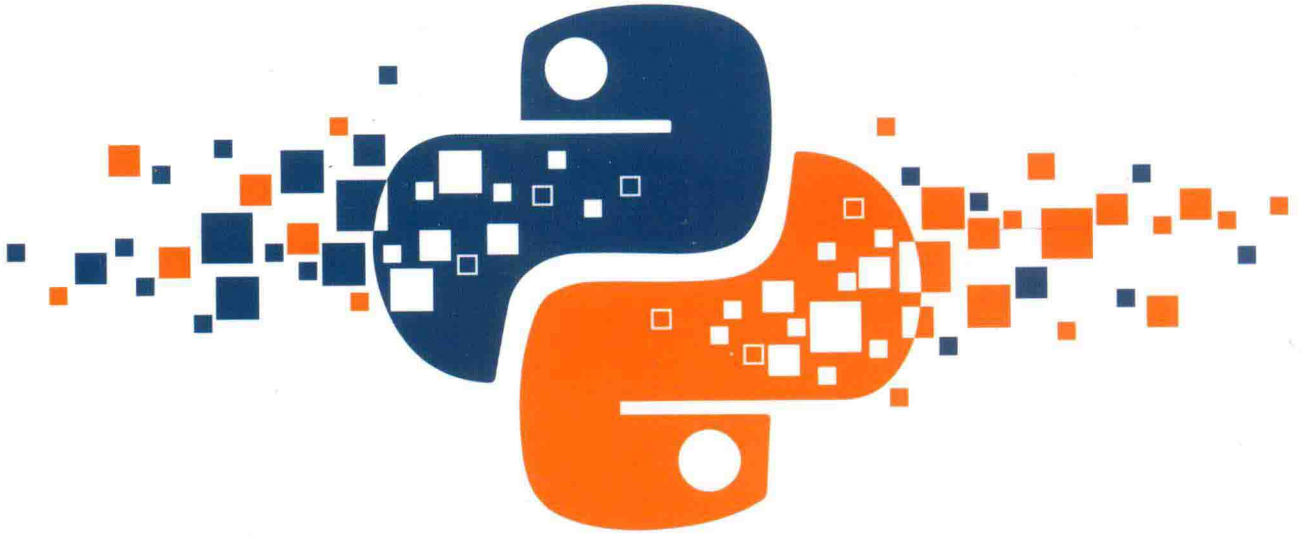
Broadview®
www.broadview.com.cn

从零开始实践Python
深入浅出运用机器学习



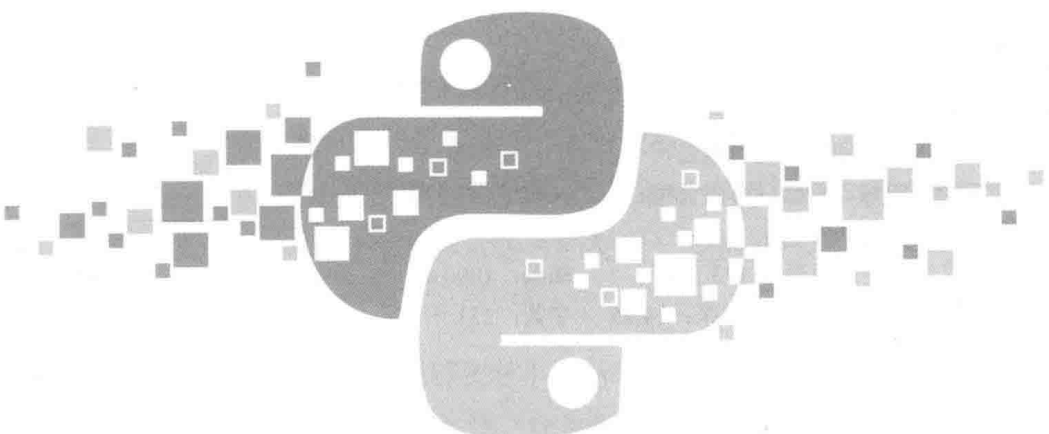
机器学习Python实践

魏贞原 著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

机器学习 Python实践

魏贞原 著



電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书系统地讲解了机器学习的基本知识，以及在实际项目中使用机器学习的基本步骤和方法；详细地介绍了在进行数据处理、分析时怎样选择合适的算法，以及建立模型并优化等方法，通过不同的例子展示了机器学习在具体项目中的应用和实践经验，是一本非常好的机器学习入门和实践的书籍。

不同于很多讲解机器学习的书籍，本书以实践为导向，使用 scikit-learn 作为编程框架，强调简单、快速地建立模型，解决实际项目问题。读者通过对本书的学习，可以迅速上手实践机器学习，并利用机器学习解决实际问题。本书非常适合于项目经理、有意从事机器学习开发的程序员，以及高校相关专业在的读学生阅读。

图书在版编目（CIP）数据

机器学习：Python 实践 / 魏贞原著. —北京：电子工业出版社，2018.1

ISBN 978-7-121-33110-7

Ⅰ. ①机… Ⅱ. ①魏… Ⅲ. ①机器学习②软件工具—程序设计 Ⅳ. ①TP311.561②TP181

中国版本图书馆 CIP 数据核字（2017）第 288527 号

策划编辑：石　倩
责任编辑：徐津平
特约编辑：赵树刚
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　　印张：14.25　　字数：251 千字
版　　次：2018 年 1 月第 1 版
印　　次：2018 年 1 月第 1 次印刷
定　　价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件到 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

序言

人工智能作为一种新技术，它的发展变迁可以用著名的 S 曲线来表示。具有划时代意义的新技术（Disruptive Innovation），往往出现在社会发展的成熟期。这时国家的 GDP 增长开始减缓，生产成本居高不下，相比之下社会需求却没有显著增加。现在的人工智能技术发展正处于这样一个时期，可以说是应运而热。但我们更关心的是它是否到了腾飞的前期。我个人觉得时机已经到来，原因有三：数字化、物联网及人工智能（AI）技术。

第一，社会生活的高度数字化进程已经让人与人之间的联结几乎完全可以用数字化的方式来描述。这一切极大地归功于智能手机和社交媒体的深度普及。关于我们的信息都被记录在信息系统里，无论是以格式化的形式（ERP、银行系统、电商系统等），还是以非格式化的形式（社交媒体上的文字或语音的交流），这些都已成为机器可以分析解读的数据，而且还在不断积累中。

第二，物联网技术使信息系统能够实时不间断地从我们的日常活动中获取新的信息，并且可以通过双向通信机制给予实时反馈。比如智能手表和智能扫地机器人等。

第三，人工智能技术本身的发展已经到了足以支撑大规模商业化应用的阶段。无论是人脸识别还是医疗文献分析，人工智能已经作为工具出现在我们的身边。

很多人还在争论人工智能是否会成为人类的敌人，尤其是在 AI 技术发展到

泛人工智能（Artificial General Intelligence，AGI）甚至超级人工智能（Artificial Super Intelligence，ASI）的时候。这是否会成为现实或者什么时候会成为现实谁也说不清楚。但在人机同行的今天，如果还不赶快学习充实自己，明天我们肯定会遇到已经用 AI 版本升级了的人类对手。这种痛也许只有百多年前抵抗英法联军的八旗子弟才领悟过，但为时已晚。如何不掉队，赶上甚至超越时代发展潮流，学习新技术是唯一手段。

就像前几次工业革命一样，人工智能带来的技术革命也有三个关键成功因素（3M）：数据（原材料：Raw Material）、技术（机器：Machine）及商业模式（Business Model）。在现实世界中，数据的金矿已经大量积累，并等待我们去开采和精炼。技术，如 Python、卷积神经网络等都已被成熟应用。商界精英更是想出了很多商业化的应用场景，在同声翻译、文本分析、医疗影像分析等领域展开了众多的投资和商业化应用。本书作为一本介绍 Python 的技术类专业书籍，立足机器学习中的监督式学习，围绕着课程、项目和方法深入浅出地介绍了如何使用 Python 来完成机器学习的相关工作。内容涵盖了人工智能的三个关键成功因素，以体系化和细致的讲解方便读者理解和学习有关机器学习的全过程。本书是一本实战性很强的参考书籍，帮助我们在人工智能时代迅速掌握新技术的精髓。

周德标
副合伙人
IBM 大中华区董事长执行助理
“周教授谈人工智能”微信公众号作者

前言

“这是最好的时代，也是最坏的时代”，这是英国文豪狄更斯的名著《双城记》开篇的第一句话，一百多年来不断被人引用。这里再次引用它来形容智能革命给我们带来的未来社会。从 2016 年 AlphaGo 在围棋比赛中战胜韩国选手李世石，到 2017 年 Master 战胜世界排名第一的围棋选手柯洁，人工智能再一次引起了世人的注意。在大数据出现之前，人工智能的概念虽然一直存在，但是计算机一直不擅长处理需要依赖人类的智慧解决的问题，现在换个思路就可以解决这些问题，其核心就是变智能问题为数据问题。由此，全世界开始了新一轮的技术革命——智能革命。

自从 1687 年艾萨克·牛顿发表了论文《自然定律》，对万有引力和三大运动定律进行了描述，人类社会进入了科学时代。在此之后，瓦特通过科学原理直接改进蒸汽机，开启了工业革命的篇章，由于机器的发明及运用成为这个时代的标志，因此历史学家称这个时代为“机器时代”。机器时代是利用机器代替人力，在原有的产业基础上加上蒸汽机形成新的产业，例如马车加上蒸汽机成为火车，改变了人的出行方式；帆船加上蒸汽机成为轮船，让货物的运输变得更加便捷。同时，原有的工匠被更加便宜的工人替代，社会的财富分配不均，社会进入动荡期，如英国大约花费了半个世纪的时间才完成了工业革命的变革。同样，第二次工业革命和信息革命，每一次变革都让财富更加集中，给社会带来动荡。第二次工业革命同样花费了半个世纪的时间，一代人才消除工业革命带来的影响，让大部分人受益。当前的智能革命也会带来财富的重新分配和社会的动荡，

当然目前的政府对这次革命的过程都有了足够的了解，能够把社会的动荡控制在最小范围，但是在变革中的人依然需要经受这次变革带来的动荡。

每一次变革都是一次思维方式的改进，工业革命是机器思维替代了农耕时代的思想；信息革命是香农博士（1916—2001 年）的信息论带来的思想方法替代机器思维，并成为社会主导思想；在这次智能革命中，以大数据为核心的思维方式将会主导这次变革。在历次的技术革命中，一个人、一家企业，甚至一个国家，可以选择的道路只有两条：要么加入变革的浪潮，成为前 2%的弄潮儿；要么观望徘徊，被淘汰。

要成为 2%的弄潮儿，需要积极拥抱这次智能变革，掌握在未来社会不会被淘汰的技能。在以大数据为基石的智能社会，利用机器学习算法对数据进行挖掘，是使机器更智能的关键，掌握数据挖掘是拥抱智能社会的举措之一。本书就将介绍如何利用机器学习算法来解决问题，对数据进行挖掘。

作　者

目录

第一部分　初始

第二部分 数据理解

第三部分 数据准备

第四部分 选择模型