HZ BOOKS 华章IT
[PACKT] PUBLISHING

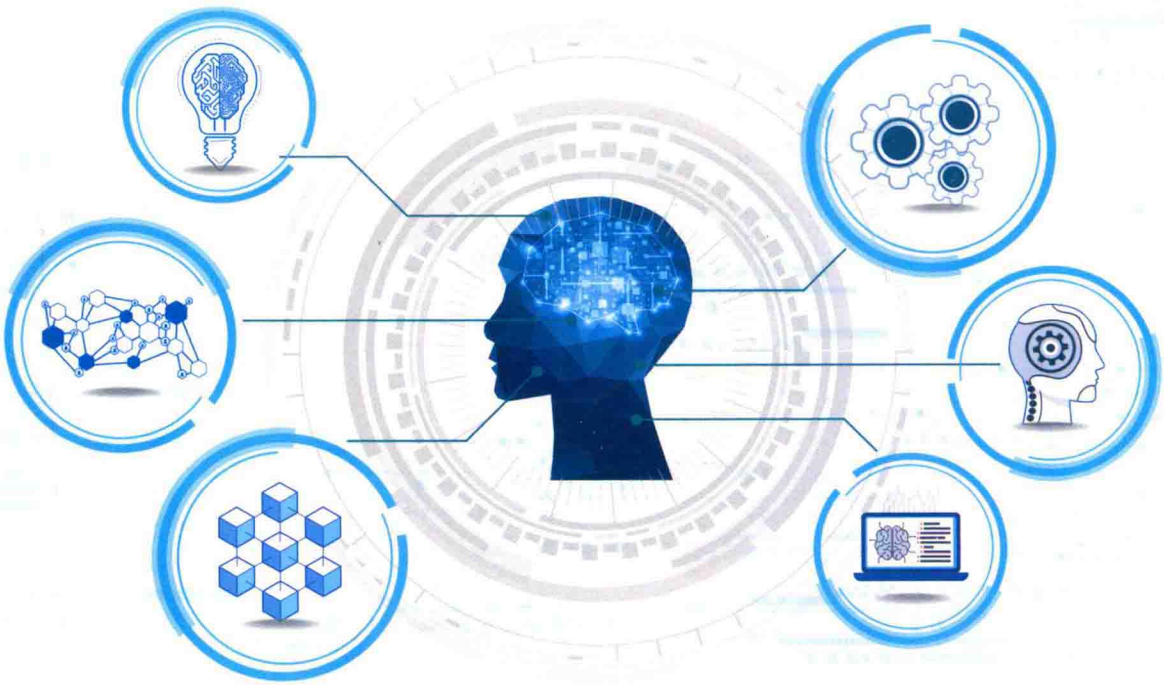
智能系统与技术丛书

Large Scale Machine Learning with Python

Python大规模机器学习

[荷] 巴斯蒂安·贾丁（Bastiaan Sjardin）
[意] 卢卡·马萨罗（Luca Massaron） 著
阿尔贝托·博斯凯蒂（Alberto Boschetti）

王贵财 刘春明 译



机械工业出版社
China Machine Press

Large Scale Machine Learning with Python

Python大规模机器学习

[荷] 巴斯蒂安·贾丁（Bastiaan Sjardin）
[意] 卢卡·马萨罗（Luca Massaron） 著
阿尔贝托·博斯凯蒂（Alberto Boschetti）

王贵财 刘春明 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 大规模机器学习 /（荷）巴斯蒂安·贾丁（Bastiaan Sjardin）等著；王贵财，刘春明译 . —北京：机械工业出版社，2019.2
（智能系统与技术丛书）
书名原文：Large Scale Machine Learning with Python

ISBN 978-7-111-62314-4

I. P… II. ①巴… ②王… ③刘… III. 软件工具－程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2019）第 053576 号

本书版权登记号：图字 01-2016-8642

Python 大规模机器学习

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：杨宴蕾 责任校对：殷 虹
印 刷：北京市荣盛彩色印刷有限公司 版 次：2019 年 5 月第 1 版第 1 次印刷
开 本：186mm×240mm 1/16 印 张：19.75
书 号：ISBN 978-7-111-62314-4 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）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 读者信箱：hzit@hzbook.com

内容简介

本书三位作者致力于人工智能和数据分析领域的工作，曾为世界各地的公司和政府机构构建数据科学和人工智能项目，本书是他们十余年实践经验的结晶。书中不仅介绍大规模机器学习的基本概念，还包含丰富的案例研究，全部内容均针对最实用的技术和工具，对理论细节不作深入讨论。不管是初学者、普通用户还是专家级用户，通过阅读本书都能了解和掌握如何利用Python进行大规模机器学习。

本书由浅入深讲解大量实例，图文并茂呈现每一步的操作结果，可帮助你更好地掌握大规模机器学习所需的Python技术，包括基于Scikit-learn可扩展学习、 Liblinear和Vowpal Wabbit快速支持向量机、基于Theano与H2O的大规模深度学习方法、TensorFlow深度学习技术与在线神经网络方法、大规模分类和回归树的可扩展解决方案、大规模无监督学习（PCA、聚类分析和主题建模等）扩展方法、Hadoop和Spark分布式环境、Spark机器学习实践以及Theano和GPU计算的基础知识等。

作者简介

巴斯蒂安·贾丁（Bastiaan Sjardin）是一位具有人工智能和数学背景的数据科学家和创始人。他获得莱顿大学和麻省理工（MIT）校园课程联合培养的认知科学硕士学位。过去五年，他从事过广泛的数据科学和人工智能项目。他擅长Python和R编程语言。目前，他是Quandbee（http://www.quandbee.com）的联合创始人，该公司主要提供大规模机器学习和人工智能应用。

卢卡·马萨罗（Luca Massaron）是一位数据科学家和市场研究总监，擅长多元统计分析、机器学习和客户洞察力研究，在解决实际问题和应用推理统计、统计、数据挖掘和算法为客户创造价值方面有十多年经验。

阿尔贝托·博斯凯蒂（Alberto Boschetti）是一位有信号处理和统计专业知识的数据科学家。他获得电信工程博士学位，目前在伦敦生活和工作。在其工作项目中，他应对过自然语言处理（NLP）、机器学习和分布式处理等多项挑战。

封面设计 锡 彬

华章IT
HZBOOKS | Information Technology

The Translator's Words 译 者 序

机器学习属于人工智能范畴。随着大数据的出现和适用性提高，尽管计算机硬件条件发生改善，但对机器学习算法效率的要求并未降低，对可扩展机器学习解决方案的需求呈指数增长，这使得人们仍然需要解决大多数学习算法扩展性较差、CPU 和内存过载等问题。然而扩展机器学习解决方案并不多，因此大数据既给大规模机器学习带来机遇，也带来挑战。

Python 是一种通用的编程语言，广泛用于科学计算、数据分析与人工智能领域。Python 具有高效、灵活、开源、功能丰富、可扩展性强、表达力强和较高的可移植性等特点，利用 Python 进行大规模机器学习不失为明智之举。

为此，本书不仅介绍大规模机器学习的基本概念，还包含丰富的案例研究。书中所选皆为最实用的技术和工具，而对理论细节未进行深入讨论，以便提供大规模机器学习方法（甚至非常规方法）。不管是初学者、普通用户还是专家级用户，通过本书都能理解并掌握如何利用 Python 进行大规模机器学习。为了让读者快速掌握核心技术，本书由浅入深讲解大量实例，图文并茂呈现每一步的操作结果，帮助读者更好地掌握大规模机器学习的 Python 工具。例如，基于 Scikit-learn 可扩展学习、Liblinear 和 Vowpal Wabbit 快速支持向量机、基于 Theano 与 H2O 的大规模深度学习方法、TensorFlow 深度学习技术与在线神经网络方法、大规模分类和回归树的可扩展解决方案、大规模无监督学习（PCA、聚类分析和主题建模等）扩展方法、Hadoop 和 Spark 分布式环境、Spark 机器学习实践以及 Theano 和 GPU 计算的基础知识。

本书作者致力于人工智能和数据分析领域的研究，为世界各地的公司和政府机构构建过数据科学和人工智能项目，积累了十多年的实践经验。在翻译过程中，我为作者对利用 Python 进行大规模机器学习的深入掌握和独到见解而深感惊讶，并由衷赞叹。同时，对我而言这也是一个学习与提高的过程。为做到专业词汇权威准确，内容忠实原书，译者查阅了大量资料，但因水平有限，加上时间仓促，错误和疏漏在所难免，恳请读者及时指出，以便再版时予以更正。翻译分工如下：中南大学地球科学与信息物理学院刘春明负责第 1 章以及附录，河南工业大学信息科学与工程学院王贵财负责其余章节。

本书的翻译工作得到湖南省自然科学基金资助项目（2015JJ2151）、河南省高校科技

创新团队支持计划“面向领域大数据的分布式计算技术”（17IRTSTHN011）、河南省高等学校重点科研项目资助计划（18A4300111）和河南工业大学校科研基金“青年支持计划”（2016QNH29）的资助。感谢参与本书资料整理的河南工业大学信息科学与工程学院郭浩、李欣欣、胡志明与李美玲等同学。特别感谢机械工业出版社编辑老师的帮助，他们的辛勤工作提高了本译著的质量。感谢家人对我的支持与鼓励，感谢儿子禾禾给予我的精神支持，让我对从事科技工作更加坚定和执着。

王贵财

2019 年 1 月

Preface 前　言

“拥有大脑的好处在于：一个人可以学习，无知可以变成有知，点滴知识可以逐渐汇聚成江海”

——Douglas Hofstadter

机器学习属于人工智能范畴，其目的是基于现有数据集（训练集）来寻找函数，以便以尽可能高的正确性预测先前未见过的数据集（测试集）的结果，这通常以标签和类别的形式（分类问题）或以连续值的形式（回归问题）出现。在实际应用中，机器学习的具体实例包括预测未来股票价格，或从一组文件中对作者性别进行分类，等等。本书介绍最重要的机器学习概念和适合更大数据集的方法，并通过 Python 的实际示例向读者进行讲解。主要讨论监督学习（分类和回归），以及适用于更大数据集的无监督学习，比如主成分分析（PCA）、聚类和主题建模。

谷歌、Facebook 和优步等大型 IT 公司都声称它们成功地大规模应用了这样的机器学习方法，从而引起世界轰动。随着大数据的出现和适用性提高，对可扩展机器学习解决方案的需求呈指数增长，导致许多其他公司甚至个人也已经开始渴望在大数据集中挖掘隐藏的相关性成果。不幸的是，大多数学习算法都不能很好扩展，会在台式计算机或较大的计算集群上导致 CPU 和内存过载。因此，即使大数据的炒作高峰已经过去，但可扩展机器学习解决方案并不充裕。

坦率地说，仍然需要解决许多瓶颈问题，即便是很难归类为大数据的数据集也如此（有的数据集高达 2GB 甚至更大）。本书的任务是提供合适的方法（有时甚至是非常规方法），以便大规模应用最强大的开源机器学习方法，而无须昂贵的企业解决方案或大型计算集群。通过本书，读者可以学习使用 Python 和其他一些可用的解决方案（这些方案与可扩展的机器学习流水线能很好地集成）。阅读这本书是一次旅程，它将让你对机器学习有一个全新的了解，从而为你开始真正的大数据分析奠定基础。

本书涵盖的内容

第 1 章以正确视角提出可扩展机器学习的问题，以便你熟悉本书中将要使用的工具。

第 2 章讨论采用随机梯度下降（SGD）策略减少内存消耗，它基于非核心学习的主题。另外演示各种数据的不同处理技术，例如散列技巧。

第 3 章介绍流算法，它能够以支持向量机的形式发现非线性。我们将介绍目前 Scikit-learn 的替代方法，如 LIBLINEAR 和 Vowpal Wabbit，虽然它们以外部 shell 命令运行，但很容易用 Python 脚本封装和定向。

第 4 章为在 Theano 框架中应用深度神经网络以及使用 H2O 进行大规模处理提供有用策略。尽管这是个热门话题，但成功应用它会相当困难，更别说提供可扩展的解决方案。另外，还将学习使用 theanets 包中的自动编码器实现无监督的预训练。

第 5 章介绍有趣的深度学习技术与在线神经网络方法。虽然 TensorFlow 还处于起步阶段，但该框架提供了非常不错的机器学习解决方案。此外，还将详解如何在 TensorFlow 环境中使用 Keras 卷积神经网络功能。

第 6 章详解随机森林、梯度增强和 XGboost 的可扩展解决方案。CART 是分类和回归树的缩写，它是一种通常应用于集成方法框架的机器学习方法。我们还将演示使用 H2O 的大规模应用实例。

第 7 章深入介绍无监督学习、PCA、聚类分析和主题建模方法，并使用正确方法对它们进行扩展。

第 8 章学习如何在虚拟机环境中设置 Spark，以便从单台机器转移到网络计算范例。Python 很容易在机器集群上集成并能增强我们的工作效率，因此很容易利用 Hadoop 集群的能力。

第 9 章演示使用 Spark 处理数据和在大数据集上构建预测模型的所有重要环节。

附录介绍 GPU 和 Theano，包括 Theano 和 GPU 计算的基础知识。如果你的系统允许，还将帮助读者学习相关安装和环境配置，以便在 GPU 上使用 Theano。

本书要求

运行书中代码示例需要在 macOS、Linux 或 Microsoft Windows 上安装 Python 2.7 或更高版本。

书中示例经常使用 Python 的基本功能库，例如 SciPy、NumPy、Scikit-learn 和 StatsModels，并且在某种程度上使用 matplotlib 和 pandas 进行科学和统计计算。也会使用称为 H2O 的非核心云计算应用程序。

本书需要 Jupyter 及其 Python 内核驱动的 Notebooks，本书使用最新版本 4.1。

第 1 章将为设置 Python 环境、核心库以及全部必需工具提供所有分步说明和某些技巧。

本书读者

本书适合数据科学从业者、开发人员以及计划使用大型复杂数据集的读者。我们努力让本书拥有尽可能好的可读性，以便适合更多读者。考虑到本书主题非常先进，我们建议读者先熟悉基本的机器学习概念，如分类和回归、误差最小化函数和交叉验证等，但不严格要求读者必须这样做。本书假设读者了解 Python、Jupyter Notebooks 和命令行运行，并有一定的数学基础，能够掌握书中的各种大型解决方案背后的概念。本书写作风格也适合使用其他语言（R、Java 和 MATLAB）的程序员。理想情况下，非常适合（但不限于）熟悉机器学习并有兴趣使用 Python 的数据科学家，因为相比于 R 或 MATLAB 而言，Python 在计算、内存和 I/O 方面有优势。

排版约定

书中代码块设置如下：

```
from sklearn import datasets
iris = datasets.load_iris()
```

大多数示例中使用 Jupyter Notebooks，所以希望在包含代码块的单元中始终带有输入（标记为 `In:`），并通常带有输出（标记为 `Out:`）。在你的计算机上，只需输入 `In:` 后面的代码，并检查结果是否与 `Out:` 后面的内容相对应：

```
In: clf.fit(X, y)
Out: SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
degree=3, gamma=0.0, kernel='rbf', max_iter=-1, probability=False,
random_state=None, shrinking=True, tol=0.001, verbose=False)
```

在终端命令行中给出命令时，会带有前缀 $>，否则，如果是 Python REPL，则以>>>开头：

```
$>python
>>> import sys
>>> print sys.version_info
```


表示警告或重要说明。


表示提示和技巧。

下载示例代码及彩色图像

本书的示例代码及所有截图和样图，可以从 http://www.packtpub.com 通过个人账号下载，也可以访问华章图书官网 http://www.hzbook.com，通过注册并登录个人账号下载。

还可以从 GitHub 获取本书代码：

https://github.com/PacktPublishing/Large-Scale-Machine-Learning-With-Python。

About the Authors 作者简介

Bastiaan Sjardin 是一位具有人工智能和数学背景的数据科学家和公司创始人。他获得莱顿大学和麻省理工学院（MIT）校园课程联合培养的认知科学硕士学位。在过去五年中，他从事过广泛的数据科学和人工智能项目。他是密歇根大学社会网络分析课程 Coursera 和约翰斯·霍普金斯大学机器学习实践课程的常客。他擅长 Python 和 R 编程语言。目前，他是 Quandbee（http://www.quandbee.com）的联合创始人，该公司主要提供大规模机器学习和人工智能应用。

Luca Massaron 是一位数据科学家和市场研究总监，擅长多元统计分析、机器学习和客户洞察力研究，在解决实际问题和应用推理、统计、数据挖掘和算法来为用户创造价值方面有十多年经验。从成为意大利网络观众分析的先驱，到跻身前十名的 Kaggler，他一直对数据分析充满热情，还向专业人士和普通大众展示数据驱动知识发现的潜力，相比不必要的复杂性，他更喜欢简洁。他相信仅仅通过基本操作就可以在数据科学中收获很多东西。

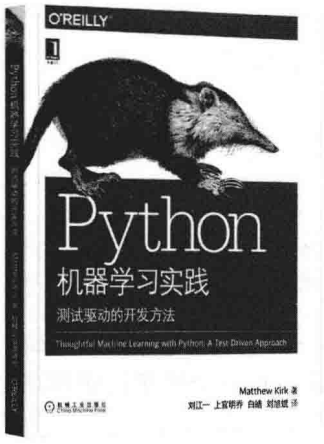
Alberto Boschetti 是一位具有信号处理和统计专业知识的数据科学家。他获得电信工程博士学位，目前在伦敦生活和工作。在其工作项目中，他面临过从自然语言处理（NLP）和机器学习到分布式处理的挑战。他在工作中充满热情，始终努力了解数据科学的最新发展，他喜欢参加聚会、会议和其他活动。

审校者简介 *About the Reviewers*

Oleg Okun 是一位机器学习专家，曾编辑出版四部著作、多篇期刊论文和会议论文。Oleg 有 25 年工作经历，其间，他曾在其祖国白俄罗斯和国外（芬兰、瑞典和德国）的学术界和工业界工作过。其工作经验包括文档图像分析、指纹生物识别、生物信息学、在线/离线营销分析和信用评分分析。他对分布式机器学习和物联网的各个方面都感兴趣。目前 Oleg 在德国汉堡生活和工作，即将担任智能系统的首席架构师。他擅长的编程语言是 Python、R 和 Scala。

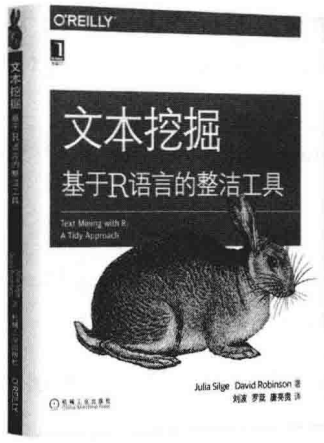
Kai Londenberg 是一位拥有多年专业经验的数据科学家和大数据专家。目前在大众汽车实验室担任数据科学家。在此之前，他有幸成为 Searchmetrics 公司的首席数据科学家，Luca Massaron 曾是他的团队成员。Kai 喜欢使用尖端技术，虽然他是一名务实的机器学习从业者和软件开发人员，但他总是乐于学习机器学习、人工智能和相关领域的最新技术和研究成果。https://www.linkedin.com/in/kailondenberg 是其 LinkedIn 个人网址。

推荐阅读



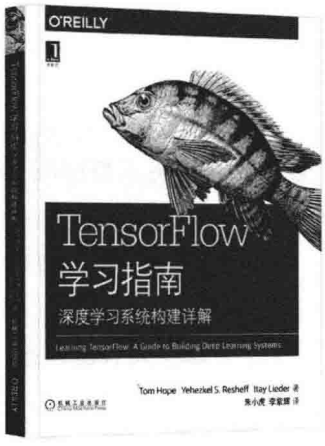
Python机器学习实践：测试驱动的开发方法

作者：Matthew Kirk ISBN：978-7-111-58166-6 定价：59.00元



文本挖掘：基于R语言的整洁工具

作者：Julia Silge , David Robinson ISBN：978-7-111-58855-9 定价：59.00元



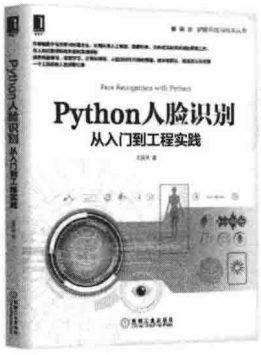
TensorFlow学习指南：深度学习系统构建详解

作者：Tom Hope, Yehezkel S. Resheff, Itay Lieder ISBN：978-7-111-60072-5 定价：69.00元

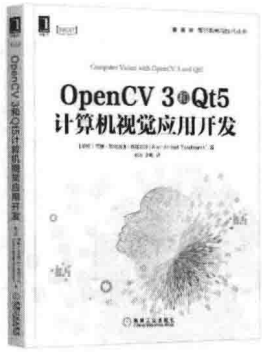
算法技术手册（原书第2版）

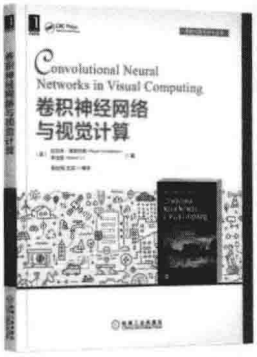
作者：George T. Heineman等 ISBN：978-7-111-56222-1 定价：89.00元

推荐阅读

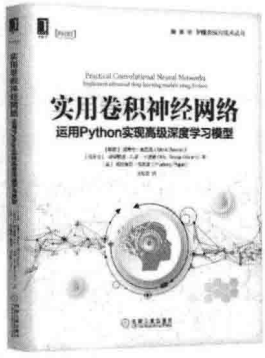

Python人脸识别
从入门到工程实践


机器学习
使用OpenCV和Python
进行智能图像处理


OpenCV 3和Qt5
计算机视觉应用开发

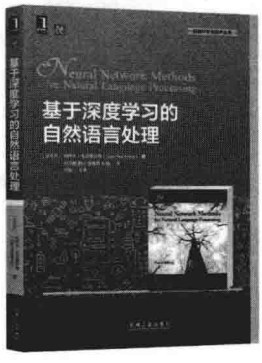

Convolutional Neural Networks in Visual Computing
卷积神经网络
与视觉计算


深度学习与计算机视觉

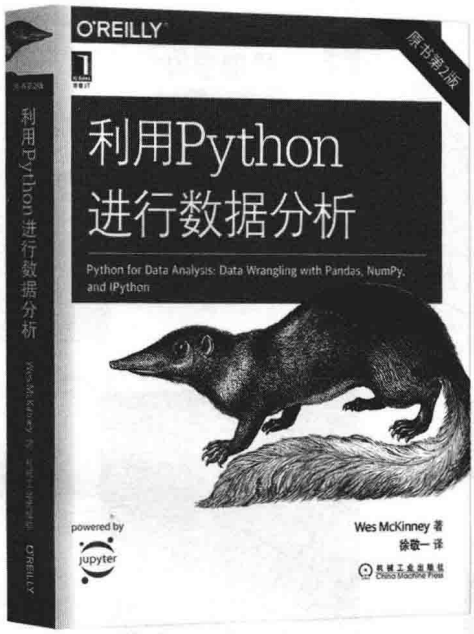

实用卷积神经网络
运用Python实现高级深度学习模型


面向自然语言处理的深度学习
用Python创建神经网络


Python自然语言处理实战
核心技术与算法


基于深度学习的
自然语言处理

推荐阅读



利用Python进行数据分析（原书第2版）

书号：978-7-111-60370-2 作者：Wes McKinney 定价：119.00元

Python数据分析经典畅销书全新升级，第1版中文版累计印刷10万册

Python pandas创始人亲自执笔，Python语言的核心开发人员鼎立推荐

针对Python 3.6进行全面修订和更新，涵盖新版的pandas、NumPy、IPython和Jupyter，并增加大量实际案例，可以帮助你高效解决一系列数据分析问题