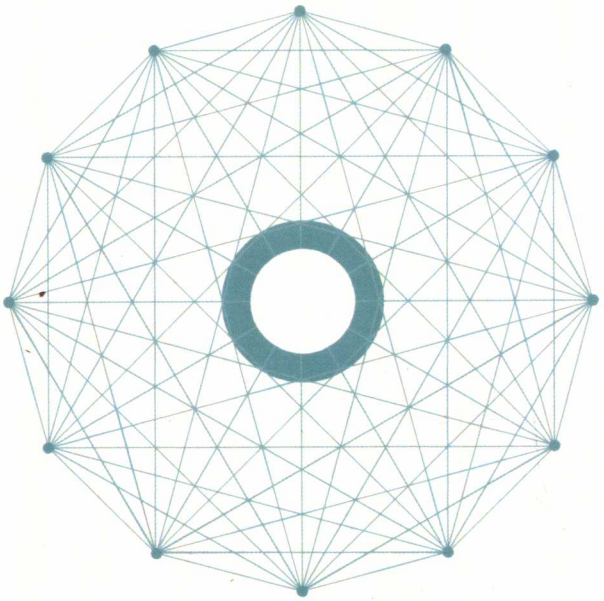

Python 机器学习实战

翟　隽　张怿檬　张目清 等◎著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

Python机器学习实战

裔 隽 张怿檬 张目清 等 著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

·北京·

图书在版编目（CIP）数据

Python机器学习实战 / 裔隽等著. —北京：科学技术文献出版社，2018.1
ISBN 978-7-5189-3808-7

Ⅰ.①P… Ⅱ.①裔… 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（2018）第012364号

Python机器学习实战

策划编辑：孙江莉　　责任编辑：宋红梅　　责任校对：文　浩　　责任出版：张志平

出 版 者　科学技术文献出版社
地　　址　北京市复兴路15号　邮编 100038
编 务 部　(010) 58882938，58882087（传真）
发 行 部　(010) 58882868，58882874（传真）
邮 购 部　(010) 58882873
官方网址　www.stdp.com.cn
发 行 者　科学技术文献出版社发行　全国各地新华书店经销
印 刷 者　虎彩印艺股份有限公司
版　　次　2018年1月第1版　2018年1月第1次印刷
开　　本　710×1000　1/16
字　　数　358千
印　　张　20.5
书　　号　ISBN 978-7-5189-3808-7
定　　价　68.00元

《Python 机器学习实战》著者名单

（以下人员工作单位均为上海汇付数据服务有限公司）

斋　隽　张怿檬　张目清　寇　莉

熊　怡　刘　聃　王新乐　袁　媛

推荐序

认识裔隽十六年了，他总是像大男孩一样，对世界饱含兴趣，对生活充满热情。他原本是读心理学的，这些年干过银行业务、电商管理、支付技术等。他喜欢背着“国家地理”的帆布包，记录城市，记录山川，记录点滴生活。他深谙硅谷管理风格，工作中幽默风趣，无论时世多么艰难，有他在，总是欢声笑语。他对很多领域的热爱远超出了一般专业水准，因此同事们喜欢亲切地叫他“裔大师”。

两周前，在一次公司内部大会后，裔大师兴奋地递给我一本书，说他们写了本有关机器学习的技术书，请我作序。我当时吃了一惊。虽然我知道他和数据中心的同事进步很快，过去几年取得了骄人的成绩，绝对值得整理分享，但每天繁重的加班使他们的业余时间所剩无几，应全部交给睡眠才是，或许还不够。

但粗略拜读后，我开始明白他们为什么要牺牲宝贵的睡眠时间，来“憋”出这么一本“经验手册”。

一是大数据和机器学习对支付、对金融，乃至未来几乎一切产业，都太重要了，是当代提升产业、撬动变革的法门。正因为如此，在几乎所有论坛、每天“朋友圈”、茶余饭后，大家都在谈论人工智能的话题。但平心而论，大多数谈话或文章，说的都是AI未来的意义，往往沉迷于科幻般的讨论和想象，其中不少人可能还是“叶公”，而真正能够亲手打造实用应用系统的人并不多，

我们太需要“know-how”这方面的知识技能和经验。我认为，这正是本书努力的目标。

二是大数据和机器学习等新兴技术需要工程师的热情分享。现代支付就是技术嬗变的产物，由早期的数据库和OLTP发端，到互联网和移动技术，而当今最大的驱动力来自大数据、机器学习和云计算。我一直认为，技术获得和提升的路径，从来就不应该是团队的闭门造车，也不靠高人的山间开悟，而应是业界热情分享，相互切磋。越乐于分享，就越多被分享；越乐于切磋，也就越容易打磨进步。我想，这就是本书写作的动力源泉。

裔大师团队把汇付过去五年以来在这方面的工作，尤其是风险管理、客户画像、精准营销等成功案例，甚至是开发的全新方法论，很好地总结分享给了读者。在书中，我看到了他们团队的激情和快乐，正是这样一种激情和快乐，使他们能迅速取得业界领先的成就。

当然，就像很多应用技术类的书籍，由于实效性要求，对文字推敲来不及做过多追求。这可能是后续需改进的地方。但瑕不掩瑜，我相信本书会帮助和启迪很多业界朋友，也会促使作者更加勇攀高峰，把新金融变得更加智能。

是为序。

周　晔

汇付天下有限公司董事长兼总裁

前言

2015年春天，下了决心，学习Python，起初是为了做公司一些项目的“黏合剂”，看中了Python方便执行、跨操作系统、有丰富的第三方扩展包等特点；之后也尝试使用Pandas来进行数据的处理，以及通过Scikit-Learn等来替代R语言进行建模；这时候我们的数据建模小组已经发现虽然R语言搭建模型进行处理等能力颇强，但是要转换到生产环境，还是不太合适。碰巧2015年机器学习热浪袭来，Python俨然是机器学习的头号主力。几个因素综合之下，我们开始学习、测试和探索Python语言在一些系统中的应用和在机器学习领域的使用。

不知不觉，两年多过去，我们团队已经用Python开发了十几个机器学习的项目，有基础的如人脸识别，也有和业务系统紧密结合的如用于交易风险控制、用户精准营销等，也没有忘记作为“黏合剂”的初衷，用于微服务的项目实践中。

从很多年学习和工作的经验来看，计算机编程语言本身的差异的确存在，各自有一定的适用范围，但更重要的是对于编程的理解、对于一个需求如何用程序实现的方式、对一个项目如何开发的方式，这些也是我们在这几年使用Python开发项目过程中体会到的。

市场上有很多Python从入门到高级的专著，我们写这本书的目的就是想从更加实际的角度，分享一些项目开发中总结的经

验，帮助使用 Python 进行机器学习开发的新手或者从其他语言转过来的朋友，少踩一些坑。我们不是大牛大神，水平能力和时间都有限，所以难免有疏漏之处。Python 本身博大精深，我们只是学到了一点皮毛，我们也会持续修订和跟踪 Python 在数据建模和机器学习等项目开发中的发展。

人工智能和机器学习近几年很热门，相信它终究会成为一个标准功能，应用在生活工作的方方面面。同时，计算机编程技术也会继续突飞猛进，从单机、终端大机模式、客户端服务器多层模式、API 接口模式等一直到现在的微服务、无服务模式、公有云，等等。在这个飞速发展的世界中，或许有压力，所以更需要学习的动力。

本书中所有的源代码都在 Github 上进行了分享，方便大家下载学习。所有相关的参考资料也注明了出处。

谢谢公司领导的远见和支持，使我们可以有资源来研究新事物；谢谢我的同事们，聪明而有韧性。

谢谢家人的鼓励和支持，忍受我的偏执；谢谢我的妻子，做好了所有的后勤工作，我才能安逸地学习和写作！

谢谢这个世界，化梦为马，万里驰骋！

斋 隽

编写说明

章节说明

本书没有基础的 Python 语言的入门内容，所以如果从未学习过 Python 的话，可能不适合阅读本书。

本书的主要内容分为四大部分：

（1）Python 开发程序的一些方法技巧，如虚拟环境管理、敏捷开发入门、单元测试等；

（2）Python 中的一些中级使用技巧，如列表生成式、多线程与多进程、Python 程序性能分析等；

（3）机器学习的基本概念和常用算法介绍，以及如何选择合适的算法；

（4）一些使用 Python 进行建模和机器学习的实际例子。

我们这样设计是在平时工作学习中发现，作为程序员和数据建模或者机器学习的同事在知识结构和实际应用程序上有一些差异。很多程序员无法理解模型的训练、调参等概念，因为这些和传统的不管是瀑布式还是敏捷式的开发都大相径庭；而建模人员对于一个应用项目的需求、详细设计、开发、测试、部署、性能等也很难理解。于是我们在实践中逐渐摸索并采用的方法就是大家都各自往前走一步，程序人员要了解建模的基本流程，而作为建模人员要了解开发的各个步骤的来龙去脉。

本书既能为 Python 程序开发人员夯实基础，提升编程技能，又能为使用 Python 的机器学习从业者提供大量实际案例，使其获得机器学习实战经验，帮助开发人员和建模人员取长补短，弥补各自知识结构上的欠缺，打造更优秀的具有综合能力的团队。

因为篇幅有限，只能蜻蜓点水，各方面略有涉及。

章节作者

第一部分 Python 开发实战
第一章 开发环境选择与比较：张目清
第二章 Anaconda 使用介绍：裔隽
第三章 开发规范与方法：张目清
第四章 单元测试与代码覆盖率：张怿檬

第二部分 Python 编程技巧
第五章 列表生成式：裔隽
第六章 Collections 库：裔隽
第七章 迭代器：裔隽
第八章 Python 多线程与多进程浅析：裔隽
第九章 Python 程序性能分析初步：裔隽

第三部分 Python 机器学习基础
第十章 机器学习基础：张怿檬
第十一章 主要算法概览：张怿檬
第十二章 *K* 近邻算法：张怿檬
第十三章 主成分分析：刘聃
第十四章 逻辑回归：刘聃
第十五章 朴素贝叶斯分类器：王新乐
第十六章 决策树算法：张怿檬
第十七章 支持向量机：张怿檬
第十八章 *K*-Means 聚类：王新乐
第十九章 人工神经网络：张怿檬
第二十章 如何选择合适的算法：张怿檬
第二十一章 Python 机器学习工具：张怿檬

第四部分 Python 机器学习实例
第二十二章 基于 RFM 的 P2P 用户聚类模型：熊怡
第二十三章 文本的主题分类：熊怡
第二十四章 利用机器翻译实现自然语言查询：袁媛

第二十五章　身份证汉字和数字识别：寇莉
第二十六章　人脸识别：张怿檬

审　校

裔　隽　何　雯　张怿檬

感　谢

文字校对：徐　晋　卓梵妮　马莉莉
P2P第三方资金存管模式研究论文：王杰

本书约定

在本书中出现代码时，我们使用这样的样式：

```
>>> s = 'hello world'
>>> comp = {x for x in s}
>>> print( comp)
{' ', 'h', 'd', 'o', 'l', 'e', 'w', 'r'}
```

其中，以>>>开头的行表明是代码输入，不含>>>的行是代码的输出结果。

勘误和服务

文中使用的每一段代码都可以在我们的Github上找到，包括一些由于篇幅限制未能完整在书里展示的代码。另外，我们也会在这个仓库上持续更新，跟进最新的Python编程技巧和机器学习内容：

https://github.com/chinapnr/How-to-Python-and-Machine-Learning-Book

虽然花了很多时间来检查和核实书中的文字、代码，但由于能力有限，

难免会存在一些纰漏，如果亲爱的读者发现书中的不足之处，恳请反馈给我们。反馈的方式是在本书 Github 仓库的 issue 板块上留下您的意见，也可以发送邮件至：yimengzh_book@163.com。

目 录

第一部分 Python 开发实战

第二部分 Python 编程技巧

第三部分 Python 机器学习基础

第四部分 Python 机器学习实例