O'REILLY®

TURING 图灵程序设计丛书

Python机器学习基础教程

Introduction to Machine Learning with Python

以机器学习算法实践为重点，使用scikit-learn库从头构建机器学习应用

[德] Andreas C. Müller [美] Sarah Guido 著

张亮 (hysic) 译

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机器学习基础教程

Introduction to Machine Learning with Python

[德] Andreas C. Müller [美] Sarah Guido 著
张亮 (hysic) 译

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®
O'Reilly Media, Inc.授权人民邮电出版社出版

图书在版编目（CIP）数据

Python机器学习基础教程 / (德) 安德里亚斯·穆勒 (Andreas C. Müller), (美) 莎拉·吉多 (Sarah Guido) 著 ; 张亮 (hysic) 译. -- 北京 : 人民邮电出版社, 2018.1
(图灵程序设计丛书)
ISBN 978-7-115-47561-9

Ⅰ. ①P… Ⅱ. ①安… ②莎… ③张… 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314601号

内 容 提 要

本书是机器学习入门书，以 Python 语言介绍。主要内容包括：机器学习的基本概念及其应用；实践中最常用的机器学习算法以及这些算法的优缺点；在机器学习中待处理数据的呈现方式的重要性，以及应重点关注数据的哪些方面；模型评估和调参的高级方法，重点讲解交叉验证和网格搜索；管道的概念；如何将前面各章的方法应用到文本数据上，还介绍了一些文本特有的处理方法。

本书适合机器学习从业者或有志成为机器学习从业者的人阅读。

◆ 著 [德] Andreas C. Müller [美] Sarah Guido
译 张 亮 (hysic)
责任编辑 岳新欣
执行编辑 李 敏
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
三河市书文印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：18.75
字数：443千字 2018年1月第1版
印数：1-4 000册 2018年1月河北第1次印刷
著作权合同登记号 图字：01-2017-8626号

定价：79.00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

O'Reilly Media, Inc.介绍

O'Reilly Media通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O'Reilly一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了*Make*杂志，从而成为DIY革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly现在还将先锋专家的知识传递给普通的计算机用户。无论是通过图书出版、在线服务或者面授课程，每一项O'Reilly的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar博客有口皆碑。”

——*Wired*

“O'Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——*Business 2.0*

“O'Reilly Conference是聚集关键思想领袖的绝对典范。”

——*CRN*

“一本O'Reilly的书就代表一个有用、有前途、需要学习的主题。”

——*Irish Times*

“Tim是位特立独行的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去，Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——*Linux Journal*

前言

目前，从医疗诊断和治疗到在社交网络上寻找好友，许多商业应用和研究项目都离不开机器学习。许多人以为，只有大公司的大型研究团队才能用到机器学习。在本书中，我们要向你展示，自己动手构建机器学习解决方案是多么容易的一件事，也将介绍如何将这件事做到最好。学完本书中的知识，你可以自己构建系统，研究 Twitter 用户的情感，或者对全球变暖做出预测。机器学习的应用十分广泛，如今的海量数据使得其应用范围更是远超人们的想象。

目标读者

本书是为机器学习从业者或有志成为机器学习从业者的人准备的，他们在为现实生活中的机器学习问题寻找解决方案。这是一本入门书，不需要读者具备机器学习或人工智能（artificial intelligence，AI）的相关知识。我们主要使用 Python 和 `scikit-learn` 库，一步一步构建一个有效的机器学习应用。我们介绍的方法适用于科学家和研究人员，也会对开发商业应用的数据科学家有所帮助。如果你对 Python 以及 NumPy 和 `matplotlib` 库有所了解的话，将能够更好地掌握本书的内容。

我们刻意不将数学作为重点，而是将机器学习算法的实践作为重点。数学（尤其是概率论）是机器学习算法的基石，所以我们不会详细分析算法的细节。如果你对机器学习算法的数学部分感兴趣，我们推荐阅读 Trevor Hastie、Robert Tibshirani 和 Jerome Friedman 合著的《统计学习基础》（*Elements of Statistical Learning*，Springer 出版社）一书，可以在几位作者的网站上免费阅读这本书（http://statweb.stanford.edu/~tibs/ElemStatLearn/）。我们也不会从头讲解如何编写机器学习算法，而是将重点放在如何应用 `scikit-learn` 库和其他库中已经实现的海量模型。

写作本书的原因

市面上已经有许多关于机器学习和 AI 的书了，但这些书都是为计算机科学专业的研究生或博士生准备的，里面全都是高等数学的内容。与之形成鲜明对比的是，在研究领域和商业应用中，机器学习是作为一般工具使用的。如今，使用机器学习并不需要拥有博士学

位。然而，能够完全涵盖在实践中实现机器学习算法的所有重要内容，而又不需要先修高等数学课程，这样的学习资源少之又少。对于那些想要使用机器学习算法而又不想花费大量时间研读微积分、线性代数和概率论的人来说，我们希望本书能够有所帮助。

本书概览

本书的结构大致如下。

- 第 1 章介绍机器学习的基本概念及其应用，并给出本书会用到的基本设置。
- 第 2 章和第 3 章介绍实践中最常用的机器学习算法，并讨论这些算法的优缺点。
- 第 4 章介绍在机器学习中待处理数据的呈现方式的重要性，以及应重点关注数据的哪些方面。
- 第 5 章介绍模型评估和调参的高级方法，重点讲解交叉验证和网格搜索。
- 第 6 章解释管道的概念。管道用于串联多个模型并封装工作流。
- 第 7 章介绍如何将前面各章讲述的方法应用到文本数据上，还介绍了一些文本特有的处理方法。
- 第 8 章对全书进行总结，还介绍了有关更高级主题的参考资料。

虽然第 2 章和第 3 章给出了实际算法，但对于初学者来说，并不需要理解所有这些算法。如果你想要尽快构建一个机器学习系统，我们建议你首先阅读第 1 章和第 2 章的开始部分，里面介绍了所有的核心概念。然后你可以翻到 2.5 节，里面提到了我们介绍的所有监督学习模型。从中选择最适合你需求的模型，然后翻回到对应小节阅读其详细内容。之后你可以使用第 5 章中的方法对你的模型进行评估和调参。

在线资源

在学习本书时，一定要参考 `scikit-learn` 官方网站（http://scikit-learn.org），查阅关于类和函数的更详细的文档，以及很多示例。此外，Andreas Müller 创建的视频课程“scikit-learn 高等机器学习”（Advanced Machine Learning with scikit-learn）可以作为本书的补充材料。你可以在 http://shop.oreilly.com/product/0636920043836.do 观看该课程。

排版约定

本书使用了下列排版约定。

- **黑体**
  表示新术语或重点强调的内容。
- 等宽字体（`constant width`）
  表示程序片段，以及正文中出现的变量、函数名、数据库、数据类型、环境变量、语句和关键字等。也用于表示命令、模块和包的名称。
- 加粗等宽字体（**`constant width bold`**）
  表示需要用户逐字输入的命令或其他文本。

- 等宽斜体（*constant width italic*）
  表示应替换成用户输入的值，或替换成根据上下文确定的值。

该图标表示提示或建议。

该图标表示一般性说明。

该图标表示警告或警示。

使用代码示例

补充材料（代码示例、IPython notebook 等）可以在 https://github.com/amueller/introduction_to_ml_with_python 下载。

本书是要帮你完成工作的。一般来说，如果本书提供了示例代码，你可以把它用在你的程序和文档中。除非你使用了很大一部分代码，否则无需联系我们获得许可。比如，用本书的几个代码片段写一个程序就无需获得许可。销售或分发 O'Reilly 图书的示例光盘则需要获得许可。引用本书中的示例代码来回答问题无需获得许可。将书中大量示例代码放到你的产品文档中则需要获得许可。

如果你在引用本书内容时注明出处，我们将不胜感激，但这并非强制要求。引用说明一般包括书名、作者、出版社和 ISBN。比如："*An Introduction to Machine Learning with Python* by Andreas C. Müller and Sarah Guido (O'Reilly). Copyright 2017 Sarah Guido and Andreas Müller, 978-1-449-36941-5."

如果你认为自己对代码示例的用法超出了合理使用的范围或上述许可的范围，敬请通过 permissions@oreilly.com 与我们联系。

Safari® Books Online

Safari Books Online 是应运而生的数字图书馆，它同时以图书和视频的形式出版世界顶级技术和商业作家的专业作品。

技术专家、软件开发人员、Web 设计师、商务人士和创意专家等，在开展调研、解决问题、学习和认证培训时，都将 Safari Books Online 视作获取资料的首选渠道。

对于企业、政府、教育机构和个人，Safari Books Online 都提供各种产品组合和灵活的定价策略。

用户可通过一个功能完备的数据库检索系统访问 O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology 等数百家出版社的上千种图书、培训视频和正式出版之前的书稿。要了解 Safari Books Online 的更多信息，请访问我们的网站。

联系我们

请把对本书的评价和问题发给出版社。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

我们为本书创建了一个网页，在上面列出了本书的勘误表、示例以及其他信息。本书的网站地址是：http://shop.oreilly.com/product/0636920030515.do。

如果你想就本书发表评论或技术性问题，请发送电子邮件到 bookquestions@oreilly.com。

想了解更多 O'Reilly 图书、培训课程、会议和新闻的信息，请访问我们的网站：
http://www.oreilly.com。

我们在 Facebook 的地址如下：http://facebook.com/oreilly

请关注我们的 Twitter 动态：http://twitter.com/oreillymedia

我们的 YouTube 视频地址如下：http://www.youtube.com/oreillymedia

致谢

来自Andreas的致谢

如果没有许多人的帮助和支持，本书永远不会出版。

我要感谢本书编辑 Meghan Blanchette 和 Brian MacDonald，特别是 Dawn Schanafelt，感谢

他们帮助我和 Sarah 共同完成这本书。

我要感谢我的审稿人 Thomas Caswell、Olivier Grisel、Stefan van der Walt 和 John Myles White，感谢他们花费时间阅读本书的早期版本，并提供宝贵的反馈意见。这些意见也成为了科学计算开源生态系统的基石。

我永远感谢热情的 Python 科学计算开源社区，特别要感谢 scikit-learn 的贡献者们。如果没有这个社区的支持和帮助，特别是 Gael Varoquaux、Alex Gramfort 和 Olivier Grisel 的支持和帮助，我永远无法成为 scikit-learn 的核心贡献者，也无法像现在这样对这个包有如此深刻的理解。我还要感谢 scikit-learn 的其他所有贡献者，他们花费了大量时间改进并维护这个包。

我还要感谢与我讨论的许多同事和同行。这些谈话帮助我理解了机器学习的挑战，并让我产生构思一本教科书的想法。我与许多人讨论过机器学习，但我要特别感谢其中的 Brian McFee、Daniela Huttenkoppen、Joel Nothman、Gilles Louppe、Hugo Bowne-Anderson、Sven Kreis、Alice Zheng、Kyunghyun Cho、Pablo Baberas 和 Dan Cervone。

我还要感谢 Rachel Rakov，她对本书的早期版本做了许多热心的测试和校对工作，在成书过程中给了我很多帮助。

就个人来说，我要感谢我的父母 Harald 和 Margot，还有我的姐姐 Miriam，感谢他们持续给予我的支持和鼓励。我还要感谢生命中的许多人，他们的爱和友谊给我能量，支持我完成这项富有挑战性的任务。

来自Sarah的致谢

我要感谢 Meghan Blanchette，没有她的帮助和指导，甚至就不会有本项目的存在。感谢 Celia La 和 Brian Carlson 早期对本书的审阅。感谢 O'Reilly 工作人员无尽的耐心。最后，感谢 DTS，感谢你永恒不变的支持。

电子书

扫描如下二维码，即可购买本书电子版。

目录

第1章

引言

机器学习（machine learning）是从数据中提取知识。它是统计学、人工智能和计算机科学交叉的研究领域，也被称为预测分析（predictive analytics）或统计学习（statistical learning）。近年来，机器学习方法已经应用到日常生活的方方面面。从自动推荐看什么电影、点什么食物、买什么商品，到个性化的在线电台和从照片中识别好友，许多现代化网站和设备的核心都是机器学习算法。当你访问像 Facebook、Amazon 或 Netflix 这样的复杂网站时，很可能网站的每一部分都包含多种机器学习模型。

除了商业应用之外，机器学习也对当前数据驱动的研究方法产生了很大影响。本书中介绍的工具均已应用在各种科学问题上，比如研究恒星、寻找遥远的行星、发现新粒子、分析 DNA 序列，以及提供个性化的癌症治疗方案。

不过，如果想受益于机器学习算法，你的应用无需像上面那些例子那样给世界带来重大改变，数据量也用不着那么大。本章将解释机器学习如此流行的原因，并探讨机器学习可以解决哪些类型的问题。然后将向你展示如何构建第一个机器学习模型，同时介绍一些重要的概念。

1.1 为何选择机器学习

在“智能”应用的早期，许多系统使用人为制订的“if”和“else”决策规则来处理数据，或根据用户输入的内容进行调整。想象有一个垃圾邮件过滤器，其任务是酌情将收到的某些邮件移动到垃圾邮件文件夹。你可以创建一个关键词黑名单，所有包含这些关键词的邮件都会被标记为垃圾邮件。这是用专家设计的规则体系来设计“智能”应用的一个示例。人为制订的决策规则对某些应用来说是可行的，特别是人们对其模型处理过程非常熟悉的应用。但是，人为制订决策规则主要有两个缺点。

- 做决策所需要的逻辑只适用于单一领域和单项任务。任务哪怕稍有变化，都可能需要重写整个系统。
- 想要制订规则，需要对人类专家的决策过程有很深刻的理解。

这种人为制订规则的方法并不适用的一个例子就是图像中的人脸检测。如今，每台智能手机都能够检测到图像中的人脸。但直到 2001 年，人脸检测问题才得到解决。其主要问题在于，计算机“感知”像素（像素组成了计算机中的图像）的方式与人类感知面部的方式有非常大的不同。正是由于这种表征差异，人类想要制订出一套好的规则来描述数字图像中的人脸构成，基本上是不可能的。

但有了机器学习算法，仅向程序输入海量人脸图像，就足以让算法确定识别人脸需要哪些特征。

1.1.1 机器学习能够解决的问题

最成功的机器学习算法是能够将决策过程自动化的那些算法，这些决策过程是从已知示例中泛化得出的。在这种叫作**监督学习**（supervised learning）的方法中，用户将成对的输入和预期输出提供给算法，算法会找到一种方法，根据给定输入给出预期输出。尤其是在没有人类帮助的情况下，给定前所未见的输入，算法也能够给出相应的输出。回到前面垃圾邮件分类的例子，利用机器学习算法，用户为算法提供大量电子邮件（作为输入），以及这些邮件是否为垃圾邮件的信息（作为预期输出）。给定一封新邮件，算法就能够预测它是否为垃圾邮件。

从输入 / 输出对中进行学习的机器学习算法叫作**监督学习算法**（supervised learning algorithm），因为每个用于算法学习的样例都对应一个预期输出，好像有一个“老师”在监督着算法。虽然创建一个包含输入和输出的数据集往往费时又费力，但监督学习算法很好理解，其性能也易于测量。如果你的应用可以表示成一个监督学习问题，并且你能够创建包含预期输出的数据集，那么机器学习很可能可以解决你的问题。

监督机器学习任务的示例如下。

识别信封上手写的邮政编码

这里的输入是扫描的手写数字，预期输出是邮政编码中的实际数字。想要创建用于构建机器学习模型的数据集，你需要收集许多信封。然后你可以自己阅读邮政编码，将数字保存为预期输出。

基于医学影像判断肿瘤是否为良性

这里的输入是影像，输出是肿瘤是否为良性。想要创建用于构建模型的数据集，你需要一个医学影像数据库。你还需要咨询专家的意见，因此医生需要查看所有影像，然后判断哪些肿瘤是良性的，哪些不是良性的。除了影像内容之外，甚至可能还需要做额外的诊断来判断影像中的肿瘤是否为癌变。

检测信用卡交易中的诈骗行为

这里的输入是信用卡交易记录，输出是该交易记录是否可能为诈骗。假设你是信用卡的发行单位，收集数据集意味着需要保存所有的交易，并记录用户是否上报过任何诈骗交易。

在这些例子中需要注意一个有趣的现象，就是虽然输入和输出看起来相当简单，但三个例子中的数据收集过程却大不相同。阅读信封虽然很辛苦，却非常简单，也不用花多少钱。与之相反，获取医学影像和诊断不仅需要昂贵的设备，还需要稀有又昂贵的专家知识，更不要说伦理问题和隐私问题了。在检测信用卡诈骗的例子中，收集数据要容易得多。你的顾客会上报诈骗行为，从而为你提供预期输出。要获取所有欺诈行为和非欺诈行为的输入/输出对，你只需等待即可。

本书会讲到的另一类算法是**无监督学习算法**（unsupervised learning algorithm）。在无监督学习中，只有输入数据是已知的，没有为算法提供输出数据。虽然这种算法有许多成功的应用，但理解和评估这些算法往往更加困难。

无监督学习的示例如下。

确定一系列博客文章的主题

如果你有许多文本数据，可能想对其进行汇总，并找到其中共同的主题。事先你可能并不知道都有哪些主题，或者可能有多少个主题。所以输出是未知的。

将客户分成具有相似偏好的群组

给定一组客户记录，你可能想要找出哪些客户比较相似，并判断能否根据相似偏好对这些客户进行分组。对于一家购物网站来说，客户分组可能是“父母”“书虫”或“游戏玩家”。由于你事先并不知道可能有哪些分组，甚至不知道有多少组，所以并不知道输出是什么。

检测网站的异常访问模式

想要识别网站的滥用或 bug，找到异常的访问模式往往是很有用的。每种异常访问模式都互不相同，而且你可能没有任何记录在案的异常行为示例。在这个例子中你只是观察流量，并不知道什么是正常访问行为和异常访问行为，所以这是一个无监督学习问题。

无论是监督学习任务还是无监督学习任务，将输入数据表征为计算机可以理解的形式都是十分重要的。通常来说，将数据想象成表格是很有用的。你想要处理的每一个数据点（每一封电子邮件、每一名客户、每一次交易）对应表格中的一行，描述该数据点的每一项属性（比如客户年龄、交易金额或交易地点）对应表格中的一列。你可能会从年龄、性别、账号创建时间、在你的购物网站上的购买频率等方面来描述用户。你可能会用每一个像素的灰度值来描述肿瘤图像，也可能利用肿瘤的大小、形状和颜色进行描述。

在机器学习中，这里的每个实体或每一行被称为一个**样本**（sample）或数据点，而每一列（用来描述这些实体的属性）则被称为**特征**（feature）。

本书后面会更详细地介绍如何构建良好的数据表征，这被称为**特征提取**（feature extraction）或**特征工程**（feature engineering）。但你应该记住，如果没有数据信息的话，所有机器学习算法都无法做出预测。举个例子，如果你只有病人的姓氏这一个特征，那么任何算法都无法预测其性别。这一信息并未包含在数据中。如果你添加另一个特征，里面包含病人的名字，那么你预测正确的可能性就会变大，因为通过一个人的名字往往可以判断其性别。