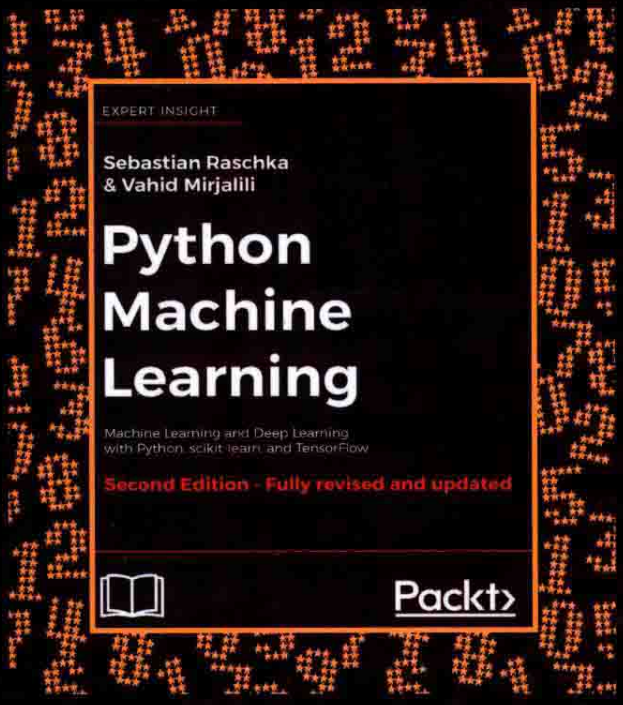
HZ BOOKS 华章IT
[PACKT] PUBLISHING
数据科学与工程技术丛书

Python机器学习

（原书第2版）

[美] 塞巴斯蒂安·拉施卡（Sebastian Raschka）
瓦希德·米尔贾利利（Vahid Mirjalili） 著
陈斌 译



PYTHON MACHINE LEARNING
SECOND EDITION

机械工业出版社
China Machine Press

数据科学与工程技术丛书

PYTHON MACHINE LEARNING
SECOND EDITION

Python机器学习

（原书第2版）

[美] 塞巴斯蒂安·拉施卡（Sebastian Raschka）
瓦希德·米尔贾利利（Vahid Mirjalili） 著
陈斌 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 机器学习（原书第 2 版）/（美）塞巴斯蒂安·拉施卡（Sebastian Raschka），（美）瓦希德·米尔贾利利（Vahid Mirjalili）著；陈斌译．—北京：机械工业出版社，2018.10
（数据科学与工程技术丛书）
书名原文：Python Machine Learning，Second Edition

ISBN 978-7-111-61150-9

I. P… II. ①塞… ②瓦… ③陈… III. 软件工具－程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2018）第 235199 号

本书版权登记号：图字 01-2017-7509

机器学习与预测分析正在改变企业和其他组织的运作方式，本书将带领读者进入预测分析的世界。全书共 16 章，除了简要介绍机器学习及 Python 在机器学习中的应用之外，还系统讲述了数据分类、数据预处理、模型优化、集成学习、情感分析、回归分析、聚类分析、深度学习等内容。书中将机器学习背后的基本理论与应用实践联系起来，通过这种方式让读者聚焦于如何正确地提出问题、解决问题。书中讲解了如何使用 Python 的核心功能以及强大的机器学习库，同时还展示了如何正确使用一系列统计模型。本书可作为学习数据科学的初学者及想进一步拓展数据科学领域认识的读者的参考书，也适合计算机等相关专业的本科生、研究生阅读。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：卢 璐 责任校对：殷 虹
印 刷：北京瑞德印刷有限公司 版 次：2019 年 1 月第 1 版第 1 次印刷
开 本：185mm×260mm 1/16 印 张：24
书 号：ISBN 978-7-111-61150-9 定 价：89.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：(010) 88378991 88361066 投稿热线：(010) 88379604
购书热线：(010) 68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

本书自第1版出版以来，备受广大读者欢迎。与同类书相比，本书除了介绍如何用Python和基于Python的机器学习软件库进行实践外，还对机器学习概念的必要细节进行讨论，同时对机器学习算法的工作原理、使用方法以及如何避免掉入常见的陷阱提供直观且翔实的解释，是Python机器学习入门必读之作。

本书将带领你进入预测分析的世界，并展示为什么Python会成为数据科学领域首屈一指的计算机语言。如果你想更好地从数据中得到问题的答案，或者想要提升并扩展现有机器学习系统的性能，那么这本基于数据科学实践的书籍非常值得一读。它的内容涵盖了众多高效Python库，包括scikit-learn、Keras和TensorFlow等，系统性地梳理和分析了各种经典算法，并通过Python语言以具体代码示例的方式深入浅出地介绍了各种算法的应用，还给出了从情感分析到神经网络的一些实践技巧，这些内容能使你快速解决你和你的团队面临的一些重要问题。

不管你是学习数据科学的初学者，还是想进一步拓展对数据科学领域的认知，本书都是一个重要且不可错过的资源，它能帮助你了解如何使用Python解决数据中的关键问题。

华章IT
HZBOOKS | Information Technology

译 者 序

人工智能的研究从20世纪40年代已经开始，在近80年的发展中经历了数次大起大落。自从2016年AlphaGo战胜顶尖的人类围棋选手之后，人工智能再一次进入了人们的视野，成为当今的热门话题。各大互联网公司都投入了大量的资源研究和开发自动驾驶、人脸识别、语音识别和机器翻译等技术。人类已经开始担忧人工智能可能带来的各种影响。人工智能的最新发展可以说是“古树发新枝”，到底是什么原因使沉寂多年的人工智能技术焕发了青春的活力呢？

首先，移动互联网的飞速发展产生了海量的数据，使我们有机会更加深入地认识社会、探索世界、掌握规律。其次，大数据技术为我们提供了有力的技术手段，使我们可以面对瞬息万变的市场，有效地存储和处理海量数据。再次，计算技术特别是GPU的广泛应用使算力有了大幅度的提升，以前需要几天的运算如今只需要几分钟或几秒钟，这也为人工智能和机器学习的普及与应用提供了计算基础。在这几项技术发展的基础之上，深度学习技术终于破茧而出，成为引领人工智能发展的重要力量。

本书英文版在美国出版后备受欢迎，究其原因，除了机器学习是所有技术人员关注的焦点以外，还在于本书系统性地梳理和分析了机器学习的各种经典算法，最为重要的是作者通过Python语言以具体代码示例深入浅出地介绍了各种算法的应用方法。如果你想了解机器学习并掌握机器学习的具体技术，那就请翻开此书，通过一个又一个案例领略机器学习的风采。所以这本书既是一本初步了解机器学习的启蒙读物，也是一本让你从初学者变成AI专家的教练示范材料。

毋庸置疑，人工智能（AI）、区块链（BlockChain）、云计算（Cloud）、大数据（Big Data）、物联网（IoE）这五项技术（简写为ABCDE）已经成为计算机和互联网技术未来发展的五大核心动力。特别是人工智能技术，它将是继蒸汽机、电力、计算机、互联网之后的又一股重要的革命性力量。之前的几次革命解放的是我们的四肢，而人工智能解放的将是我们的头脑。

关于作者

塞巴斯蒂安·拉施卡是畅销书《Python 机器学习》的作者，他在 Python 编程方面拥有多年经验，曾经就如何实际应用数据科学、机器学习和深度学习做过数次讲座，包括在 SciPy（重要的 Python 科学计算会议）上做的机器学习教学。

虽然塞巴斯蒂安的学术研究项目主要集中在解决计算生物学的问题方面，但他一般喜欢在数据科学、机器学习和 Python 方面进行写作和讨论，而且他总是乐于帮助别人在不需要机器学习的背景下开发数据驱动的解决方案。

他的工作和贡献使他得到了 2016~2017 学年系杰出研究生奖，以及《ACM 计算评论》2016 年度最佳奖。

在闲暇时间里，塞巴斯蒂安喜欢为开源项目做出一些贡献，而他所实现的方法现已成功地用于像 Kaggle 这样的机器学习竞赛。

我想借此机会感谢伟大的 Python 社区和开源软件包的开发人员，他们为我从事科学研究和数据科学创造了完美的环境。

另外，我要感谢我的父母，他们始终鼓励和支持我追求我热爱的道路和事业。

特别感谢 scikit-learn 的核心开发人员。作为这个项目的贡献者，我很高兴能够与伟大的人士合作，他们不仅在机器学习方面非常博学，而且也是优秀的程序员。

瓦希德·米尔贾利利拥有机械工程博士学位，从事大规模分子结构计算模拟新方法的研究。目前他在密歇根州立大学计算机科学与工程系工作，致力于把机器学习应用到各种计算机视觉研究项目中。

瓦希德以 Python 作为编程语言的首选，在学术和研究生涯中积累了丰富的 Python 编码经验。他为密歇根州立大学的工程专业教授 Python 编程，这让他有机会帮助学生理解不同的数据结构并在 Python 中开发高效的代码。

虽然瓦希德的广泛研究兴趣集中在深度学习和计算机视觉应用方面，但他对利用深度学习技术来扩展生物识别数据（如人脸图像）中的隐私保护尤其感兴趣，目的是确保信息不会超出用户想要披露的范围。

此外，他还与一群从事自动驾驶汽车工作的工程师合作，设计了用于检测行人的多光谱图像融合神经网络模型。

我想感谢博士导师阿伦·罗斯博士，他为我提供了在其实验室研究新问题的机会。我还要感谢毗湿奴·伯德利博士，他激发了我对深度学习的兴趣，并让我揭开了核心概念的神秘面纱。

关于审校人员

贾里德·霍夫曼是一位企业家、游戏玩家、讲故事的人以及机器学习的狂热分子和数据库迷。在过去的10年中，他致力于开发软件和分析数据。之前的工作涉及网络安全、金融系统、商业智能、Web服务、开发者工具和业务战略等不同的领域。作为Minecraft公司数据科学团队的创始人，他近几年主要关注大数据和机器学习。工作以外，他会玩游戏或者与朋友和家人一起享受美丽的太平洋西北地区的美好生活。

感谢Packt给我机会参与编纂如此伟大的著作，感谢太太的长期鼓励，
感谢女儿当我在深夜里审校此书和调试代码时能安静地酣睡。

孙怀恩于台湾交通大学取得统计学硕士学位。目前，他作为数据科学家在PEGATRON公司分析生产线的数据。机器学习和深度学习是他主要的研究领域。

前　言

通过新闻媒体的报道，你可能已经了解到机器学习已经成为当代最激动人心的技术。像谷歌、Facebook、苹果、Amazon 和 IBM 这样的大公司基于各自的考虑，已经在机器学习的研究和应用方面投入了巨资。机器学习似乎已经成为流行词，但这绝不是昙花一现。这个激动人心的领域开启了许多新的可能性，已经在日常生活中不可或缺。智能手机的语音助手、为客户推荐合适的产品、防止信用卡欺诈、过滤垃圾邮件、检测和诊断疾病等都是明证。

如果有志于从事深度学习，想更好地解决问题或开展深度学习方面的研究，那么这本书就是为你而写。然而，深度学习背后的理论概念可能艰深难懂。但近几年已经出版了许多机器学习方面的著作，阅读它们有助于通过研发强大的机器学习算法走上机器学习之路。

熟悉机器学习的示例代码及应用是深入该领域的捷径。通过具体的示例学以致用有助于阐明宽泛的概念。请记住，能力越大责任越大！除了用 Python 和基于 Python 的机器学习软件库掌握实践经验外，本书还介绍了机器学习算法背后的数学概念，这对于成功地使用机器学习必不可少。这使得本书有别于其他的纯实战书籍。本书将对机器学习概念的必要细节进行讨论，同时对机器学习算法的工作原理、使用方法以及最为重要的如何避免掉入最常见的陷阱，提供直观且翔实的解释。

如果在谷歌专业网站以“机器学习”作为关键词进行搜索，结果会找到 180 万个出版物。当然我们无法对过去 60 年来所出现的各种不同算法和应用逐一进行考证。然而，本书将开始一个激动人心的旅程，涵盖所有重要的主题和概念，让你在该领域捷足先登。如果你发现所提供的知识还不能解渴，没关系，本书还引用了许多其他有用的资源，供你追踪该领域的精要突破。

如果已经详细研究了机器学习理论，那么本书可以教你如何把知识付诸实践。如果以前用过机器学习技术，想更深入地了解其工作原理，那么本书就是为你而备。如果机器学习对你是全新的领域，那么不必担心，你更有理由为此感到兴奋。我保证机器学习将会改变你解决问题的思路，并让你看到如何通过释放数据的力量来解决问题。

在深入机器学习领域之前，先回答一个最重要的问题：“为什么要用 Python？”答案很简单：Python 功能强大且易于取得。Python 已成为数据科学最常用的编程语言，因为它可以

让我们忘记编程的冗长乏味，同时提供了可以把想法落地、概念直接付诸行动的环境。

我们认为，对机器学习的研究使我们成为更好的科学家、思想家和问题解决者。本书将与你分享这些知识。知识是要靠学习获得的。学习的关键在于热情，而要真正掌握技能只能通过实践。前面的路或许崎岖不平，有些话题可能颇具挑战性，但我们希望你能抓住这个机会，更多地考虑本书所带来的回报。请记住，我们共同踏上这个旅程，本书将为你的军火库添加许多强大的武器，让你以数据驱动的方式来解决最棘手的问题。

本书内容

第 1 章介绍了机器学习在解决不同问题时的主要应用领域。另外，还讨论了构建典型的机器学习模型所需要的基本步骤，从而形成一条导引后续各章节的管道。

第 2 章追溯了机器学习的起源，介绍了二元感知器、分类器和自适应线性神经元。对模式分类的基本原理作了简单介绍，同时关注算法优化和机器学习的交互。

第 3 章描述了基本的机器学习分类算法，并用最流行和全面的开源机器学习软件库 scikit-learn 提供了实际案例。

第 4 章讨论了如何解决未处理数据集中最常见的问题，如数据缺失。也讨论了用来识别数据集中信息量最大特性的几种方法，并教你如何将不同类型的变量作为机器学习算法的适当输入。

第 5 章描述了减少数据集中的特征数，同时保留大部分有用和识别性信息的基本技术。讨论了基于主成分分析的标准降维方法，并将其与有监督学习和非线性变换技术进行了比较。

第 6 章讨论了在预测模型的性能评价中该做和不该做什么。此外，还讨论了模型性能评估的不同度量以及优化机器学习算法的技术。

第 7 章介绍了有效结合多种学习算法的不同概念，讲解了如何建立专家小组来克服个别学习者的弱点，从而产生更准确更可靠的预测。

第 8 章讨论了将文本数据转换为有意义的机器学习算法，以根据文本内容预测人们意见的基本步骤。

第 9 章继续使用前一章中的预测模型，并介绍了使用嵌入式机器学习模型开发网络应用的基本步骤。

第 10 章讨论根据目标和响应变量之间的线性关系建模，从而进行连续预测的基本技术。在介绍了不同的线性模型之后，还讨论了多项式回归和基于树的建模方法。

第 11 章将焦点转移到机器学习的其他子领域，即无监督学习。 用来自于三个基本聚类家族的算法来寻找一组拥有一定程度相似性的对象。

第 12 章扩展了基于梯度的优化概念，该概念在第 2 章中介绍过，用来在 Python 中构建基于常见的强大的多层神经网络的反向传播算法。

第 13 章基于前一章的知识，为更有效地训练神经网络提供实用指南。 该章的重点是

TensorFlow，这是一个开源的 Python 软件库，允许我们充分利用现代的多核 GPU。

第 14 章更详细地介绍了 TensorFlow 的计算图和会话的核心概念。另外，该章还介绍了如何保存会话以及可视化神经网络图等主题，这对本书其他章节的学习会非常有用。

第 15 章讨论了深度神经网络的结构体系，这些结构体系已成为计算机视觉和图像识别领域（卷积神经网络）的新标准。本章讨论了作为特征提取器的卷积层之间的主要概念，并将卷积神经网络体系结构应用于图像识别，以获得近乎完美的识别准确度。

第 16 章介绍了深度学习的另外一种常用的神经网络结构体系，它特别适合于处理序列数据和时间序列数据。在该章中，我们应用不同的递归神经网络体系结构来处理文本数据。作为热身练习，我们将从一个情感分析开始，并学习如何生成全新的文本。

阅读本书需要的材料

要执行本书的示例代码，需要在 MacOS、Linux 或者 Microsoft Windows 操作系统上安装 Python 3.6.0 或更新的版本。本书将持续使用包括 SciPy、NumPy、scikit-learn、Matplotlib 和 pandas 在内的 Python 的科学计算软件库。

第 1 章将为建立 Python 环境及其核心库提供指令和有用的提示。我们将逐渐添加更多的软件库。另外，会分别在不同的章节提供安装指令：用于自然语言处理的 NLTK 库（第 8 章），Flask 网络框架库（第 9 章），Seaborn 统计数据可视化库（第 10 章）和有关图像处理单元的有效神经元网络训练的 TensorFlow（第 13 ～ 16 章）。

本书的目标读者

如果你想知道如何开始用 Python 回答数据方面的关键问题，那就开始学习本书吧！不论是从头学起，还是要扩展数据科学方面的知识，本书都是不可或缺的重要资源。

下载示例代码及彩色图像

本书的示例源码及所有截图和样图，可以从 `http://www.packtpub.com` 通过个人账号下载，也可以访问华章公司官网 `http://www.hzbook.com`，通过注册并登录个人账号下载。

本书的代码包也托管在 GitHub 上，地址如下：

`https://github.com/PacktPublishing/Python-Machine-Learning-Second-Edition`。书中用到的彩色图像截图或者图表的 PDF 文件也可以从 `http://www.packtpub.com/sites/default/files/downloads/PythonMachineLearningSecondEdition_ColorImages.pdf` 下载。

作者简介

塞巴斯蒂安·拉施卡（Sebastian Raschka）

密歇根州立大学博士，他在计算生物学领域提出了几种新的计算方法，还被科技博客Analytics Vidhya评为GitHub上最具影响力的数据科学家。他在Python编程方面积累了丰富经验，曾为如何实际应用数据科学、机器学习和深度学习做过数次讲座，包括在SciPy（重要的Python科学计算会议）上做的机器学习教程。正是因为在数据科学、机器学习以及Python等领域拥有丰富的演讲和写作经验，他才有动力完成本书的撰写，以帮助那些不具备机器学习背景的人设计出有数据驱动的解决方案。他因其工作和贡献获得了2016-2017学年系杰出研究生奖，以及《ACM 计算评论》 2016年度最佳奖。

瓦希德·米尔贾利利（Vahid Mirjalili）

密歇根州立大学计算机视觉与机器学习研究员，致力于把机器学习应用到各种计算机视觉研究项目。他在学术和研究生涯中积累了丰富的Python编程经验，其主要研究兴趣为深度学习和计算机视觉应用。

译者简介

陈斌（Chuck Chen）

现任易宝支付CTO。1989年取得吉林大学硕士学位；1992年任新加坡航空公司高级系统分析师；1999年投身于硅谷互联网技术发展浪潮，曾任日立美国系统集成总监，Abacus首席架构师和Nokia美国首席工程师；2008年任eBay资深架构师，负责移动应用的架构设计。丰富的海外经历，多年的架构经验，深谙移动互联网对传统行业的影响；2014年再次投身易宝，提出大、平、移、商的战略方针，全力推动移动互联网技术，引领行业变革。

目　录