职资深讲师。

审阅者简介

Kushal Khandelwal 是一位数据科学家和全栈开发人员。他的兴趣包括构建可扩展的机器学习和图像处理的软件应用。他擅长 Python 编码，并对各种开源项目做出了积极的贡献。他目前担任 Truce.in 的技术主管，这是一家以农民为中心的创业公司，Kushal 致力于创建可扩展的 Web 应用程序来帮助农民。

译者简介

黄申博士，现任 IBM 研究院资深科学家，毕业于上海交通大学计算机科学与工程专业，师从俞勇教授。微软学者、IBM ExtremeBlue 天才计划成员。长期专注于大数据相关的搜索、推荐、广告以及用户精准化领域。曾在微软亚洲研究院、eBay 中国、沃尔玛 1 号店（现京东 1 号店）和大润发飞牛网担任要职，带队完成了若干公司级的战略项目。同时发表了 20 多篇国际论文，并拥有 10 多项国际专利，《计算机工程》特邀审稿专家，《Elasticsearch 实战》中文版的译者，2016 年出版的《大数据架构商业之路》一书销量和口碑双赢，续作《大数据架构和算法实现之路》将于 2017 年中出版。2015 年，因对业界做出卓越贡献，获得美国政府颁发的“美国杰出人才”称号。

译者序

谈到为什么要翻译这本书，还是一段机缘巧合。那是2015年的下半年，当时我正在撰写自己的原创书籍《大数据架构商业之路：从业务需求到技术方案》。在那本书中，我希望结合一个创业的故事，展示各个阶段可能遇到的大数据课题、业务需求，以及相对应的技术方案，甚至是实践解析。其中，最挑战的部分莫过于案例的分析到技术方案，再到框架编码的逐步展开。因为之前对于这种写作模式没有相关的经验，让人很是苦恼。我也搜寻了市面上相关的中英文书籍，可惜并未发现特别好的范例作为参考。

一次偶然的机会，我在Amazon.com上发现了Alexander T. Combs的《Python Machine Learning Blueprints》。当时此书尚未出版，还是试读本。在阅读样章之后我发现这种写作模式就是我想要的，没有太多的理论和说教，而是结合我们日常生活都会经历的方方面面，包括房产、金融、旅游和电子商务等，提供了可以直接上手的教学内容，让读者可以身临其境，乐在其中，轻松了解机器学习的实用知识。这正是我想要学习的风格！于是我采纳了这种模式，并结合自己的项目经验，一口气完成了《大数据架构商业之路：从业务需求到技术方案》一书。上市之后，读者对这种理论和案例相结合的方式很是赞许。所以，我对《Python Machine Learning Blueprints》一书心存感激，对它何时上市也很是关注。

终于，2016年的7月底，该书的英文版正式发行。我迫不及待地阅读完了原版，和当初试读的感觉一样，这是一本很有创意的书，而且Python和机器学习都是最近几年的技术热点，如果能将这么棒的内容介绍给广大国内的读者，那是多么令人激动的事情！于是，我抱着试试看的心态，联系了人民邮电出版社的编辑陈冀康老师。很幸运，当时此书还没有译者，陈老师审阅我的试译稿之后也表示满意，于是我很荣幸地成为了此书的译者。

不过在翻译的过程中，我也发现了不少细节上的疑问，于是我主动联系了原书的作者Alexander，他总是非常仔细地解答这些问题，使得我信心大增，可以确保译文尽可能地贴近原文。而编辑陈老师也对此举表示了充分的肯定。在此，我要对Alexander和陈老师的

帮助表示衷心的感谢。当然，我也要感谢父母和妻儿的支持，为了此书，我陪伴你们的时间更少了，而你们丝毫没有怨言，让我可以安心地完成每次的写作。

在翻译此书的岁月中，Python、机器学习及其应用在国内外都获得了空前的关注，相关的社区也保持了非常好的活跃度，相信这个技术方向在将来还有很大的空间。希望本书能帮助到每一位热爱 Python 和机器学习的朋友，为中国的人工智能事业尽一份绵薄之力。如果您对本书中的技术细节感兴趣，可以通过如下渠道联系我，很期待和大家的互动和交流。

QQ 36638279

微信 18616692855

邮箱 s_huang790228@hotmail.com

LinkedIn https://cn.linkedin.com/in/shuang790228

扫一扫就能微信联系作者：

个人

公众号

前言

机器学习正在迅速成为数据驱动型世界的一个必备模块。许多不同的领域如机器人、医学、零售和出版等，都需要依赖这门技术。在这本书中，你将学习如何一步步构建真实的机器学习应用程序。

通过易于理解的项目，你将学习如何处理各种类型的数据，如何以及何时应用不同的机器学习技术，包括监督学习和无监督学习。

本书中的每个项目都同时提供了教学和实践。例如，你将学习如何使用聚类技术来发现低价的机票，以及如何使用线性回归找到一间便宜的公寓。本书以通俗易懂、简洁明了的方式，教你如何使用机器学习来收集、分析并操作大量的数据。

本书涵盖的内容

第 1 章，Python 机器学习的生态系统，深入 Python，它有一个深度活跃的开发者社区，而且许多开发者来自科学社区。这为 Python 提供了丰富的科学计算库。在本章中，我们将讨论这些关键库的特性以及如何准备你的环境，以最好地利用它们。

第 2 章，构建应用程序，发现低价的公寓，指导我们构建第一个机器学习应用程序，我们从一个最小但实际的例子开始：建设应用程序来识别低价的公寓。到本章结束，我们将创建一个应用程序，使得寻找合适的公寓变得更容易点。

第 3 章，构建应用程序，发现低价的机票，演示了如何构建应用程序来不断地监测票价。一旦出现异常价格，应用程序将提醒我们，可以快速采取行动。

第 4 章，使用逻辑回归预测 IPO 市场，展示了我们如何使用机器学习决定哪些 IPO 值得仔细研究，而哪些可以直接跳过。

第 5 章，创建自定义的新闻源，介绍如何构建一个系统，它会了解你对于新闻的品味，而且每天都可以为你提供个性化的新闻资讯。

第 6 章，预测你的内容是否会广为流传，检查一些被大家广泛分享的内容，并试图找到这种内容相对于其他人们不愿分享的内容有哪些特点。

第 7 章，使用机器学习预测股票市场，讨论如何构建和测试交易策略。当你试图设计属于自己的系统时，有无数的陷阱要避免，这是一个几乎不可能完成的任务。但是，这个过程有很多的乐趣，而且有的时候，它甚至可以帮你盈利。

第 8 章，建立图像相似度的引擎，帮助你构建高级的、基于图像的深度学习应用。我们还将涵盖深度学习的算法来了解为什么它们是如此的重要，以及为什么它们成为了最近研究的热点。

第 9 章，打造聊天机器人，演示如何从头构建一个聊天机器人。读完之后，你将了解更多关于该领域的历史及其未来前景。

第 10 章，构建推荐引擎，探讨不同类型的推荐系统。我们将看到它们在商业中是如何实现和运作的。我们还将实现自己的推荐引擎来查找 GitHub 资料库。

阅读本书需要准备什么

你需要的是 Python 3.x 和建立真实机器学习项目的渴望。你可以参考随本书的详细代码列表。

本书的读者

本书的目标读者包括了解数据科学的 Python 程序员、数据科学家、架构师，以及想要构建完整的、基于 Python 的机器学习系统的人员。

约定

在这本书中，你会发现许多文本样式，以区分不同种类的信息。这里是某些样式的例子和它们的含义。

文本中的代码、数据库表名称、文件夹名称、文件名、文件扩展名、路径名、虚构的 URL、用户输入和 Twitter 句柄如下所示：“这点可以通过在我们的数据框上调用.corr()来实现。”

代码块的格式设置如下。

```
<category>
   <pattern>I LIKE TURTLES</pattern>
   <template>I feel like this whole <set name="topic">turle</set>
   thing could be a problem. What do you like about them?</template>
</category>
```

任何命令行输入或输出的写法如下。

```
sp = pd.read_csv(r'/Users/alexcombs/Downloads/spy.csv')
sp.sort_values('Date', inplace=True)
```

新术语和重要词语以粗体显示。

读者反馈

我们非常欢迎读者的反馈。让我们知道你对这本书有什么想法——你喜欢哪些内容或不喜欢哪些内容。读者的反馈对我们而言很重要，因为它有助于我们打造各种主题，而且让你获益更多。

对于一般的反馈，通过电子邮件 feedback@packtpub.com 发送，并在消息的主题中提及书的标题。

如果你擅长某个专业的主题，并且你有兴趣撰写或合著一本书，请参阅我们的作者指南 www.packtpub.com/authors。

客户支持

现在你是一名自豪的 Packt 书籍所有者，我们将做一些事情来帮助你从这次购买中获得最大收益。

下载示例代码

在 http://www.packtpub.com，你可以通过自己的账户来下载此书的示例代码文件。如果你在其他地方购买此书，你可以访问 http://www.packtpub.com/support 并注册，我们将文件直接发送给你。

你可以通过以下步骤下载代码文件。

1. 使用你的电子邮件地址和密码登录或注册我们的网站。
2. 将光标指针悬停在顶部的 SUPPORT 选项卡上。
3. 单击 Code Downloads & Errata。
4. 在搜索框 Search 中输入书籍的名称。
5. 选择你要下载代码文件的图书。
6. 在下拉菜单中，选择你在哪里购买的此书。
7. 单击 Code Download。

在 Packt Publishing 的网站上，你也可以单击该书主页上的 Code Files 按钮来下载代码文件。可以在搜索框中输入图书的名称来访问其主页。请注意，你需要登录到你的 Packt 账户。

一旦文件下载完毕，请确保使用以下软件的最新版本来解压缩或提取文件夹。

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

该书的代码包也托管在 GitHub 上：https://github.com/packtpublishing/pythonmachinelearningblueprints。我们还有丰富的来自其他书籍的代码包和视频，位于 https://github.com/PacktPublishing/。去看一下吧！

勘误

虽然我们已经采取一切谨慎的措施，以确保内容的准确性，但错误在所难免。如果你在我们的书中发现一个错误——也许在正文中，也许在代码中——请向我们报告，我们将非常感激。这样，你可以让其他读者避免挫折，并帮助我们改进本书的后续版本。如果你发现任何错误，请访问这个链接进行报告：http://www.packtpub.com/submit-errata，选择你的书，单击 Errata Submission Form 链接，然后输入错误的详细信息。一旦此勘误通过验证，你的提交将被接受，勘误信息将被上传到我们的网站或添加到任何该主题 Errata 部分现有的勘误表。

要查看以前提交的勘误，请访问 https://www.packtpub.com/books/content/support 并在搜索字段中输入书籍的名称。所需信息将出现在 Errata 部分中。

盗版行为

在互联网上出现正版材料的盗版，是所有媒体面临的一个持续性的问题。在 Packt，我们非常重视版权和许可的保护。如果你在互联网上，发现我们作品任何形式的非法副本，请立即向我们提供地址或网站名称，以便我们请求补偿。

请通过 copyright@packtpub.com 与我们联系，并提供疑似盗版材料的链接。

我们感谢你的帮助，这样可以保护我们的作者，并让我们继续为你提供宝贵的内容。

疑问

如果你对本书的任何方面有问题，可以通过 questions@packtpub.com 与我们联系，我们将尽最大努力解决这个问题。

目录

第 1 章
Python 机器学习的生态系统

机器学习正在迅速改变我们的世界。作为人工智能的核心，我们几乎每天都会读到机器学习如何改变日常的生活。一些人认为它会带领我们进入一个风格奇异的高科技乌托邦；而另一些人认为我们正迈向一个高科技天启时代，将与窃取我们工作机会的机器人和无人机敢死队进行持久的战争。不过，虽然权威专家们可能会喜欢讨论这些夸张的未来，但更为平凡的现实是，机器学习正在快速成为我们日常生活的固定装备。随着我们微小但循序渐进地改进自身与计算机以及周围世界之间的互动，机器学习正在悄悄地改善着我们的生活。

如果你在 Amazon.com 这样的在线零售商店购物，使用 Spotify 或 Netflix 这样的流媒体音乐或电影服务，甚至只是执行一次 Google 搜索，你就已经触碰到了机器学习的应用。使用这些服务的用户会产生数据，这些数据会被收集、汇总并送入模型，而模型最终会为每个用户创建个性化的体验来完善服务。

想要深入到机器学习应用的开发中，现在就是一个理想的时机。你会发现，Python 是开发这些应用的理想选择。Python 拥有一个深度的、活跃的开发者社区，许多开发者也来自科学家的社区。这为 Python 提供了一组丰富的科学计算库。在本书中，我们将讨论并使用这些来自 Python 科学栈的库。

在接下来的章节中，我们将一步步学习如何建立各种不同的机器学习应用。但是，在真正开始之前，我们将使用本章剩下的篇幅讨论这些关键库的特性，以及如何准备能充分利用它们的环境。

我们将在本章中介绍以下主题。

- 数据科学/机器学习的工作流程。