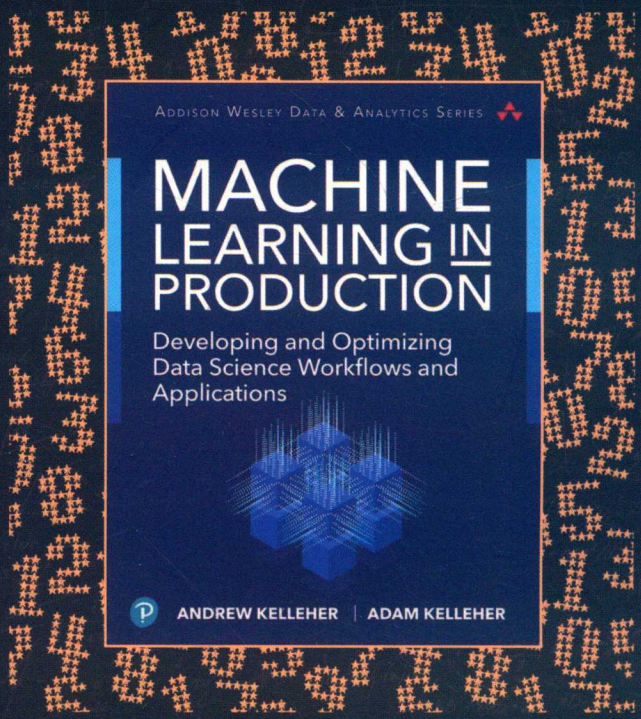
HZ BOOKS 华章IT
Pearson
数据科学与工程技术丛书

机器学习实践

数据科学应用与工作流的开发及优化

[美] 安德鲁·凯莱赫（Andrew Kelleher）
亚当·凯莱赫（Adam Kelleher）
著
陈子墨 刘瀚文 译



MACHINE LEARNING IN PRODUCTION

DEVELOPING AND OPTIMIZING DATA SCIENCE WORKFLOWS AND APPLICATIONS

机械工业出版社
China Machine Press

机器学习实践
数据科学应用与工作流的开发及优化

MACHINE LEARNING IN PRODUCTION
DEVELOPING AND OPTIMIZING DATA SCIENCE WORKFLOWS AND APPLICATIONS

本书可作为数据科学与机器学习速成课程的参考教材，面向需要在生产环境中解决实际问题的技术人员。两位作者展示了如何快速交付重要的生产价值，如何持续极大化投资回报率，避免使用被过度夸大的工具和不必要的复杂性，利用极简单、低风险的方法来达成目的。

作者利用他们丰富的经验，帮助你提出十分有用的问题，从无到有地完成你的生产项目。书中展示了利用简单的查询、聚合和可视化方法可以做些什么，并且讲述了不可或缺的误差分析方法来帮助你避免做出错误结论。全书涵盖了主要的机器学习方法，比如线性回归、随机森林、分类、聚类以及贝叶斯推断，这些能够帮助你在面对实际问题时选择正确的算法。本书后面关于硬件、架构、分布式系统的章节对如何在生产环境中优化性能提供了非常宝贵的参考。

通过学习本书，你将能够：

- 利用敏捷原则缩小项目范围，保持高效开发。
- 从实用Python代码示例中学习。
- 从简单的启发式方法开始，并随着数据管道的成熟而改进它们。
- 利用基本的数据可视化技巧来表达你的结果。
- 精通主要的机器学习方法，包括线性回归、随机森林、分类、聚类与过拟合。
- 学习图模型与贝叶斯推断的基础。
- 理解机器学习模型中的相关性与因果性。

扫码查看更多数字资源

Pearson
www.pearson.com

上架指导：计算机\人工智能
ISBN 978-7-111-65136-9
9 787111 651369
定价：99.00元

投稿热线：(010) 88379604
读者信箱：hzit@hzbook.com
客服电话：(010) 88361066 88379833 68326294

华章网站：www.hzbook.com
网上购书：www.china-pub.com
数字阅读：www.hzmedia.com.cn

数据科学与工程技术丛书

MACHINE LEARNING IN PRODUCTION

DEVELOPING AND OPTIMIZING DATA SCIENCE WORKFLOWS AND APPLICATIONS

机器学习实践

数据科学应用与工作流的开发及优化

[美] 安德鲁·凯莱赫（Andrew Kelleher）
亚当·凯莱赫（Adam Kelleher） 著

陈子墨 刘瀚文 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

机器学习实践：数据科学应用与工作流的开发及优化 /（美）安德鲁·凯莱赫（Andrew Kelleher），（美）亚当·凯莱赫（Adam Kelleher）著；陈子墨，刘瀚文译．—北京：机械工业出版社，2020.4
（数据科学与工程技术丛书）
书名原文：Machine Learning in Production: Developing and Optimizing Data Science Workflows and Applications

ISBN 978-7-111-65136-9

I. 机… II. ①安… ②亚… ③陈… ④刘… III. 机器学习－研究 IV. TP181

中国版本图书馆 CIP 数据核字（2020）第 048227 号

本书版权登记号：图字 01-2019-6632

机器学习实践
数据科学应用与工作流的开发及优化

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：李忠明　　责任校对：殷　虹
印　　刷：大厂回族自治县益利印刷有限公司　　版　　次：2020 年 4 月第 1 版第 1 次印刷
开　　本：185mm×260mm　1/16　　印　　张：15.25
书　　号：ISBN 978-7-111-65136-9　　定　　价：99.00 元

客服电话：（010）88361066　88379833　68326294　　投稿热线：（010）88379604
华章网站：www.hzbook.com　　读者信箱：hzit@hzbook.com

内容简介

这本实用书籍同时介绍了机器学习和数据科学，填补了数据科学家和工程师之间的空白，并帮助将这些技术应用于生产。它致力于确保你做的努力能够真正解决你的问题，并覆盖了真实世界生产环境设置中的性能优化问题。

——Paul Dix，丛书编辑

本书共分三部分，19章。第一部分（第1~6章）介绍框架原则，涵盖数据科学领域的背景知识、项目工作流程及其与敏捷开发原则的关系、误差测量的概念和量化、数据编码与预处理、统计假设检验、数据可视化和探索性数据分析。第二部分（第7~14章）描述算法和架构，包括算法和架构的概述、相似性度量方法、有监督机器学习、离散的有监督模型和无监督机器学习的基础知识、贝叶斯网络和贝叶斯模型、因果推断，以及高级机器学习技术。第三部分（第15~19章）讲解瓶颈和优化，涵盖硬件方面的基本瓶颈、软件设计的基础知识、分布式系统中的体系结构模式、CAP定理，以及逻辑网络拓扑节点。

作者简介

安德鲁·凯莱赫

（Andrew Kelleher）

Venmo的一名高级软件工程师和分布式系统架构师。他毕业于克莱姆森大学并获得物理学学士学位，曾任BuzzFeed的软件工程师，关注现代优化的数据管道和算法实现。

亚当·凯莱赫

（Adam Kelleher）

BuzzFeed的首席数据科学家，曾是巴克莱银行的首席研究数据科学家，并在哥伦比亚大学教授因果推断和机器学习产品课程。他毕业于克莱姆森大学，获得物理学学士学位，并在北卡罗来纳大学教堂山分校获得宇宙学博士学位。

译者简介

陈子墨

原ThoughtWorks数据分析师，负责机器学习方案搭建与实验。现为PayPal数据科学家，主要负责支付风险定量分析、反欺诈建模与决策方案优化。

刘瀚文

ThoughtWorks算法工程师，专注为大型企业提供机器学习平台及算法的研发和咨询服务，深谙工业级机器学习之道。

译 者 序

不管你的职业是什么，如果你在工作中会遇到真实世界的数据科学问题，那么本书将会对你提供巨大的帮助。它不仅描绘了广阔的机器学习算法世界，还教导你如何用合适的工程方法在其中翱翔。除了数学公式和图表，本书切合实际的代码和检验方法将有助于确保你专注于解决问题本身，而非研究高深莫测的算法理论。

两位作者——安德鲁·凯莱赫（Andrew Kelleher）和亚当·凯莱赫（Adam Kelleher）在工作中分别扮演着数据科学家和工程师的角色，默契的兄弟俩将机器学习和计算机工程巧妙地结合在一起，基于在 BuzzFeed 的工作经验，写出了这本机器学习工程指南。第一部分介绍的框架原则是数据科学世界坚实的基础；第二部分介绍解决现实问题的常用算法，帮助读者迅速解决实际问题，以及避免被数据误导，产生结论错误；第三部分则着眼于工程实践，基于工程角度突破瓶颈，让算法能够在现实条件中得以实现。

因本书着眼于利用数据科学解决实际问题，所以无论你是初学者还是经验丰富的工程师，都能受益良多。

序

这本实用书籍同时介绍了机器学习和数据科学，填补了数据科学家和工程师之间的空白，并帮助将这些技术应用于生产。它致力于确保你做的努力能够真正解决你的问题，并覆盖了真实世界生产环境设置中的性能优化问题。本书包含 Python 代码示例和可视化示例来解释算法中的概念。验证、假设检验和可视化的部分在本书开始就引入了，以确保你在数据科学上的努力能够真正解决问题。本书的第三部分在数据科学和机器学习书籍中是独一无二的，因为它侧重于现实世界对性能优化的关注。思考硬件、基础设施和分布式系统都是将机器学习和数据科学技术引入生产实践的步骤。

安德鲁·凯莱赫（Andrew Kelleher）和亚当·凯莱赫（Adam Kelleher）分别总结了他们在 BuzzFeed 工作时在工程领域和数据科学方面的经验，他们在大型生产环境中解决问题的实际经验为本书所涉及的主题以及在何内容上提供广度或深度提供了依据。本书介绍了用于比较、分类、聚类和降维的算法，并分别提供了可以解决特定问题的示例。在奠定了基本机器学习任务的框架之后，将提供对更高阶主题（如贝叶斯网络或深度学习）的探索。

本书提供了对数据科学和机器学习的充分介绍，关注于解决实际问题。对于那些希望将机器学习应用于其生产环境的具有传统数学或科学背景的任何工程师或“意外程序员”来说，本书是一个很好的资源。

——保罗·迪克斯

前　言

本书大部分内容是 Andrew 和 Adam 一起在 BuzzFeed 工作时写的。Adam 是数据科学家，Andrew 是工程师，他们在同一个团队中工作了很长时间。最让人感到惊奇和有趣的是，他俩不只是工作伙伴，还是三胞胎中的一对兄弟。

写这本书的想法是 2014 年 8 月我们参加了纽约的 PyGotham[⊖]之后产生的。当时有好几场相对广义的关于“数据科学”的讨论，我们发现许多数据科学家的职业生涯始于对事物的好奇心和学习新事物的兴奋感。他们会发现一些新工具，在这之中发展出自己偏爱使用的某种技术或算法，然后将这些工具应用到他们正在处理的问题上。每个人都喜欢用自己最熟悉的方式去解决问题，这种做法很高效。比如使用神经网络（我们将会在第 14 章中讨论），因为它是一个更为高效的解决工具。我们想通过为数据科学家，尤其是初入职场的新人提供一个完整的工具箱，从而推动数据科学的发展。有人可能会质疑，第一部分的内容和误差分析实际上比第三部分讨论的技术更重要。但实际上第三部分才是我们写这本书的动力。如果数据集中充斥着大量噪声或系统误差，那么算法几乎是不可能成功的。我们希望这本书可以提供一些正确的参考来帮助读者解决在实际项目中遇到的问题，从而帮助他们在职业生涯中取得成功。

机器学习领域、计算机科学领域甚至数据科学领域不乏好书，但我们希望本书可以作为一本比较严谨、全面的数据科学入门书籍。这是一本根据我们自身实践经验写成的轻量级工具书，我们尽可能规避了研究型的问题。假如作为一名初级数据科学家，你正在解决研究型问题，那这可能已经超出了我们关心的范围。

数据科学有一个与机器学习分开的关键部分，那就是工程学。这一点我们会在第三部分着重讨论。我们会讨论你有可能遇到的问题并提供解决它们所需要的基础知识。可以这么说，第三部分基本上可作为计算机科学速成课程（初级课程）参考。因为即使你知道在开发什么，但在落实到生产的路上依然有很多注意事项，这意味着必须要理解这

⊖ Python 社区在纽约举办的一个以 Python 为主题的大会。——译者注

些知识本身，而不仅仅是把它们当作某种工具。

本书受众

在过去几年优秀工程师一直有很大缺口。2008 年在一个会议上我们第一次听到了“意外程序员”这个词。它用来描述那些不是科班出身的工程师——他们只是误打误撞到了那个位置并开始做相关工作。十多年后的今天对于开发人员依然有大量需求，并且这种需求开始逐渐扩展到数据科学家这个职位上。谁将充当“意外数据科学家”的角色？通常情况下是开发人员或者是物理或数学专业本科生，虽然他们没有接受过太多数据科学家所需的正规培训，但拥有成功所需的好奇心和雄心，对工具箱有需求。

本书旨在打造一套速成课程，通过从头到尾过一遍数据项目的基本开展步骤来鼓励数据科学家使用手里的数据而非工具，并以此作为起点。由数据本身驱动的数据科学是成功的关键。数据科学最大的公开秘密就是，虽然建模很重要，但数据科学最基础的日常工作依然是数据的查询、聚合和可视化。许多行业仍然处在收集和使用数据的比较原始的阶段，因此快速交付一些复杂度较低的东西是非常有意义的。

建模很重要，但也很难。我们相信敏捷开发的原则是可以应用到数据科学中的，我们将在第 2 章中讨论这一点。比如我们可以从最小的解决方案开始，有一个基于聚合数据的点子，当数据管道稳定且成熟的时候套用一些模型慢慢延伸它，然后在你手头没有那么多别的重要的事情时慢慢改进模型。我们会提供基于此方法的真实案例。

本书内容

在开头我们提供了一些数据科学领域的基本背景。第一部分的第 1 章是了解数据行业的引子。

第 2 章将数据科学置于敏捷开发流程下考虑，这是一种有助于保持小范围有效开发的理念。让自己不去尝试最新的机器学习框架或基于云平台的工具很难，但从长远来看是值得的。

第 3 章提供了关于误差分析的基本介绍。许多数据科学都在做一些简单的统计报告，如果不理解统计误差，则很有可能会得出无效的结论。误差分析是一项基本技能，并且是一项必备技能。

第 4 章提供了一些编码现实世界数据的方法。这会让我们提出一些现实世界中被数据驱动的问题。回答这类问题的框架是假设检验，我们会在第 5 章中说明。

到现在为止我们还没有看到很多图表，所以还缺乏将分析结果与外部（非技术）世

界沟通的渠道。我们会在第 6 章中解决这个问题。我们会把讨论限定在比较小的范围，主要针对那些我们知道如何计算误差的数量图，或者那些使数据可视化产生细微差别的图。虽然这些工具不像 d3 的交互式可视化图那样酷炫（d3 非常值得学习），但它们也是与非技术人员沟通的基础。

在介绍了基本的数据处理方法之后，我们将继续研究更高级的概念，也就是第二部分。我们首先在第 7 章中简要介绍数据结构，然后在第 8 章中介绍机器学习的基本概念。到这时候你已经有了一些可以上手的方法来衡量对象的相似性。

从现在开始我们已经可以进行简单的机器学习了。第 9 章中，我们开始引入回归的概念并从一个最重要的模型线性回归开始。在如今这个神经网络和非线性机器学习时代，从介绍这种简单模型开始确实有些奇怪，但线性回归绝对是一个相当优秀的模型。正如稍后将详述的那样，它是可解释的、稳定的，能提供一个非常好的基准。另外，通过一些小技巧，它也可以用于非线性情况，并且最近的研究结果表明，多项式回归（线性回归的简单变形）在一些应用中的表现甚至可以胜过深度前馈网络！

接下来我们还描述了回归模型中的另一个主力模型：随机森林。随机森林依赖“bagging”技术，这是一种基于统计技巧的非线性算法，可以为各种不同的问题提供出色的基准。如果想要一个简单的模型来开始项目并且线性回归不太合适，那么随机森林是一个不错的候选。

在介绍了回归并提供了一些机器学习工作流程的基本案例之后，将继续学习第 10 章。有很多方法都适用于向量和图形数据，我们在这部分提供关于图的基本背景知识和贝叶斯推断的简要介绍。在下一章我们会深入研究贝叶斯推断和因果关系。

第 11 章的内容既非常规又比较难。从因果关系的角度来看，贝叶斯网络是最直观（尽管不一定最简单）的因果图。因此我们引入贝叶斯网络的基础介绍并把它作为理解因果推断的基础。第 12 章中，我们以基础贝叶斯网络理解 PCA 和潜在因子模型的其他变体。主题建模是隐变量模型的一个重要例子，我们提供了一个基于新数据集的详细例子。

作为下一个以数据为中心的章节，我们将重点放在第 13 章中的因果推断问题上。它的重要性是无法低估的。数据科学通常的目标是告知企业如何行事，假设数据能告诉你某个行为的结果，只有当分析出因果关系而不仅仅是相关关系时，这个结果才会成立。从这个意义上说，理解因果关系是数据科学家工作的基础。不幸的是，为了尽量保持工作范围最小化，它也常常第一个被削减。在规划项目时，平衡利益相关者的期望是很重要的，而因果推断工作可能需要花一些时间。我们希望让数据科学家做出明智的决策，而不是轻易接受相关结果。

在最后一个以数据为中心的章节（第 14 章）中，我们提供了更先进的机器学习技术

的一些细微差别。我们使用神经网络作为讨论过拟合和模型能力的工具。重点应放在尽可能使用简单的解决方案，抵制以神经网络作为第一模型开始的冲动。简单的回归方法几乎总能为第一个解决方案提供足够好的基线。

到目前为止，我们介绍的都是背景知识，这是开始数据科学项目的起点，但不是我们的主要关注点，至少现在不是。本书的第三部分也是最后一部分将深入研究硬件、软件及其组成的系统。

第 15 章首先全面介绍计算机硬件。该章介绍一个我们日常会用的基本资源的工具箱，并提供一个框架来讨论我们在实际操作中受到的约束。这些约束是可能的物理限制，以及这些限制在硬件中的实现。

第 16 章提供了软件的基础知识和数据传输的基本描述，其中一节讨论“提取 – 传输 / 转换 – 加载”，通常称为 ETL。

接下来，我们在第 17 章中概述了软件架构的设计注意事项。架构是整个系统如何组合在一起的设计。它包括用于数据存储、数据传输和计算的组件，以及它们之间如何相互通信。有些架构比其他架构更有效率，并且客观上也比其他架构做得更好。但是，鉴于时间和资源的限制，效率较低的解决方案可能更实用。我们希望提供足够的上下文，以便你可以做出明智的决定。即使你是数据科学家而不是工程师，我们也希望提供足够的知识，让你至少可以了解数据平台的状况。

然后，我们继续研究工程学中的一些更高阶的主题。第 18 章涵盖了数据库性能的一些基本界限。最后，在最后一章（第 19 章）讨论网络拓扑时，我们讨论了所有元素如何组合在一起。

继续

我们希望你不仅可以运用数据科学中的机器学习这部分，还可以了解自己数据平台的局限性。这样你才可以了解你需要构建什么，并找到按需构建基础设施的有效途径。我们希望借助完整的工具箱，你可以最终意识到这些工具只是解决方案的一部分。它们是解决实际问题的一种手段，而实际问题总是会受到资源的限制。

如果要从本书中吸取教训，那就是你应该始终将资源用于解决投资回报率最高的问题。解决你的问题是一个真正的约束。有时候，最好的机器学习模型无法解决所有问题。那这时候要问的问题是，这个就是要解决的最佳问题，还是有一个更简单的、风险更低的任务。

最后，尽管我们希望本书能涉及生产类机器学习的所有方面，但目前它更像是一本生产类数据科学书籍。在后续版本中，我们打算涵盖本版遗漏的内容，尤其是在机器学习基础设施方面。新的资料将包括：并行模型训练和预测的方法；Tensorflow、Apache

Airflow、Spark 以及其他框架和工具的基础知识；几个真正的机器学习平台的详细信息，包括 Uber 的 Michelangelo、Google 的 TFX 和我们自己在类似系统上的工作；以及避免和处理机器学习系统中的耦合。我们鼓励读者同时搜索涉及这些主题的书籍、论文和博客文章，并在本书的网站（adamkelleher.com/ml_book）上查看更新。

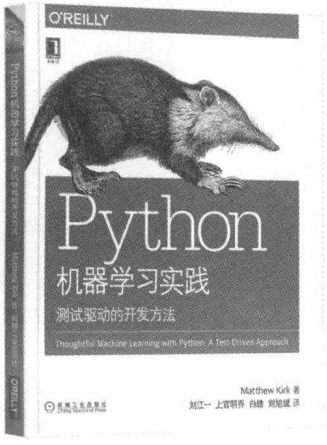
希望你会像我们一样喜欢学习这些工具，并且希望这本书可以节省你的时间和精力。

作者简介

安德鲁·凯莱赫（Andrew Kelleher）是 Venmo 的一名高级软件工程师和分布式系统架构师。他以前是 BuzzFeed 的一名高级软件工程师，并且致力于数据管道和算法实现的最新优化。他毕业于克莱姆森大学，在那里获得物理学学士学位。他在纽约市举行了一次研讨会，研究了在生产应用环境中分布式系统背后的基础知识，并连续两年被评为 FastCompany 最具创造力的人之一。

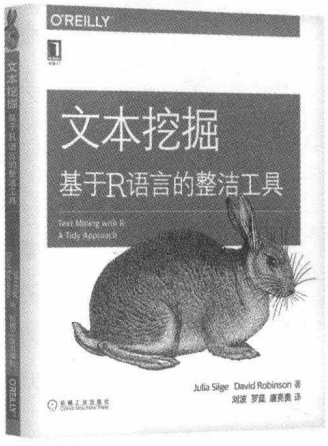
亚当·凯莱赫（Adam Kelleher）在 BuzzFeed 担任首席数据科学家，且在纽约哥伦比亚大学做兼职教授期间写下了这本书。截至 2018 年 5 月，他是巴克莱银行的首席研究数据科学家，并在哥伦比亚大学教授因果推断和机器学习产品。他毕业于克莱姆森大学，获得物理学学士学位，并在北卡罗来纳大学教堂山分校获得宇宙学博士学位。

推荐阅读



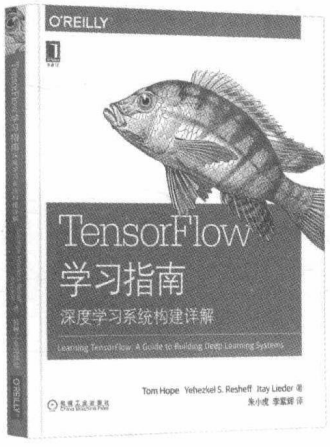
Python机器学习实践：测试驱动的开发方法

作者：Matthew Kirk ISBN：978-7-111-58166-6 定价：59.00元



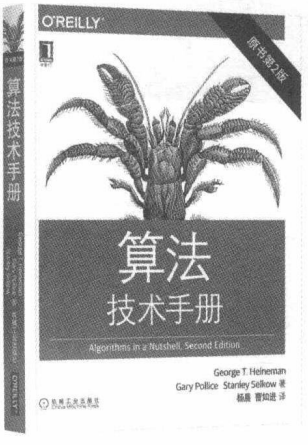
文本挖掘：基于R语言的整洁工具

作者：Julia Silge , David Robinson ISBN：978-7-111-58855-9 定价：59.00元



TensorFlow学习指南：深度学习系统构建详解

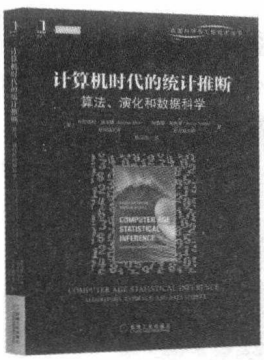
作者：Tom Hope, Yehezkel S. Resheff, Itay Lieder ISBN：978-7-111-60072-5 定价：69.00元



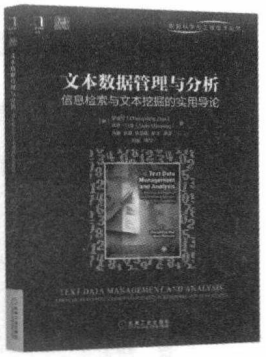
算法技术手册（原书第2版）

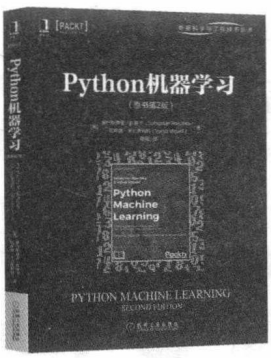
作者：George T. Heineman等 ISBN：978-7-111-56222-1 定价：89.00元

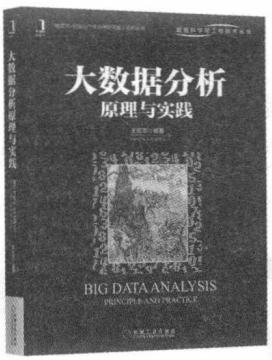
推荐阅读

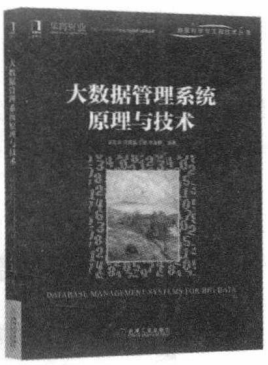

计算机时代的统计推断
算法、演化和数据科学
COMPUTER AGE STATISTICAL INFERENCE

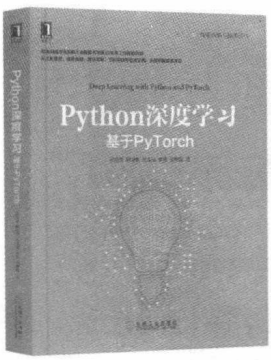

统计反思
用R和Stan例解贝叶斯方法

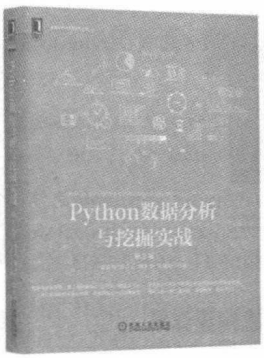

文本数据管理与分析
信息检索与文本挖掘的实用导论
TEXT DATA MANAGEMENT AND ANALYSIS

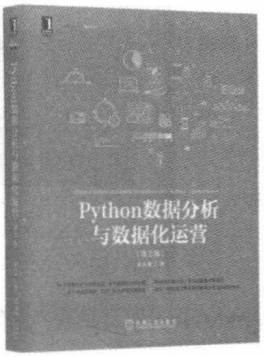

Python机器学习
PYTHON MACHINE LEARNING


大数据分析
原理与实践
BIG DATA ANALYSIS
PRINCIPLES AND PRACTICE


大数据管理系统
原理与技术


Python深度学习
基于PyTorch


Python数据分析
与挖掘实战


Python数据分析
与数据化运营

目　录