计算机科学先进技术译丛

MANNING

实用机器学习

Real-World Machine Learning

（美）亨里克·布林克（Henrik Brink）
（美）约瑟夫 W.理查兹（Joseph W. Richards）
（美）马克·弗特罗夫（Mark Fetherolf）
著

程继洪　孙玉梅　潘佩佩　译

从实战角度剖析机器学习的知识原理
无需担心遇到难懂的数学公式和推导
含3个机器学习真实应用的综合案例

在线提供源代码

机械工业出版社
CHINA MACHINE PRESS

计算机科学先进技术译丛

实用机器学习

Real-World Machine Learning

[美] 亨里克·布林克（Henrik Brink）
[美] 约瑟夫 W. 理查兹（Joseph W. Richards） 著
[美] 马克·弗特罗夫（Mark Fetherolf）

程继洪 孙玉梅 潘佩佩 译

机 械 工 业 出 版 社

本书介绍了实用机器学习的工作流程，主要从实用角度进行了描述，没有数学公式和推导。本书涵盖了数据收集与处理、模型构建、评价和优化、特征的识别、提取和选择技术、高级特征工程、数据可视化技术以及模型的部署和安装，结合3个真实案例全面、详细地介绍了整个机器学习流程。最后，还介绍了机器学习流程的扩展和大数据应用。

本书可以作为程序员、数据分析师、统计学家、数据科学家解决实际问题的参考书，也可以作为机器学习爱好者学习和应用的参考书，还可以作为非专业学生的机器学习入门参考书，以及专业学生的实践参考书。

北京市版权局著作权合同登记　图字 01-2016-9106

图书在版编目(CIP)数据

实用机器学习/(美)亨里克·布林克(Henrik Brink)著;程继洪,孙玉梅,潘佩佩译. —北京:机械工业出版社,2017.6
(计算机科学先进技术译丛)
书名原文:Real - world Machine Learning
ISBN 978-7-111-56922-0

Ⅰ. ①实… Ⅱ. ①亨… ②程… ③孙… ④潘… Ⅲ. ①机器学习
Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2017)第 094007 号

机械工业出版社(北京市百万庄大街 22 号　邮政编码　100037)
策划编辑:丁　诚　　责任编辑:丁　诚　陈瑞文
责任校对:张艳霞　　责任印制:李　昂
三河市宏达印刷有限公司印刷
2017 年 6 月第 1 版·第 1 次印刷
184 mm×260 mm·14 印张·1 插页·339 千字
0001-3000 册
标准书号:ISBN 978-7-111-56922-0
定价:69.00 元

凡购本书,如有缺页、倒页、脱页,由本社发行部调换

电话服务	网络服务
服务咨询热线:010-88361066	机 工 官 网:www. cmpbook. com
读者购书热线:010-68326294	机 工 官 博:weibo. com/cmp1952
010-88379203	教育服务网:www. cmpedu. com
封面无防伪标均为盗版	金　书　网:www. golden - book. com

推荐序

在过去的几年里，机器学习已成为一个很大的产业：公司用它赚钱，对它的应用研究遍布工业领域和高校，好奇的开发人员到处寻找能提高他们机器学习技巧的方法。但是这种新发现的学习需求已经大大超过了市场上能够提供的学习这些技术如何在实际中应用的好方法，而本书恰好填补了这一空白。

应用机器学习包含对等的数学原理和技巧，也就是说，它是一门真正的学问。特别专注于任何一个方面都不是好办法，掌握平衡是非常必要的。

人们在相当长的时间内认为学习机器学习最好的方式，而且是唯一的方式，是在统计学和优化技术两个方面（很大程度上是相互独立的）追求更高的学位。主要是学习核心算法，包含它们的理论基础和边界以及该领域的特征域问题。相应地，同样有价值的经验积累在非官方渠道传递：会场外走廊上，研究实验室的团队智慧和数据处理手稿在成员之间传递。这种口头传递的知识往往有助于工作的完成，包括：针对情况进行算法选择，每一步数据如何打磨以及如何把各个部分组成一个有机整体。

如今我们生活在一个开源的时代，GitHub 上有大多数机器学习算法的高质量实现方法，还有综合的、结构良好的、能组合各个部分的框架可以使用。但在这繁荣的背后，这些口头传递的知识还是为大多数人所不知的。本书作者提供了这项服务，把这些知识收集在一起。这正是机器学习从深奥的学科转向工程应用技能所缺失的重要环节。

另一点需要强调的是，现在大多数广泛使用的机器学习算法并不完美，我们需要设计更完善的解决方案。这些算法对接收的数据很挑剔，如果目标不明确，则算法可能给出过分自信的预测。输入的细微改变会导致学习模型巨大的、莫名其妙的变化。它们的结果难以解释且难以做进一步处理。现代机器学习可被看作管理和减轻底层优化与统计学习方法的一个应用。

本书正是按照读者所面临的现实而组织的。首先，在深入研究这个流程应用于实际之前，描述了机器学习典型的工作流程。通过阅读你会发现一些方程（别处也可看到，主要包含在本领域的经典文档中），更重要的是如何将机器学习应用到实际项目和解决方案中。

作为宝贵的数学和形式化知识的有益补充，现在是学习这本书的时候了。它是许多老手希望在回头时看到的至关重要的另一本书。

—Beau Cronin
Head of Data，21 Inc.
Berkeley，CA

作者序

作为物理学和天文学专业的学生，我的大部分时间都用于处理测量和模拟数据，对数据进行分析、可视化和建模，以期得到具有科学价值的数据。作为程序员，我很快就把程序设计的技巧用于数据处理。第一次接触机器学习时，对我来说，它不仅是数据处理的有力的潜在工具，还是我感兴趣的两个领域的完美结合——数据科学和程序设计。

机器学习成为我在物理科学研究中的一个重要组成部分，并把我带到加州大学伯克利分校天文系，那里的统计人员、物理学家和计算机科学家们正在一起努力，把机器学习作为理解宇宙的一个越来越重要的工具。

在时域信息中心，我遇到了约瑟夫 W. 理查兹，他是一个统计学家，而且还是本书的作者之一。我们了解到，不但可以利用数据科学和机器学习技术进行科学研究，非学术界的公司和企业对此也越来越感兴趣。我们与其他 3 位同事共同创立了 Wise. io 来抓住这一机会。

在过去的 4 年里，Wise. io 已经与数不清的公司通过机器学习对他们的数据处理流程进行优化、提高和实现自动化。我们已经构建了大规模机器学习应用平台，每月为客户进行几亿次预测。我们了解到实用领域的数据仍是比较混乱的，这使得我们十分吃惊。我们希望把实用数据处理和构建下一代智能机器学习软件的知识传授给读者。

马克·弗特罗夫，本书的第 3 位作者，是多个创业公司的创始人兼 CTO（Chief Technology Officer，首席技术官），他在传统的统计和定量方法的基础上，进行系统管理和业务分析。在石油化工精炼过程的测试和优化中，他和他的团队认识到应用在生产过程中的技术可用于提高数据库、计算机系统和网络的性能。他们的分布式系统管理技术已嵌入到系统管理领先的产品中，后续将应用于电信和客户交互管理系统的管理和优化。

几年后，他沉迷于 Kaggle 比赛并完成了向机器学习领域的转换。他领导了一个有线电视推荐项目，由于工作需要，学习了许多关于大数据的技术，适用于并行计算的算法和人们对机器推荐的反应方式。最近几年，他是机器学习应用的顾问，并进行数字广告、电信、半导体制造、系统管理和客户体验优化等方面的实用预测分析。

Henrik Brink

致谢

感谢 Manning 出版社和所有参与本书编写的人员，特别是苏珊娜·克莱恩，感谢她在本书写作过程中的耐心指导。

感谢 Beau Cronin 为本书作序，也感谢 Valentin Crettaz 对所有章节进行了深入的技术校对。许多其他审核人员给我们提供了有益的反馈，具体有 Alain Couniot，Alessandrini Alfredo，Alex Iverson，Arthur Zubarev，David H. Clements，Dean Iverson，Jacob Quant，Jan Goyvaerts，Kostas Passadis，Leif Singer，Louis Luangkesorn，Massimo Ilario，Michael Lund，Moran Koren，Pablo DomínguezVaselli，Patrick Toohey，RavishankarRajagopalan，Ray Lugo，Ray Morehead，Rees Morrison，Rizwan Patel，Robert Diana 和 UrsinStauss。

马克·弗特罗夫感谢克雷格·卡迈克尔分享他对机器学习的痴迷；对妻子帕特里夏和女儿艾米多年的支持表示感谢。

亨里克·布林克对 Wise. io 团队和创立者分享他们使用机器学习解决实际问题的热情表示感谢。同时，感谢他的父母 Edith 和 Jonny，还有他的兄弟姐妹传递知识和话语的热情，更重要的是要感谢他的妻子 Ida 和儿子 Harald，感谢他们的爱与支持。

约瑟夫 W. 理查兹也要感谢 Wise. io 团队，感谢他们对机器学习分享的热情和无尽的努力，这使得他每天都过得无比充实。他特别感谢他的父母 Susan 和 Carl，教他终身学习的快乐，并灌输要努力工作和同情的价值观。最重要的是感谢他的妻子 Trishna，感谢她无尽的爱、激情与支持。

译者序

虽然机器学习已经有几十年的发展历史，但对于大多数人来说，它还是那么神秘！它是只有在顶尖的象牙塔里或者尖端企业中才会研究和使用的技术。大量的数学公式令人望而生畏，它就站在高高的云端，俯视着芸芸众生。让大家都能使用和享受这一“尖端”技术，仍然任重而道远。Brink 的这本《Real - World Machine Learning》给大家带来了曙光，这里没有令人望而生畏的数学公式和证明推导，而是将整个机器学习应用过程娓娓道来，分享他的实际工作经验和技术。只要你有一定的编程基础，就可以通过学习本书来使用“机器学习”这么“高深”的技术了。“旧时王谢堂前燕，飞入寻常百姓家”，让高精尖的“机器学习”技术来到我们中间，让我们近距离体验它的魅力吧！

作为计算机专业的教师，我喜欢尝试新的东西，兴趣驱使着我不断地学习和探索新的领域。当看到《Real - World Machine Learning》这本书时，我真正感觉到适合自己的书稿来了，我一定要完成它的翻译。经过不懈的努力，终于迎来了胜利的曙光。翻译过程痛并快乐着，一方面感受作者的博大精深，一方面惶恐于自己的才疏学浅。在整个翻译过程中，我都惴惴不安，唯恐不能准确传达作者的真实意图。对于各种专业术语，不论是统计学的、算法的，还是汽车的、广告的，都在网上搜索了很多遍，并向专家反复请教，推敲多遍方才确定。对于每一字每一句，都反复琢磨很多遍，对翻译片段反复组合，优中选优，力求忠于原文并符合中文的表达习惯，避免在阅读过程中让读者感到生硬和不连贯。

“纸上得来终觉浅，绝知此事要躬行。”这是我多年来教学过程中领悟最深的一句话。无论学习还是做事，必须落实到行动上，不要做“语言的巨人，行动的矮子”。读书更是如此，推荐大家在阅读本书时，除了逐字逐句地认真阅读以外，还必须要付诸实践。把作者描述的过程、方法和注意事项在实际中应用一遍，哪怕是做一道 Kaggle 竞赛试题或重复一遍别人已经完成的任务，都是不错的选择。记住，重要的是“做一遍”，而不仅仅是“看一遍”！有了这些“应用基础”之后，再回头学习机器学习的基本理论和数学基础就比较容易了。俗话说，“万事开头难”，只要入了门，一切都好办了。衷心希望大家都能通过本书成为机器学习的高手和巨人。

在翻译过程中得到了来自各方面的帮助，数理部的孙静波老师和商学院的王灿老师在统计学方面给予了大量帮助；公共外语教学部的蒋彩艳老师，在英美习惯用法和个别翻译中给予了一定帮助，在此表示感谢！

虽然在翻译过程中，经过反复思考，力图传达作者确切的意图，但由于译者水平有限，错误疏漏之处在所难免，望各位读者、专家和业内人士不吝提出宝贵意见。

程继洪

关于本书

《实用机器学习》的读者对象是针对想要把机器学习应用于实际问题的人。它详细阐述了机器学习的主要组成部分：工作流程、算法和工具。关注点是著名算法的实际应用，而不是创建一个算法。构建和使用机器学习模型的每个步骤都有详细描述，并有从简单到中等复杂的实例与之对应。

主要内容

第 1 部分，“机器学习工作流程”介绍基本的机器学习工作流程，并分章节对每个步骤加以介绍。

第 1 章，“什么是机器学习”介绍机器学习的应用领域和用途。

第 2 章，“实用数据处理”，详细介绍机器学习流程中的数据处理和准备工作。

第 3 章，“建模和预测”，介绍构建简单的机器学习模型，并利用应用广泛的算法和库进行预测。

第 4 章，“模型评估和优化”，深入研究机器学习模型，并对其进行评估和性能优化。

第 5 章，“基础特征工程”，介绍利用领域知识对原始数据进行提高的常用方法。

第 2 部分，“实际应用”，介绍模型规模化和从文本、图片和时间序列数据中提取特征的技术，来提高绝大多数现代机器学习的性能。本部分包括 3 个有完整实例的章节。

第 6 章，“实例：NYC 出租车数据”，这是第一个完整实例章节，会预测乘客的倾向性行为。

第 7 章，“高级特征工程”，包含高级特征工程过程，介绍从自然语言的文本、图片和时序序列数据中提取有价值的数据。

第 8 章，“NLP 高级案例：电影评论情感预测”，运用高级特征工程知识预测在线电影评论的情感。

第 9 章，“扩展机器学习流程”，介绍扩大机器学习系统的数据规模、预测吞吐量和降低预测间隔的技术。

第 10 章，“案例：数字显示广告”，构建大型数据的模型，预测数字广告点击行为。

如何使用本书

如果你是机器学习新手，第 1 ~5 章将引导你学习研究和准备数据、特征工程、建模和

模型评估过程。Python 实例采用流行的数据处理、pandas 和 Scikit - Learn 机器学习库。第 6 ~ 10 章，包括 3 个实际机器学习案例、高级特征工程和优化的话题。由于学习库封装了大部分的复杂性，因此代码示例可以很容易地应用到你自己的机器学习系统中。

目标读者

本书可以使程序员、数据分析师、统计学家、数据科学家和其他专业人士将机器学习应用于实际问题，或者简单地理解它。他们将获得实用数据建模、优化和开发机器学习系统的经验，而没必要了解特定算法的理论推导。机器学习的数学基础是针对感兴趣的人的，某些算法在较高的层次上进行解释，本书提供给那些想深入学习的人，我们的焦点是获得实际结果以解决手头的问题。

代码约定，下载和软件需求

本书包含许多示例源代码，或者以编号的清单出现，或者嵌入在正文中，但无论哪种情况，都以固定宽度的这种字体显示，以区别于正常的文本。

源代码使用 Python，pandas 和 Scikit - Learn 编写。与章节相应的 iPython 笔记文件可在 GitHub 上下载，地址为 https://github.com/brinkar/real - world - machine - learning，也可以通过关注机械工业出版社计算机分社官方微信订阅号“IT 有得聊”，输入 5 位数号“56922”后获得资源下载链接，还可以登录 golden-book.com 搜索本书并进行下载。

笔记文件(扩展名为 .ipynb)与章节相对应。样本数据包含在 data 文件夹中，只要必需的库随 iPython 一起安装，那么所有的笔记文件都能执行。图形由 matplotlib 和 Seaborn 的 pyplot 模块生成。

在有些情况下，由 iPython 产生的图形被提取出来作为本书的插图(为了适应打印质量和电子书显示，有些已经做了修改)。

作者简介

Henrik Brink(亨里克·布林克)是一名数据科学家,对应用机器学习进行工业和学术应用开发有着丰富的经验。

Joseph Richards(约瑟夫 W. 理查兹)是一位资深的数据科学家,具有应用统计和预测分析方面的专业知识。Henrik 和 Joseph 是 Wise. io 的联合创立者,Wise. io 是一家提供工业机器学习解决方案的开发商。

Mark Fetherolf(马克·弗特罗夫)是数据管理和预测分析公司 Numinary Data Science 的创始人和总裁。他曾在社会科学研究、化学工程、信息系统性能、容量规划、有线电视和在线广告应用等方面担任统计师和分析数据库开发人员。

关于封面插图

《实用机器学习》封面插图标题为“中国战士(ChinoisCombattant)”或“中国武士(Chinese fighter)”。插图取自19世纪法国出版的Sylvain Maréchal编撰的《区域服饰习俗第四卷》,其中的每幅图都是精心绘制并手工着色。Maréchal丰富的收藏给我们生动地展示了200年前不同城市和地区的文化差异。由于相互隔离,人们说着不同的方言和语言。无论在城市的街道、小城镇或乡村,都可以很容易地通过他们的穿着分辨出他们在哪里生活以及他们的生活习惯。

服饰密码从那时起已经改变,那个时候的人们根据区域和阶级的不同拥有的服饰特色现在已经逐渐消失。现在人们已经很难通过服饰区分不同大洲的居民,更不用说不同的城镇或地区了。也许我们已经将文化多样性换成了一种更加多样化的个人生活——当然是为了更加多样化和快节奏的科技生活。

当计算机图书多到无法区分时,本书采用Maréchal的两世纪以前的区域生活的多样性图片作为图书封面的方式,庆祝计算机图书的创造性和主动性。

目录

第 2 部分 实际应用

第 1 部分

机器学习工作流程

在本书的第 1 部分，将介绍基本的机器学习工作流程。每章都涵盖流程中的一个工作步骤。

第 1 章，介绍机器学习的用途，以及为什么要阅读本书。

第 2 章，研究基本机器学习流程中的数据处理，读者将学习到一些通用方式，从现实世界和纷乱的数据中清理和提取有价值的信息。

第 3 章，随着一些模型算法及其应用的学习，开始构建简单的机器学习模型。

第 4 章，深入研究机器学习模型，并对它们进行评估和性能优化。

第 5 章，致力于特征工程。从数据中提取特征是构建和优化机器学习系统的重要组成部分。