O'REILLY®

HZ BOOKS
华章IT

Python

机器学习实践

测试驱动的开发方法

Thoughtful Machine Learning with Python: A Test-Driven Approach

Matthew Kirk 著
刘江一 上官明乔 白皓 刘旭斌 译

机械工业出版社
China Machine Press

O'REILLY

Python 机器学习实践

高效智能的开发方法

Matthew Kirk 著

[illegible]

Python机器学习实践：测试驱动的开发方法

Matthew Kirk 著

刘江一 上官明乔 白皓 刘旭斌 译

opol • Tokyo O'REILLY®

edia, Inc.授权机械工业出版社出版

机械工业出版社

图书在版编目（CIP）数据

Python 机器学习实践：测试驱动的开发方法 /（美）马修·柯克（Matthew Kirk）著；刘江一等译 . —北京：机械工业出版社，2017.10
（O'Reilly 精品图书系列）
书名原文：Thoughtful Machine Learning with Python: A Test-Driven Approach
ISBN 978-7-111-58166-6

I. P… II. ①马… ②刘… III. 软件工具－程序设计 IV. TP311.561
中国版本图书馆 CIP 数据核字（2017）第 245948 号

北京市版权局著作权合同登记
图字：01-2017-0912 号

书　　名 / Python 机器学习实践：测试驱动的开发方法
书　　号 / ISBN 978-7-111-58166-6
责任编辑 / 陈佳媛
封面设计 / Randy Comer，张健
出版发行 / 机械工业出版社
地　　址 / 北京市西城区百万庄大街 22 号（邮政编码 100037）
印　　刷 / 中国电影出版社印刷厂
开　　本 / 178 毫米 ×233 毫米　16 开本　13 印张
版　　次 / 2018 年 1 月第 1 版　2018 年 1 月第 1 次印刷
定　　价 / 59.00 元（册）

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：(010)88379426；88361066
购书热线：(010)68326294；88379649；68995259
投稿热线：(010)88379604
读者信箱：hzit@hzbook.com

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。无论是通过书籍出版，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

译者序

人类与人工智能共存的未来

AI（Artificial Intelligence），即人工智能，在不久以前还仿佛只存在于科幻大片中，例如《黑客帝国》《机械公敌》中神一样无所不能的超级AI，《她》《机械姬》中进步神速的AI个体，《人工智能》中惹人怜惜的机器小男孩……

好像小宇宙爆炸一般，在刚刚过去的“人工智能元年”——2016年，人们惊讶地发现AI已经渗透在生活和生产的方方面面。它决定你打开网站时会看到一些你正好喜欢的内容，它在你自拍时帮你选择效果最棒的相机参数，它能打败人类的顶尖棋手，它能像朋友陪你聊天，像医生给你看病，像调度员安排最优的物流路径，像资深的顾问给公司提供管理建议……在某些方面，它傻乎乎的，还不如幼儿园小朋友厉害；可在它擅长的领域，毫无争议地俯视着人类“低微”的能力。即使想象力再有限的人，也知道在不久的将来，AI将会逐步地、深远地改变人类社会，改变我们以及子孙后代的生活方式。

在这样的时代背景下，作为技术工作者、学习者、爱好者，我们自然而然地对人工智能产生了浓厚的兴趣。

机器学习作为人工智能的核心算法，已经存在了几十年。从论文上的数学表达，到铺天盖地的实际应用，说起来简单的步骤却是它几十年走过的漫长旅途。

所以当我们看到《Thoughtful Machine Learning with Python》这个书名时，立刻就被吸引了。

我们知道，Python是当前最受欢迎、效率最高的编程语言之一，而且有丰富的算法库可供调用。使用Python作为编写机器学习软件的最初尝试，是一个非常明智的选择。

机器学习算法的效果并不总是很好，因为有一些可调的参数起着至关重要的作用。在选择最优或者较好参数的过程中，选择合适的测试数据、科学的测试方法，可以更快地获得优良的算法效果。

基于以上原因，本书一开始就立足于软件编写、算法测试的实践指导，为读者理解示例代码、动手编写自己的程序做必要的铺垫。

然后，作者才开始简明扼要地介绍机器学习算法的定义，以及读者必须知道的算法类别，并轻轻点出这些算法各自有何神通和每个算法的死穴。

第 3 章到第 9 章深入详实地讲解了几种最有代表性的机器学习算法：K- 最近邻、朴素贝叶斯分类、决策树和随机森林、隐马尔可夫模型、支持向量机、神经网络，以及聚类。这些章节不但讲解了算法核心部分的数学表达，而且用机智、形象的语言描述了算法如何在实际生活中解决问题，并给出了关键的 Python 代码示例和算法训练、测试过程。

在测试数据的选择和处理上，在模型改进的方法论上，作者毫无保留地将点滴宝贵的经验传授予读者，这对于读者进一步在实践中开发自己的软件，提供了实实在在的帮助。这一点不仅体现在每一个介绍算法的章节，还体现在第 10 章“模型改进与数据提取”中。

只要是有一定数学知识和编程基础的读者，都能够在本书中汲取到机器学习算法的精华。由于篇幅所限，必然有些遗珠未能详述，但本书可引导读者走进琳琅满目的人工智能应用背后那个神秘的世界，不失为一个登堂入室的台阶。也许本书能开启你和人工智能的未来！

读者对象

本书适合以下几类读者阅读：

- 有志于成为人工智能专家的从业人员
- 有一定算法和编程基础的技术爱好者
- 计算机、模式识别等相关专业的在读人士

勘误支持

鉴于译者的知识和视角，本书中难免出现用词错误、技术适用性的问题。在此，译者恳请读者不吝指教。欢迎读者发送邮件到 judy_l_2017@sina.com，与译者交流。

译者分工

本书由来自 IBM 中国开发中心的软件工程师及项目经理联合翻译完成。其中，

- 刘江一（目前就职于 IBM 中国开发中心，e-mail: judy_l_2017@sina.com）
  翻译了第 2 章、第 3 章、第 7 章、第 8 章。
- 上官明乔（目前就职于 IBM 中国开发中心， e-mail:shgmq@163.com）
  翻译了第 4 章和第 9 章。
- 白皓（原 IBM 软件工程师，目前就职于京东，e-mail: salanhess@126.com）
  翻译了第 6 章和第 10 章。
- 刘旭斌（目前就职于 IBM 中国开发中心，e-mail: liuxb76@sina.com ）
  翻译了第 5 章和第 11 章。
- 第 1 章由上面四位译者共同翻译。

致谢

感谢华章公司引进了原著版权，这本书得以面市的最核心要素。

感谢华章公司的和静和陈佳媛编辑，她们专业的编辑能力为本书提供了重要的质量保证。

翻译本书的过程是一种学习与思考的结合，也是和伙伴合作与交流的经历。非常庆幸遇到了睿智又勤奋的伙伴，即使在繁忙的工作和节奏极快的生活中，也努力完成了翻译和审阅任务。

目录

前言

我写本书第 1 版时，因为同事缺乏纪律感到很无奈。2009 年我正在做大量机器学习的项目，发现一旦引入了对向量机、神经网络或者其他东西的支持，一般的编程实践突然间就出问题了。

本书第 1 版是我对这些问题的答复。当时我的代码都是用 Ruby 编写的，第 1 版也是基于 Ruby 语言编写的。你可以想象那有多困难，所以我很高兴可以推出本书基于 Python 语言的新版本。我检查了本书的大部分章节，替换了示例代码，并使它与时俱进，对写机器学习程序的人更有用处。我希望你们能喜欢它。

我在第 1 版中就声明，欢迎随时与我沟通。如果你有什么想和我讨论的，欢迎来信，我的邮箱是：*matt@matthewkirk.com*。如果你会来西雅图，我也很乐意跟你在咖啡馆相聚。

本书约定

在本书中下列字体有特殊的意义：

斜体字

表示术语、链接、电子信箱地址、文件名和文件后缀名。

等宽字体（`Constant width`）

用于程序清单，也用于在段落中引用程序元素，例如变量名、函数名、数据库、数据类型、环境变量、程序语句和关键词。

加粗等宽字体（**Constant width Bold**）

表示命令，或者其他应该由用户输入的文字信息。

斜体的等宽字体（*constant width italic*）

表示此处应该替换为由用户提供的数值，或者根据上下文确定的数值。

这个图案表示通用的注意事项。

如何使用示例代码

补充材料（示例代码、练习，等等）都可以在链接 *http://github.com/thoughtfulml/examples-in-python* 中下载。

你可以在自己的程序和文档中使用本书提供的示例代码。你不用联系我们来取得许可，除非你想大量地复制这些代码。例如，在程序中使用本书中的几段代码无需取得许可。但把 O'Reilly 书籍里的示例代码刻录成光盘就需要取得许可。引用本书内容和例子来回答问题，无需取得许可。但在你的产品文档里大量包含本书的示例就需要取得许可。

我们期望，但不强求，引用时注明出处。出处通常包括标题、作者、出版商和国际标准书号。例如：《Thoughtful Machine Learning with Python》（《Python 机器学习实践：测试驱动的开发方法》），作者 Matthew Kirk（O'Reilly 出版），书号 978-1-491-92413-6。”

如果你觉得你对示例代码的使用不在以上所述的许可范围内，请通过邮件联系我们（*permissions@oreilly.com*）。

Safari 在线图书

Safari Books Online 针对企业，政府、教育机构和个人提供了不同的购买计划，你可以根据实际需求进行选购。

用户已经访问了上千种图书、培训视频学习路径、互动教材和专业的播放列表，这些内容来自超过 250 个出版商，例如，O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Focal Press、Cisco Press、John

Wiley& Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FTPress、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology 等，关于 Safari 在线图书的更多信息，请访问 *http://oreilly.com/safari.*。

联系方式

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

我们有个关于本书的网页，上面有勘误表、示例和所有的附加信息。可以通过以下链接访问：*http://bit.ly/thoughtful-machine-learningwith-python*。

关于本书的评论和技术问题，请发邮件给 *bookquestions@oreilly.com*。

关于本书的更多信息，如教程、会议、新闻，请参见网站：

http://www.oreilly.com

http://www.oreilly.com.cn

致谢

完成本书前我等待了一年有余。我被诊断出睾丸癌，而后我的父亲突然离世，这些迫使我退而反思，然后才能重拾写作。尽管用的时间比预计的时间久，但是我对结果甚为满意。

感谢那些在本书写作过程中给予我支持的人：O'Reilly 出版社的所有帮助过我的人。编辑 Shannon Cutt 不但坚定不移。主力技术审校者 Liz Rush 始终和我一起并肩奋战。Stephen Elston 给了我许多有用的反馈。Mike Loukides 迎合我的想法并把它融入了两本已出版的书籍中。

感谢我的朋友们，尤其是 Curtis Fanta。我们从五岁就相识了。感谢他总是为了我腾

出时间（并且从来不惧怕我繁忙的日程）。

感谢我的家人。感谢我的两个侄子——好奇心强又喜欢惊叹的 Zoe 和 Darby。感谢我的兄弟 Jake，总是给我推荐新的音乐和电影，使我身心愉悦。感谢我的妈妈 Carol，她鼓励我自己寻找答案，建议我多做运动（尽管我从来不做）。他们对我来说非常重要。

感谢 Le 一家，把我当作家人来对待。感谢 Liliana 跟我相约玩乐高玩具，感谢 Sayone 和 Alyssa 给我的生活带来欢乐。感谢 Martin 和 Han 一直都给我爱和支持。感谢 Thanh（Le 的爸爸）和 Kim（Le 的妈妈）劝我吃了太多美食，还给我万用表和关于放大器的书。感谢他们成为我生活的一部分。

感谢我的奶奶，在书快要出版时一直询问我。她一直鼓励我努力奋斗，无论是童子军的时候还是当我自己做事业的时候。感谢她一直都在。

感谢我的妻子 Sophia。就在一年前，我们还在病房里，我大量地吃止痛片……现在我们挺过来了。她就是我成年以后最坚强的后盾。每当我立下雄心壮志（例如写本书）时，她总是把自己的需求放一边，把我照顾周全。她就是我的全世界。

最后，感谢我的爸爸，我怀念他来看我的时刻，还有我们在树林里露营的时光。我多希望他能在这里和我分享这一切，我珍视我们在一起的所有时光。这本书献给他。

第 1 章

可能近似正确的软件

如果你曾乘飞机飞行，那你已经体验了世界上最安全的旅行方式之一。死于飞机事故的概率是 2940 万分之一，这意味着如果你想成为一名飞行员，可以在 40 年的职业生涯中，不遭遇一次坠机。考虑到飞机是如此复杂，那些概率真是惊人的。但也并非总是如此。

2014 年对航空业来说是糟糕的一年，有 824 起航空相关的死亡，其中包括失联的马航飞机。而在 1929 年，有 257 起航空伤亡。 我们的航空业看起来变差了，但是你要意识到仅在美国每年就有超过 1000 万次航班，而在 1929 年则少得多，大约 5 万到 10 万次航班。这意味着，从 1929 年到 2014 年，飞机失事遇难的概率从 0.25% 降到了 0.00824%。

搭乘飞机旅行多年来变化很多，对于软件开发同样如此。我们知道在 1929 年的时候，软件开发尚不存在，在过去的 85 年中，我们有很多成功的软件项目，也有很多失败的软件项目。

最近的例子如 healthcare.gov 这个软件项目的推出，这是个财政灾难，因为它耗资约 6 亿 3400 万美元。更糟糕的是那些有着重大缺陷的软件项目。 2013 年由于软件故障而闭市的纳斯达克，被罚款 1000 万美元。2014 年发现的 Heartbleed bug 感染，让许多使用 SSL 的网站容易受到攻击。后来 CloudFlare 撤销了 10 余万 SSL 证书，据说花费高达数百万美元。

软件和飞机有个共同点：它们都很复杂，而且一旦出了故障，都是公共性的和灾难性的。飞机已经能确保安全飞行，减少航空事故概率超过 96%。遗憾的是，在软件行业并非