异步图书
www.epubit.com

scikit-learn 机器学习 第2版

Mastering Machine Learning with scikit-learn
Second Edition

[美] 加文·海克（Gavin Hackeling）著　张浩然 译

Packt›

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

scikit-learn 机器学习 第2版

Mastering Machine Learning with scikit-learn
Second Edition

[美] 加文• 海克（Gavin Hackeling）著　张浩然 译

人民邮电出版社
北京

图书在版编目（CIP）数据

scikit-learn机器学习 : 第2版 / (美) 加文·海克 (Gavin Hackeling) 著 ; 张浩然译. -- 北京 : 人民邮电出版社, 2019.2
ISBN 978-7-115-50340-4

Ⅰ. ①s… Ⅱ. ①加… ②张… Ⅲ. ①机器学习 Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2018)第284538号

版权声明

◆ 著 [美] 加文·海克（Gavin Hackeling）
译 张浩然
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
涿州市京南印刷厂印刷

◆ 开本：800×1000 1/16
印张：13.5
字数：260 千字 2019 年 2 月第 1 版
印数：1－2 400 册 2019 年 2 月河北第 1 次印刷

著作权合同登记号 图字：01-2017-9190 号

定价：59.00 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：京东工商广登字 20170147 号

内容提要

近年来，Python 语言成为了广受欢迎的编程语言，而它在机器学习领域也有着卓越的表现。scikit-learn 是一个用 Python 语言编写的机器学习算法库，它可以实现一系列常用的机器学习算法，是一个不可多得的好工具。

本书通过 14 章内容，详细地介绍了一系列机器学习模型和 scikit-learn 的使用技巧。本书从机器学习的基础理论讲起，涵盖了简单线性回归、K-近邻算法、特征提取、多元线性回归、逻辑回归、朴素贝叶斯、非线性分类、决策树回归、随机森林、感知机、支持向量机、人工神经网络、K-均值算法、主成分分析等重要话题。

本书适合机器学习领域的工程师学习，也适合想要了解 scikit-learn 的数据科学家阅读。通过阅读本书，读者将有效提升自己在机器学习模型的构建和评估方面的能力，并能够高效地解决机器学习难题。

作者简介

加文·海克（Gavin Hackeling）是一名数据科学家和作家。他研究过各种各样的机器学习问题，包括自动语音识别、文档分类、目标识别以及语义切分。他毕业于北卡罗来纳大学和纽约大学，目前他和妻子以及小猫生活在布鲁克林。

感谢我的妻子 Hallie，以及 scikit-learn 社区。

审稿人简介

奥列格·奥肯（Oleg Okun）是一位机器学习专家，他还是4本书、许多期刊文章和会议论文的作者/编辑。他的职业生涯已经超过四分之一个世纪。他受雇于包括他的祖国（白罗斯）和国外（芬兰、瑞典和德国）的学术机构和企业。他的工作经验涉及文本图片分析、指纹生物技术、生物信息学、在线/离线市场分析、信用评估和文本分析领域。

他对分布式机器学习和物联网感兴趣，目前居住在德国汉堡市。

我想对父母为我做的一切表示最深切的感激。

前言

近些年来，机器学习已经成为大家热衷的话题。在机器学习领域，各式各样的应用层出不穷。其中的一些应用（例如垃圾邮件过滤器）已经被广泛使用，却反而因为太成功而变得平淡无奇。很多其他的应用直到近些年才纷纷出现，它们无一不在昭示着机器学习带来的无限可能。

在本书中，我们将分析一些机器学习模型和学习算法，讨论一些常用的机器学习任务，同时也会学习如何衡量机器学习系统的性能。我们将使用一个用 Python 编程语言编写的类库 scikit-learn，它包含了最新机器学习算法的实现，其 API 也很直观通用。

本书涵盖内容

第 1 章，机器学习基础。本章给出了机器学习的定义：机器学习是对如何通过从经验中学习来改善工作性能的研究和设计。该定义提纲挈领地引出了后续的章节，在后续的每个章节中，我们都将分析一种机器学习模型，将其运用于现实工作中，并衡量其性能。

第 2 章，简单线性回归。本章讨论了将单个特征同连续响应变量联系起来的模型。我们将学习代价函数，以及使用范式函数优化模型的相关知识。

第 3 章，用 K-近邻算法分类和回归。本章介绍了一个用于分类和回归任务的简单的非线性模型。

第 4 章，特征提取。本章介绍了将文本、图片以及分类变量表示为机器学习模型可用特征的技术。

第 5 章，从简单线性回归到多元线性回归。本章讨论了简单线性回归模型的扩展——多元线性回归模型，它能在多个特征上对连续响应变量进行回归。

第 6 章，从线性回归到逻辑回归。本章将多元线性回归模型做了进一步推广，并介绍了一个用于二元分类任务的模型。

第 7 章，朴素贝叶斯。本章讨论了贝叶斯定理和朴素贝叶斯分类器，同时对生成模型和判别模型进行了对比。

第 8 章，非线性分类和决策树回归。本章介绍了决策树这种用于分类和回归任务的简单模型。

第 9 章，集成方法：从决策树到随机森林。本章讨论了 3 种用于合并模型的方法，它们分别是套袋法（bagging）、推进法（boosting）和堆叠法（stacking）。

第 10 章，感知机。本章内容介绍了一种用于二元分类的简单在线模型。

第 11 章，从感知机到支持向量机。本章讨论了一种可用于分类和回归的强大的判别模型——支持向量机，同时还介绍了一种能有效将特性投影到高维度空间的技巧。

第 12 章，从感知机到人工神经网络。本章介绍了一种建立在人工神经元图结构基础上，用于分类和回归任务的强大的非线性模型。

第 13 章，K-均值算法。本章讨论了一种在无标记数据中发现结构的算法。

第 14 章，使用主成分分析降维。本章讨论了一种用于降低数据维度以缓和维度灾难的方法。

准备工作

运行本书中的例子需要 Python 版本 2.7 或者 3.3，以及 pip—PyPA 工作组推荐使用的 Python 包安装工具。书中的例子预期在 Jupyter notebook 环境中或者 IPython 解释器环境中运行。第 1 章详细说明了如何在 Ubuntu、MacOS 和 Windows 环境下安装 scikit-learn 0.18.1 版本类库及其依赖项目和其他类库。

目标读者

本书的目标读者是希望了解机器学习算法是如何运行，想培养机器学习使用直觉的软

件工程师。本书的目标读者也包含希望了解 scikit-learn 类库 API 的数据科学家。读者不需要熟悉机器学习基础和 Python 编程语言，但具备相关基础对阅读本书很有帮助。

排版约定

在本书中，你会发现一些不同的文本样式，用以区别不同种类的信息。下面对一些样式及其意义举例进行说明。

代码片段、数据库表名、目录名、文件名、文件扩展名、路径名、URL、用户输入、以及推特用户名会如下印刷：“由于 scikit-learn 不是一个有效的 Python 包名称，该类库被命名为 sklearn”。

```
# In[1]:
import sklearn
sklearn.__version__

# Out[1]:
'0.18.1'
```

新术语和**重要语句**会加粗印刷。

这个图标表示警告或需要特别注意的内容。

这个图标表示提示或者技巧。

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

配套资源

本书提供配套源代码，要获得该配套资源，请在异步社区本书页面中单击 配套资源，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

如果您是教师，希望获得教学配套资源，请在社区本书页面中直接联系本书的责任编辑。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，单击“提交勘误”，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。

扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“**异步社区**”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“**异步图书**”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

目录

第 1 章
机器学习基础

在本章中，我们将回顾机器学习中的基础概念，比较监督学习和无监督学习，讨论训练数据、测试数据和验证数据的用法，并了解机器学习应用。最后，我们将介绍 scikit-learn 库，并安装后续章节中需要的工具。

1.1 定义机器学习

长久以来，我们的想象力一直被那些能够学习和模仿人类智慧的机器所吸引。尽管具有一般人工智能的机器（比如阿瑟·克拉克笔下的 HAL 和艾萨克·阿西莫夫笔下的 Sonny）仍然没有实现，但是能够从经验中获取新知识和新技能的软件正在变得越来越普遍。我们使用这些机器学习程序去寻找自己可能喜欢的新音乐，找到自己真正想在网上购买的鞋子。机器学习程序允许我们对智能手机下达命令，并允许用恒温控制器自动设置温度。机器学习程序可以比人类更好地破译书写凌乱的邮寄地址，并更加警觉地防止信用卡欺诈。从研发新药到估计一个头条新闻的页面访问量，机器学习软件正在成为许多行业的核心部分。机器学习甚至已经侵占了许多长久以来一直被认为只有人类能涉及的领域，例如撰写一篇关于杜克大学篮球队输给了北卡大学篮球队的体育专栏报道。

机器学习是对软件工件的设计和学习，它使用过去的经验去指导未来的决策。机器学习是对从数据中学习的软件的研究。机器学习的基础目标是归纳，或者从一种未知规则的应用例子中归纳出未知规则。机器学习的典型例子是垃圾邮件过滤。通过观察已经被标记为垃圾邮件或非垃圾邮件的电子邮件，垃圾邮件过滤器可以分类新消息。研究人工智能的先锋科学家亚瑟·萨缪尔曾说过机器学习是“给予计算机学习的能力而无须显式地编程的研究”。在 20 世纪 50 年代到 20 世纪 60 年代之间，萨缪尔开发了多个下棋程序。虽然下棋的规则很简单，但是要战胜技艺高超的对手需要复杂的策略。萨缪尔从来没有显式地编程过这些策略，