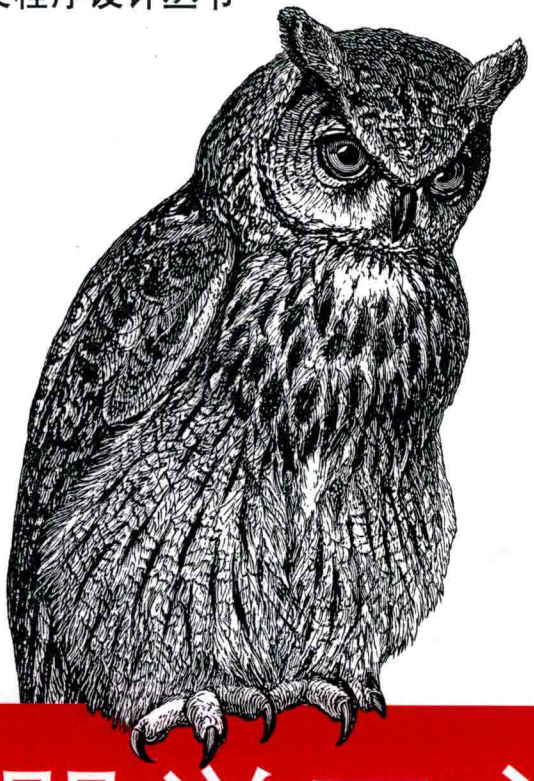
O'REILLY®

TURING 图灵程序设计丛书



机器学习实践

测试驱动的开发方法

Thoughtful Machine Learning

[美] Matthew Kirk 著
段菲 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

机器学习实践
测试驱动的开发方法

Thoughtful Machine Learning
A Test-Driven Approach

[美] Matthew Kirk 著
段菲 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北　京

图书在版编目（CIP）数据

机器学习实践 : 测试驱动的开发方法 / （美）柯克(Kirk,M.）著 ; 段菲译. -- 北京 : 人民邮电出版社, 2015.8
（图灵程序设计丛书）
ISBN 978-7-115-39618-1

Ⅰ. ①机… Ⅱ. ①柯… ②段… Ⅲ. ①机器学习－研究 Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2015)第148307号

内 容 提 要

本书主要介绍如何将测试驱动开发运用于机器学习算法。每一章都通过示例介绍了机器学习技术能够解决的有关数据的具体问题，以及求解问题和处理数据的方法。具体涵盖了测试驱动的机器学习、机器学习概述、K 近邻分类、朴素贝叶斯分类、隐马尔可夫模型、支持向量机、神经网络、聚类、核岭回归、模型改进与数据提取等内容。通过学习本书，你将能够利用机器学习技术解决涉及数据的现实问题。

本书适合开发人员、CTO 和商业分析师阅读。

◆ 著　[美] Matthew Kirk
译　段　菲
责任编辑　岳新欣
责任印制　杨林杰

◆ 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
三河市中晟雅豪印务有限公司印刷

◆ 开本：800×1000　1/16
印张：12.75
字数：302千字　2015年8月第1版
印数：1－3 500册　2015年8月河北第1次印刷
著作权合同登记号　图字：01-2015-3679号

定价：49.00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：京崇工商广字第 0021 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*

前言

这是一本介绍如何解决棘手问题的书。机器学习是计算技术的一项令人叹为观止的应用，因为它要解决的很多问题都源自科幻小说。机器学习算法可用于解决语音识别、映射、推荐以及疾病检测等复杂问题。机器学习的应用领域浩瀚无垠，也正因为如此它才这样引人入胜。

然而，这种灵活性也使得机器学习技术令人望而却步。它的确可以解决许多问题，但如何知晓我们求解的是否是正确的问题，或者是否应最先求解某个问题呢？此外，令人沮丧的是，大部分学术编码标准都不够严密。

即便是在今天，人们仍未对如何编写高质量的机器学习代码给予足够的关注，这是无比遗憾的。将一种观念在整个行业广泛传播的能力取决于有效沟通的能力。如果我们编写的代码本身就质量低劣，那么恐怕不会有很多人愿意聆听我们的讨论。

本书便是我对于该问题的回答。我试图按照容易理解的方式为大家讲授机器学习。这门学科本身难度就不小，况且我们还需要阅读代码，尤其是那些极难理解的古老 C 实现，无疑更是雪上加霜。

很多读者会对本书采用 Ruby 而非 Python 感到非常困惑。我的理由是用 Ruby 编写测试程序是一种诠释你所写代码的美妙方式。本质上，这本通篇采用测试驱动方法的书讲述的是如何沟通，具体说来是如何与机器学习这个奇妙的世界沟通。

从本书可学到的知识

应该说，本书对机器学习的介绍并不全面。因此，我向你强烈推荐 Peter Flach 编著的 *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*（Cambridge University Press）[1]。如果你希望读一些数学味道较浓的书，可参阅 Tom Mitchell 的 Machine

注 1：中文版即将由人民邮电出版社出版。——编者注

Learning 系列。此外，你还可参考 Stuart Russel 和 Peter Norvig 编写的有关人工智能的名著 *Artificial Intelligence: A Modern Approach*, 3rd Edition（Prentice Hall）。

阅读完本书之后，你并不会获得机器学习的博士学位，但我希望本书能向你传授足够的知识，帮助你开始研究如何利用机器学习技术解决涉及数据的现实问题。对于求解问题的方法以及如何在基础层面上使用它们，本书会提供大量相关示例。

你还将掌握如何解决那些比普通的单元测试更为模糊的复杂问题。

本书的阅读方法

阅读本书的最佳方法是找到一些能够让你感到兴奋的例子。我力图每一章都介绍一些这样的例子，虽然有时无法尽如人意。我不希望本书过于偏重理论，而是希望通过示例向你介绍机器学习能够解决的一些具体问题，以及我处理数据的方法。

在本书的大多数章中，我都试图在一开始便引入一个商业案例，之后开始研究一个可求解的问题，直到结尾。本书是一部短篇读物，因为我希望你能够专注于理解书中的代码，并深入思考这些问题，而非沉浸在理论当中。

本书读者

本书面向的读者包括三个群体：开发人员、CTO 以及商业分析师。

开发人员已经熟知如何编写代码，且想更多地了解机器学习这个激动人心的领域。他们拥有在计算环境中求解问题的背景，可能使用过 Ruby 编写代码，也可能从未接触过这种语言。他们是本书的主要读者群，但本书在写作时也兼顾了 CTO 和商业分析师。

CTO 是那些真正希望了解如何利用机器学习技术提升公司业务的人。他们可能听说过 *K* 均值算法、*K* 近邻算法，但不知道这些算法的适用性。商业分析师与之相似，只是不像 CTO 那样关注技术细节。为了这两个读者群，我在每章的开头都准备了一个商业案例。

作者联系方式

如果你喜欢我之前所做的演讲，或想为某个问题寻求帮助，欢迎通过电子邮件与我联系。我的邮箱是 matt@matthewkirk.com。为了增进我们之间的联系，如果你来到西雅图地区，且我们的日程允许的话，我非常乐意请你喝一杯咖啡。

如果希望查看本书的所有代码，可从 GitHub 站点 http://github.com/thoughtfulml 免费下载。

排版约定

本书使用了下列排版约定。

- 楷体
  表示新术语或突出强调的内容。
- 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。
- 加粗等宽字体（**`constant width bold`**）
  表示应该由用户输入的命令或其他文本。

该图标表示提示或建议。

该图标表示一般注记。

该图标表示警告或警示。

该图标表示非常重要的警告，请仔细阅读。

使用代码示例

补充材料（代码示例、练习等）可以从 http://github.com/thoughtfulml 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或文档中。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回答问题无需获得许可，将书中大量的代码放到你的产品文档中则需要获得许可。

我们很希望但并不强制要求你在引用本书内容时加上引用说明。引用说明一般包括书名、作者、出版社和 ISBN。比如："*Thoughtful Machine Learning* by Matthew Kirk (O'Reilly). Copyright 2015 Matthew Kirk, 978-1-449-37406-8"。

如果你觉得自己对示例代码的用法超出了上述许可的范围，欢迎你通过 permissions@oreilly.com 与我们联系。

Safari® Books Online

Safari Books Online（http://www.safaribooksonline.com）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务作家的专业作品。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。

对于组织团体、政府机构和个人，Safari Books Online 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 以及其他几十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，我们网上见。

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

O'Reilly 的每一本书都有专属网页，你可以在那儿找到本书的相关信息，包括勘误表、示例代码以及其他信息。本书的网站地址是：

http://shop.oreilly.com/product/0636920032298.do

对于本书的评论和技术性问题，请发送电子邮件到：bookquestions@oreilly.com

要了解更多 O'Reilly 图书、培训课程、会议和新闻的信息，请访问以下网站：

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

- Mike Loukides，对于我的利用测试驱动开发来编写机器学习代码的想法，他表现出了浓厚的兴趣。
- 我的编辑 Ann Spencer，在我撰写本书的数月间，她教给了我大量编辑知识和技巧，并针对本书的设计提出了宝贵意见。

我要感谢 O'Reilly 团队的所有成员，是他们成就了本书。我尤其要感谢下面这些人。

我的审稿人：

- Brad Ediger，当我提出要编写一本关于基于测试驱动的机器学习代码的书籍时，他对于这一古怪想法兴奋不已，并对本书的初稿提出了大量极有价值的意见。
- Starr Horne，他在审阅过程中向我提供了大量真知灼见。感谢他在电话会议中与我讨论了机器学习、错误报告等。
- Aaron Sumner，他针对本书的代码结构提出了极有价值的意见。

我极为出色的合作者，以及在本书编写过程中提供帮助的朋友们：Edward Carrel、Jon-Michael Deldin、Christopher Hobbs、Chris Kuttruff、Stefan Novak、Mike Perham、Max Spransy、Moxley Stratton 以及 Wafa Zouyed。

如果没有家人的支持和鼓励，我不可能顺利完成本书。

- 献给我的妻子 Sophia，是她帮助我坚持梦想，并帮助我将对本书的设想变成现实。
- 献给我的祖母 Gail，在我幼年时，是她引导我对学习产生了浓厚的兴趣。我还记得在一次公路旅行中，她心无旁骛地向我询问一些关于我正在看的那本“咖啡书”（实际上是讲 Java 的书）的问题。

- 献给我的父亲 Jay 和母亲 Carol，是他们教会了我如何分析系统，以及如何为其注入人类的情感。
- 献给我的弟弟 Jacob 以及侄女 Zoe 和 Darby，感谢他们教会了我如何通过孩子的眼光重新认识世界。

最后，我将本书献给科学和对知识不断求索的人。

目录

第1章

测试驱动的机器学习

伟大的科学家既是梦想家，也是怀疑论者。在现代历史中，科学家们取得了一系列重大突破，如发现地球引力、登上月球、发现相对论等。所有这些科学家都有一个共同点，那就是他们都有着远大的梦想。然而，在完成那些壮举之前，他们的工作无不经过了周密的检验和验证。

如今，爱因斯坦和牛顿已离我们而去，但所幸我们处在一个大数据时代。随着信息时代的到来，人们迫切需要找到将数据转化为有价值的信息的方法。这种需求的重要性已日益凸显，而这正是数据科学和机器学习的使命。

机器学习是一门充满魅力的学科，因为它能够利用信息来解决像人脸识别或笔迹检测这样的复杂问题。很多时候，为完成这样的任务，机器学习算法会采用大量的测试。典型的测试包括提出统计假设、确定阈值、随着时间的推移将均方误差最小化等。理论上，机器学习算法具备坚实的理论基础，可从过去的错误中学习，并随时间的推移将误差最小化。

然而，我们人类却无法做到这一点。机器学习算法虽能将误差最小化，但有时我们可能“指挥失误”，没能令其将“真正的误差”最小化，我们甚至可能在自己的代码中犯一些不易察觉的错误。因此，我们也需要通过一些测试来发现自己所犯的错误，并以某种方式来记录我们的进展。用于编写这类测试的最为流行的方法当属*测试驱动开发*（Test-Driven Development，TDD）。这种“测试先行”的方法已成为编程人员的一种最佳实践。然而，这种最佳实践有时在开发环境中却并未得到运用。

采用驱动测试开发（为简便起见，下文中统称 TDD）有两个充分的理由。首先，虽然在主动开发模式中，TDD 需要花费至少 15%~35% 的时间，却能够排除多达 90% 的程序缺陷