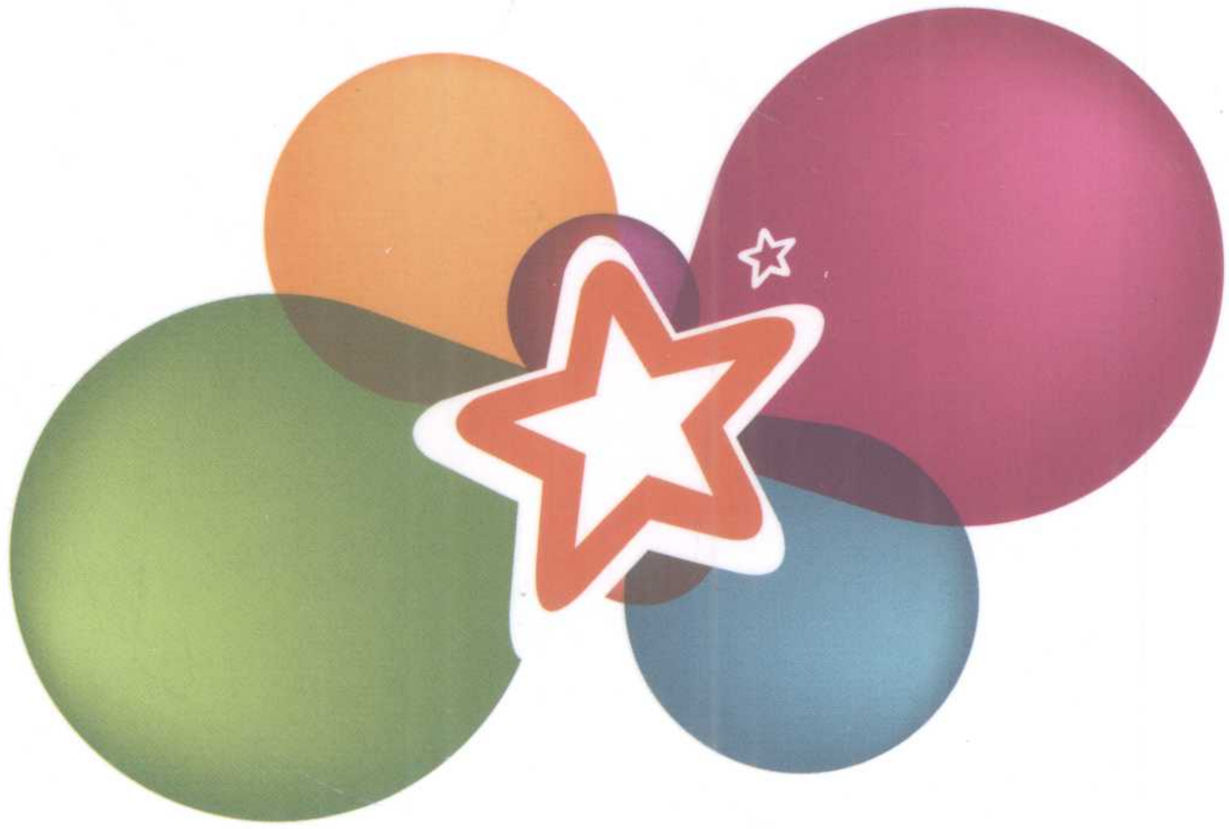
MLlib是Apache Spark机器学习库。本书入门简单，实例丰富，详解协同过滤、线性回归、分类、决策树、保序回归、聚类、关联、数据降维、特征提取和转换等MLlib主要算法，用实例说明MLlib大数据机器学习算法的运用。



Spark MLlib Machine Learning Practice

# Spark MLlib

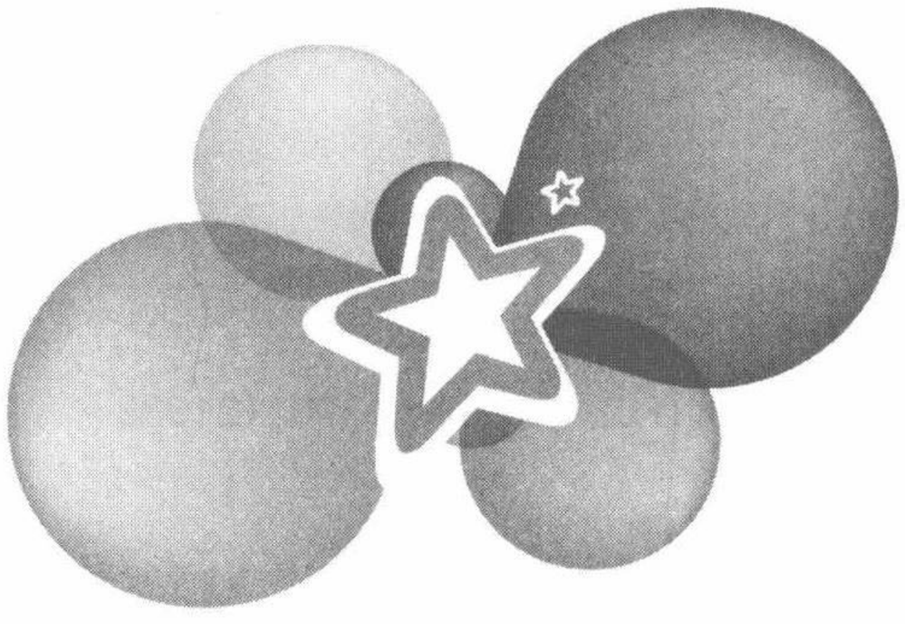
# 机器学习实践

（第2版）

王晓华 著

清華大學出版社

本书示例
源代码下载



# Spark MLlib
# 机器学习实践
（第2版）

王晓华 著

清华大学出版社
北京

# 内 容 简 介

Spark 作为新兴的、应用范围最为广泛的大数据处理开源框架引起了广泛的关注，它吸引了大量程序设计和开发人员进行相关内容的学习与开发，其中 MLlib 是 Spark 框架使用的核心。本书是一本细致介绍 Spark MLlib 程序设计的图书，入门简单，示例丰富。

本书分为 13 章，从 Spark 基础安装和配置开始，依次介绍 MLlib 程序设计基础、MLlib 的数据对象构建、MLlib 中 RDD 使用介绍，各种分类、聚类、回归等数据处理方法，最后还通过一个完整的实例，回顾了前面的学习内容，并通过代码实现了一个完整的分析过程。

本书理论内容由浅而深，采取实例和理论相结合的方式，讲解细致直观，适合 Spark MLlib 初学者、大数据分析和挖掘人员，也适合高校和培训学习相关专业的师生教学参考。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售
版权所有，侵权必究。侵权举报电话：010-62782989　13701121933

图书在版编目（CIP）数据

Spark MLlib 机器学习实践 / 王晓华著. — 2 版. —北京：清华大学出版社，2017
ISBN 978-7-302-46508-9

I. ①S… II. ①王… III. ①数据处理软件—机器学习 IV. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 025411 号

**责任编辑**：夏毓彦
**封面设计**：王　翔
**责任校对**：闫秀华
**责任印制**：李红英

**出版发行**：清华大学出版社
**网　　址**：http://www.tup.com.cn，http://www.wqbook.com
**地　　址**：北京清华大学学研大厦 A 座　　**邮　　编**：100084
**社 总 机**：010-62770175　　**邮　　购**：010-62786544
**投稿与读者服务**：010-62776969，c-service@tup.tsinghua.edu.cn
**质 量 反 馈**：010-62772015，zhiliang@tup.tsinghua.edu.cn
**印 装 者**：清华大学印刷厂
**经　　销**：全国新华书店
**开　　本**：190mm×260mm　　**印　张**：12.75　　**字　　数**：326 千字
**版　　次**：2015 年 12 月第 1 版　2017 年 3 月第 2 版　　**印　　次**：2017 年 3 月第1 次印刷
**印　　数**：1～3000
**定　　价**：49.00 元

产品编号：073286-01

# 前　言

Spark 在英文中是火花的意思，创作者希望它能够像火花一样点燃大数据时代的序幕。它，做到了。

大数据时代是一个充满着机会和挑战的时代，就像一座未经开发的金山，任何人都有资格去获得其中的宝藏，仅仅需要的就是有一把得心应手的工具——MLlib 就是这个工具。

## 本书目的

本书的主要目的是介绍如何使用 MLlib 进行数据挖掘。MLlib 是 Spark 中最核心的部分，它是 Spark 机器学习库，经过无数创造者卓越的工作，MLlib 已经成为一个优雅的、可以运行在分布式集群上的数据挖掘工具。

MLlib 充分利用了现有数据挖掘的技术与手段，将隐藏在数据中不为人知，但又包含价值的信息从中提取出来，并通过相应的计算机程序，无须人工干预自动地在系统中进行计算，以发现其中的规律。

通常来说，数据挖掘的难点和重点在于两个方面：分别是算法的学习和程序的设计。还有的是需要使用者有些相应的背景知识，例如统计学、人工智能、网络技术等。本书在写作上以工程实践为主，重点介绍其与数据挖掘密切相关的算法与概念，并且使用浅显易懂的语言将其中涉及的算法进行概括性描述，从而可以帮助使用者更好地了解和掌握数据挖掘的原理。

作者在写作本书的时候有一个基本原则，这本书应该体现工程实践与理论之间的平衡。数据挖掘的目的是为了解决现实中的问题，并提供一个结果，而不是去理论比较哪个算法更高深，看起来更能吓唬人。本书对算法的基本理论和算法也做了描述，如果读者阅读起来觉得困难，建议找出相应的教材深入复习一下，相信大多数的读者都能理解相关的内容。

## 本书内容

本书主要介绍 MLlib 数据挖掘算法，编写的内容可以分成三部分：第一部分是 MLlib 最基本的介绍以及 RDD 的用法，包括第 1~4 章；第二部分是 MLlib 算法的应用介绍，包括第 5~12 章；第三部分通过一个经典的实例向读者演示了如何使用 MLlib 去进行数据挖掘工作，

为第 13 章。

各章节内容如下：

第 1 章主要介绍了大数据时代带给社会与个人的影响，并由此产生的各种意义。介绍了大数据如何深入到每个人的生活之中。MLlib 是大数据分析的利器，能够帮助使用者更好地完成数据分析。

第 2 章介绍 Spark 的单机版安装方法和开发环境配置。MLlib 是 Spark 数据处理框架的一个主要组件，因此其运行必须要有 Spark 的支持。

第 3 章是对弹性数据集（RDD）进行了讲解，包括弹性数据集的基本组成原理和使用，以及弹性数据集在数据处理时产生的相互依赖关系，并对主要方法逐一进行示例演示。

第 4 章介绍了 MLlib 在数据处理时所用到的基本数据类型。MLlib 对数据进行处理时，需要将数据转变成相应的数据类型。

第 5 章介绍了 MLlib 中协同过滤算法的基本原理和应用，并据此介绍了相似度计算和最小二乘法的原理和应用。

第 6~12 章每章是一个 MLlib 分支部分，其将 MLlib 各个数据挖掘算法分别做了应用描述，介绍了其基本原理和学科背景，演示了使用方法和示例，对每个数据做了详细的分析。并且在一些较为重要的程序代码上，作者深入 MLlib 源码，研究了其构建方法和参数设计，从而帮助读者更深入地理解 MLlib，也为将来读者编写自有的 MLlib 程序奠定了基础。

第 13 章是本文的最后一章，通过经典的鸢尾花数据集向读者演示了一个数据挖掘的详细步骤。从数据的预处理开始，去除有相关性的重复数据，采用多种算法对数据进行分析计算，对数据进行分类回归，从而最终得到隐藏在数据中的结果，并为读者演示了数据挖掘的基本步骤与方法。

## 本书特点

- 本书尽量避免纯粹的理论知识介绍和高深技术研讨，完全从应用实践出发，用最简单的、典型的示例引申出核心知识，最后还指出了通往“高精尖”进一步深入学习的道路；
- 本书全面介绍了 MLlib 涉及的数据挖掘的基本结构和上层程序设计，借此能够系统地看到 MLlib 的全貌，使读者在学习的过程中不至于迷失方向；
- 本书在写作上浅显易懂，没有深奥的数学知识，采用了较为简洁的形式描述了应用的理论知识，让读者轻松愉悦地掌握相关内容；
- 本书旨在引导读者进行更多技术上的创新，每章都会用示例描述的形式帮助读者更好地学习内容；



试读结束：需要全本请在线购买： www.ertongbook.com

- 本书代码遵循重构原理，避免代码污染，引导读者写出优秀的、简洁的、可维护的代码。

## 读者与作者

- 准备从事或者从事大数据挖掘、大数据分析的工作人员
- Spark MLlib 初学者
- 高校和培训学校数据分析和处理相关专业的师生

本书由王晓华主编，其他参与创作的作者还有李阳、张学军、陈士领、陈丽、殷龙、张鑫、赵海波、张兴瑜、毛聪、王琳、陈宇、生晖、张喆、王健，排名不分先后。

## 示例代码下载

本书示例代码可以从下面地址（注意数字和字母大小写）下载：

http://pan.baidu.com/s/1hqtuutY

如果下载有问题，请联系电子邮箱 booksaga@163.com，邮件主题为“MLlib 代码”。

编　者

2017 年 1 月

# 目　录

试读结束：需要全本请在线购买： www.ertongbook.com

# 第 1 章

# ◂ 星星之火 ▸

星星之火，可以燎原吗？

当我们每天面对扑面而来的海量数据，是战斗还是退却，是去挖掘其中蕴含的无限资源，还是就让它们自生自灭？我的答案是：“一切都取决于你自己”。对于海量而庞大的数据来说，在不同人眼里，既可以是一座亟待销毁的垃圾场，也可以是一个埋藏有无限珍宝的金银岛，这一切都取决于操控者的眼界与能力。本书的目的就是希望所有技术人员都有这种挖掘金矿的能力！

本章主要知识点：

- 什么是大数据？
- 数据要怎么分析？
- MLlib 能帮我们做些什么？

## 1.1 大数据时代

什么是“大数据”？一篇名为“互联网上一天”的文章告诉我们：

一天之中，互联网上产生的全部内容可以刻满 1.68 亿张 DVD，发出的邮件有 2940 亿封之多（相当于美国两年的纸质信件数量），发出的社区帖子达 200 万个（相当于《时代》杂志 770 年的文字量），卖出的手机数量为 37.8 万台，比全球每天出生的婴儿数量高出 37.1 万。

正如人们常说的一句话：“冰山只露出它的一角”。大数据也是如此，“人们看到的只是其露出水面的那一部分，而更多的则是隐藏在水面下”。随着时代的飞速发展，信息传播的速度越来越快，手段也日益繁多，数据的种类和格式也趋于复杂和丰富，并且在存储上已经突破了传统的结构化存储形式，向着非结构存储飞速发展。

大数据科学家 JohnRauser 提到一个简单的定义：“大数据就是任何超过了一台计算机处理能力的庞大数据量”。亚马逊网络服务（AWS）研发小组对大数据的定义：“大数据是最大的宣传技术、是最时髦的技术，当这种现象出现时，定义就变得很混乱。”Kelly 说：“大数据是

可能不包含所有的信息，但我觉得大部分是正确的。对大数据的一部分认知在于它是如此之大，分析它需要多个工作负载，这是 AWS 的定义。当你的技术达到极限时也就是数据的极限”。

飞速产生的数据构建了大数据，海量数据的时代我们称为大数据时代。但是，简单地认为那些掌握了海量存储数据资料的人是大数据强者显然是不对的。真正的强者是那些能够挖掘出隐藏在海量数据背后获取其中所包含的巨量数据信息与内容的人，是那些掌握专门技能懂得怎样对数据进行有目的、有方向地处理的人。只有那些人，才能够挖掘出真正隐藏的宝库，拾取金山中的珍宝，从而实现数据的增值，实现大数据的为我所用。

## 1.2 大数据分析时代

随着“大数据时代”的到来，掌握一定的知识和技能，能够对大数据信息进行锤炼和提取越来越受到更多的数据分析人员所器重。可以说，大数据时代最重要的技能是掌握对大数据的分析能力。只有通过对大数据的分析，提炼出其中所包含的有价值内容才能够真正做到为我所用。换言之，如果把大数据比作一块沃土，那么只有强化对土地的“耕耘”能力，才能通过“加工”实现数据的“增值”。

一般来说，大数据分析需要涉及以下 5 个方面，如图 1-1 所示。


图 1-1　大数据分析的 5 个方面

### 1. 有效的数据质量

任何数据分析都来自于真实的数据基础，而一个真实数据是采用标准化的流程和工具对数据进行处理得到的，可以保证一个预先定义好的高质量的分析结果。

### 2. 优秀的分析引擎

对于大数据来说，数据的来源多种多样，特别是非结构化数据来源的多样性给大数据分析带来了新的挑战。因此，我们需要一系列的工具去解析、提取、分析数据。大数据分析引擎就是用于从数据中提取我们所需要的信息。

### 3. 合适的分析算法

采用合适的大数据分析算法能让我们深入数据内部挖掘价值。在算法的具体选择上，不仅仅要考虑能够处理的大数据的数量，还要考虑到对大数据处理的速度。

### 4. 对未来的合理预测

数据分析的目的是对已有数据体现出来的规律进行总结，并且将现象与其他情况紧密连接在一起，从而获得对未来发展趋势的预测。大数据分析也是如此。不同的是，在大数据分析中，数据来源的基础更为广泛，需要处理的方面更多。

### 5. 数据结果的可视化

大数据的分析结果更多是为决策者和普通用户提供决策支持和意见提示，其对较为深奥的数学含义不会太了解。因此必然要求数据的可视化能够直观地反映出经过分析后得到的信息与内容，能够较为容易地被使用者所理解和接受。

因此可以说，大数据分析是数据分析最前沿的技术。这种新的数据分析是目标导向的，不用关心数据的来源和具体格式，能够根据我们的需求去处理各种结构化、半结构化和非结构化的数据，配合使用合适的分析引擎，能够输出有效结果，提供一定的对未来趋势的预测分析服务，能够面向更广泛的用户快速部署数据分析应用。

## 1.3 简单、优雅、有效——这就是 Spark

Apache Spark 是加州大学伯克利分校的 AMPLabs 开发的开源分布式轻量级通用计算框架。与传统的数据分析框架相比，Spark 在设计之初就是基于内存而设计，因此其比一般的数据分析框架有着更高的处理性能，并且对多种编程语言，例如 Java、Scala 及 Python 等提供编译支持，使得用户在使用传统的编程语言即可对其进行程序设计，从而使得用户的学习和维护能力大大提高。

简单、优雅、有效——这就是 Spark！

Spark 是一个简单的大数据处理框架，可以使程序设计人员和数据分析人员在不了解分布式底层细节的情况下，就像编写一个简单的数据处理程序一样对大数据进行分析计算。

Spark 是一个优雅的数据处理程序，借助于 Scala 函数式编程语言，以前往往几百上千行

的程序，这里只需短短几十行即可完成。Spark 创新了数据获取和处理的理念，简化了编程过程，不再需要使用以往的建立索引来对数据分类，通过相应的表链接将需要的数据匹配成我们需要的格式。Spark 没有臃肿，只有优雅。

Spark 是一款有效的数据处理工具程序，充分利用集群的能力对数据进行处理，其核心就是 MapReduce 数据处理。通过对数据的输入、分拆与组合，可以有效地提高数据管理的安全性，同时能够很好地访问管理的数据。

Spark 是建立在 JVM 上的开源数据处理框架，开创性地使用了一种从最底层结构上就与现有技术完全不同，但是更加具有先进性的数据存储和处理技术，这样使用 Spark 时无须掌握系统的底层细节，更不需要购买价格不菲的软硬件平台，借助于架设在普通商用机上的 HDFS 存储系统，可以无限制地在价格低廉的商用 PC 上搭建所需要规模的评选数据分析平台。即使从只有一台商用 PC 的集群平台开始，也可以在后期任意扩充其规模。

Spark 是基于 MapReduce 并行算法实现的分布式计算，其拥有 MapReduce 的优点，对数据分析细致而准确。更进一步，Spark 数据分析的结果可以保持在分布式框架的内存中，从而使得下一步的计算不再频繁地读写 HDFS，使得数据分析更加快速和方便。


需要注意的是，Spark 并不是“仅”使用内存作为分析和处理的存储空间，而是和 HDFS 交互使用，首先尽可能地采用内存空间，当内存使用达到一定阈值时，仍会将数据存储在 HDFS 上。

除此之外，Spark 通过 HDFS 使用自带的和自定义的特定数据格式（RDD），Spark 基本上可以按照程序设计人员的要求处理任何数据，不论这个数据类型是什么样的，数据可以是音乐、电影、文本文件、Log 记录等。通过编写相应的 Spark 处理程序，帮助用户获得任何想要的答案。

有了 Spark 后，再没有数据被认为是过于庞大而不好处理或存储的了，从而解决了之前无法解决的、对海量数据进行分析的问题，便于发现海量数据中潜在的价值。

## 1.4 核心——MLlib

如果将 Spark 比作一个闪亮的星星的话，那么其中最明亮最核心的部分就是 MLlib。MLlib 是一个构建在 Spark 上的、专门针对大数据处理的并发式高速机器学习库，其特点是采用较为先进的迭代式、内存存储的分析计算，使得数据的计算处理速度大大高于普通的数据处理引擎。

MLlib 机器学习库还在不停地更新中，Apache 的相关研究人员仍在不停地为其中添加更多的机器学习算法。目前 MLlib 中已经有通用的学习算法和工具类，包括统计、分类、回归、聚类、降维等，如图 1-2 所示。



试读结束：需要全本请在线购买： www.ertongbook.com