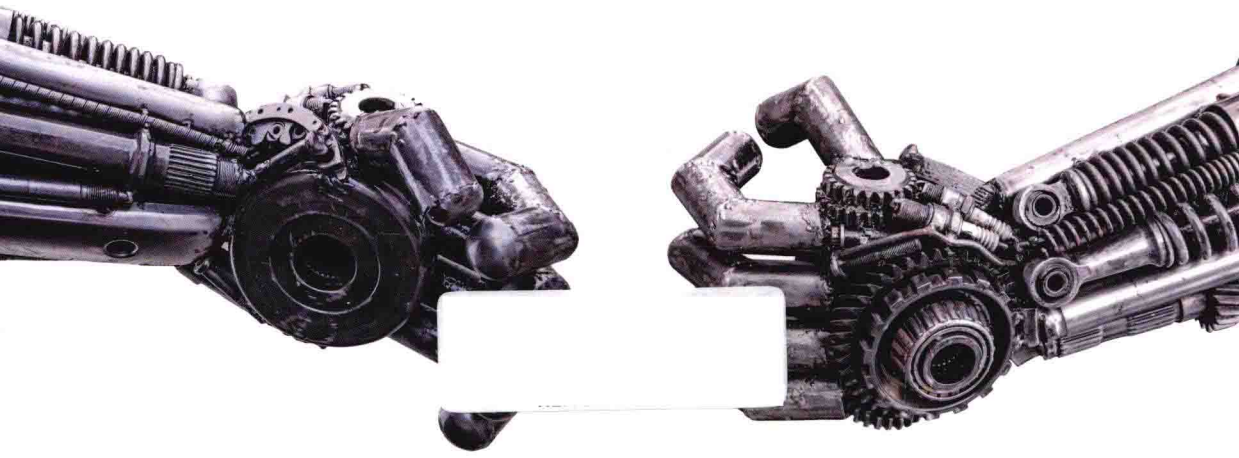
Broadview®
www.broadview.com.cn

Spark MLlib 机器学习

算法、源码及实战详解

黄美灵 著



本书系统、全面、深入地解析Spark MLlib机器学习的相关知识
以源码为基础，兼顾算法、理论与实战，帮助读者在实际工作中进行MLlib的应用开发和定制开发

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Spark MLlib 机器学习

算法、源码及实战详解

黄美灵 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书以 Spark 1.4.1 版本源码为切入点，全面并且深入地解析 Spark MLlib 模块，着力于探索分布式机器学习的底层实现。

书中本着循序渐进的原则，首先解析 MLlib 的底层实现基础：数据操作及矩阵向量计算操作，该部分是 MLlib 实现的基础；接着对各个机器学习算法的理论知识进行讲解，并且解析机器学习算法如何在 MLlib 中实现分布式计算；然后对 MLlib 源码进行详细的讲解；最后进行 MLlib 实例的讲解。相信通过本书的学习，读者可全面掌握 Spark MLlib 机器学习，能够进行 MLlib 实战、MLlib 定制开发等。

本书适合大数据、Spark、数据挖掘领域的从业人员阅读，同时也为 Spark 开发者和大数据爱好者展现了分布式机器学习的原理和实现细节。

图书在版编目（CIP）数据

Spark MLlib 机器学习：算法、源码及实战详解 / 黄美灵著.—北京：电子工业出版社，2016.4
ISBN 978-7-121-28214-0

Ⅰ. ①S… Ⅱ. ①黄… Ⅲ. ①数据处理软件－机器学习 Ⅳ. ①TP274②TP181

中国版本图书馆 CIP 数据核字（2016）第 039755 号

策划编辑：付　睿
责任编辑：李云静
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　　印张：25.25　　字数：563 千字
版　　次：2016 年 4 月第 1 版
印　　次：2016 年 4 月第 1 次印刷
印　　数：3000 册　　定价：79.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前言

机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、计算复杂性理论等多门学科，其中大部分理论来源于 18、19 世纪，例如贝叶斯定理，是 18 世纪英国数学家托马斯·贝叶斯（Thomas Bayes）提出的重要概率论理论；而 21 世纪则侧重于如何将机器学习理论运用在工业化中，帮助改进性能及提升其效率。

机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。在算法设计方面，机器学习理论关注可以实现的、行之有效的学习算法；机器学习研究的不是求解精确的结果，而是研究开发容易处理的近似求解算法。尤其是在 21 世纪，知识和数据量爆发的时代，机器学习面临大数据的求解难题。

随着数据量的增加，从传统的单机计算发展到大规模的集群计算，以至发展到今天的一种大规模、快速计算的集群平台——Apache Spark。Spark 是一个开源集群运算框架，最初由加州大学伯克利分校 AMP 实验室开发。相对于 Hadoop 的 MapReduce 会在执行完工作后将中介资料存放到磁盘中，Spark 使用了内存内运算技术，能在资料尚未写入硬盘时即在内存内分析运算。Spark 在内存上的运算速度比 Hadoop MapReduce 的运算速度快 100 倍，即便是在磁盘上运行也能快 10 倍。Spark 允许将数据加载至集群内存，并多次对其进行查询，非常适合用于机器学习算法。

本书侧重讲解 Spark MLlib 模块。Spark MLlib 是一种高效、快速、可扩展的分布式计算框架，实现了常用的机器学习，如聚类、分类、回归等算法。本文循序渐进，从 Spark 的基础知识、矩阵向量的基础知识开始，然后再讲解各种算法的理论知识，以及 Spark 源码实现和实例

实战，帮助读者从基础到实践全面掌握 Spark MLlib 分布式机器学习。

学习本书需要的基础知识包括：Spark 基础入门、Scala 入门、线性代数基础知识。

本书面向的读者：Spark 开发者、大数据工程师、数据挖掘工程师、机器学习工程师、研究生和高年级本科生等。

本书学习指南：

<table>
<tr><td>第一部分 Spark MLlib基础</td><td rowspan="4">Spark MLlib机器学习的基础包括：Spark数据操作、矩阵向量，它们都是各个机器学习算法的底层实现基础</td><td rowspan="4">通过这部分掌握：RDD的基础操作、矩阵和向量的运算、数据格式等</td></tr>
<tr><td>第1章 Spark机器学习简介</td></tr>
<tr><td>第2章 Spark数据操作</td></tr>
<tr><td>第3章 Spark MLlib矩阵向量</td></tr>
<tr><td>第二部分 Spark MLlib回归算法</td><td rowspan="16">Spark MLlib机器学习算法的全面解析。包含常见机器学习：回归、分类、聚类、关联、推荐和神经网络</td><td rowspan="16">通过这部分掌握：机器学习算法理论知识、机器学习算法的分布实现方法、MLlib源码解析、实例解析

其中第14、15章是基于Spark MLlib上实现或者定制开发机器学习算法，读者可掌握分布式机器学习的开发

分布式机器学习的学习路径：
理论→分布式实现逻辑→开发→实例</td></tr>
<tr><td>第4章 Spark MLlib线性回归算法</td></tr>
<tr><td>第5章 Spark MLlib逻辑回归算法</td></tr>
<tr><td>第6章 Spark MLlib保序回归算法</td></tr>
<tr><td>第三部分 Spark MLlib分类算法</td></tr>
<tr><td>第7章 Spark MLlib贝叶斯分类算法</td></tr>
<tr><td>第8章 Spark MLlib SVM支持向量机算法</td></tr>
<tr><td>第9章 Spark MLlib决策树算法</td></tr>
<tr><td>第四部分 Spark MLlib聚类算法</td></tr>
<tr><td>第10章 Spark MLlib KMeans聚类算法</td></tr>
<tr><td>第11章 Spark MLlib LDA主题模型算法</td></tr>
<tr><td>第五部分 Spark MLlib关联规则挖掘算法</td></tr>
<tr><td>第12章 Spark MLlib FPGrowth关联规则算法</td></tr>
<tr><td>第六部分 Spark MLlib推荐算法</td></tr>
<tr><td>第13章 Spark MLlib ALS交替最小二乘算法</td></tr>
<tr><td>第14章 Spark MLlib 协同过滤推荐算法</td></tr>
<tr><td>第七部分 Spark MLlib神经网络算法</td></tr>
<tr><td>第15章 Spark MLlib神经网络算法综述</td></tr>
</table>

在本书的编写过程中，何娟、何丹、王蒙、叶月媚参与了全书的编写、整理及校对工作，刘程辉、李俊、廖宏参与了 Spark 集群运维和第 2 章数据操作的实例部分工作，刘晓宏、方佳武、于善龙参与了全书的实例部分工作。

由于笔者水平有限，编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读

者批评指正。您也可以通过博客 http://blog.csdn.net/sunbow0、邮箱 humeli317@163.com 和 QQ 群 487540403 联系到我，期待能够得到读者朋友们的真挚反馈，在技术之路上互勉共进。

本书在写作的过程中，得到了很多朋友及同事的帮助和支持，在此表示衷心感谢！

感谢久邦数码大数据团队的同事们。在两年的工作中，笔者得到了很多同事的指导、支持和帮助，尤其感谢杨树清、周小平、梁宁、刘程辉、刘晓宏、方佳武、于善龙、王蒙、叶月媚、廖宏、谭钊承、吴梦玲、邹桂芳、曹越等。

感谢电子工业出版社的付睿编辑，她不仅积极策划和推动本书的出版，而且在写作过程中还给出了极为详细的改进意见。感谢电子工业出版社的李云静编辑为本书做了非常辛苦和专业的编辑工作。

感谢我的父母和妻子，有了你们的帮助和支持，我才有时间和精力去完成写作。

谨以此书献给热爱大数据技术的朋友们！

目录

第一部分　Spark MLlib 基础

第三部分 Spark MLlib 分类算法

第四部分 Spark MLlib 聚类算法

第七部分　Spark MLlib 神经网络算法

第一部分

Spark MLlib 基础

第 1 章

Spark 机器学习简介

1.1 机器学习介绍

机器学习（Machine Learning，ML）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。在算法设计方面，机器学习理论关注可以实现的、行之有效的学习算法。

机器学习可以分成下面几种类别。

- 监督学习：输入数据被称为训练数据，它们有已知的标签或者结果，比如垃圾邮件/非垃圾邮件或者某段时间的股票价格。模型的参数确定需要通过一个训练的过程，在这个过程中模型将会要求做出预测，当预测不符时，则需要做出修改。常见的监督学习算法包括回归分析和统计分类。
- 无监督学习：输入数据不带标签或者没有一个已知的结果。通过推测输入数据中存在的结构来建立模型。常见的无监督学习算法有聚类。
- 半监督学习：输入数据由带标签的和不带标签的组成。合适的预测模型虽然已经存在，

但是模型在预测的同时还必须能通过发现潜在的结构来组织数据。这类问题包括分类和回归。

- 强化学习：输入数据作为来自环境的激励提供给模型，且模型必须做出反应。反馈并不像监督学习那样来自训练的过程，而是作为环境的惩罚或者奖赏。例如，系统和机器人控制。算法的例子包括 Q 学习和时序差分学习。

当你处理大量数据来对商业决策建模时，通常会使用监督和无监督学习。目前的一个热门话题是半监督学习，比如会应用在图像分类中，涉及的数据集很大但是只包含极少数标签的数据。

常见的机器学习算法如下：

- 分类与回归——线性回归、逻辑回归、贝叶斯分类、决策树分类等；
- 聚类——KMeans 聚类、LDA 主题、KNN 等；
- 关联规则——Apriori、FPGrowth 等；
- 推荐——协同过滤、ALS 等；
- 神经网络——BP、RBF、SVM 等；
- 深度神经网络等算法。

1.2　Spark 介绍

Spark 是一个基于内存计算的开源集群计算系统，是由加州大学伯克利分校 AMP 实验室使用 Scala 语言开发的，目前已是 Apache 的顶级开源项目，是 Apache 社区最火热的项目之一。Spark 提供了一个更快、更通用的数据处理平台，和 Hadoop 相比，Spark 可以让你的程序在内存中运行时速度提升 100 倍，或者在磁盘上运行时速度提升 10 倍。

Spark 是基于 MapReduce 算法实现的分布式计算，拥有 Hadoop MapReduce 所具有的优点；但不同于 MapReduce 的是中间过程的输出和结果可以保存在内存中，从而不再需要读写 HDFS，因此 Spark 能更好地适用于数据挖掘与机器学习等需要迭代的 MapReduce 的算法。

Spark 核心由一组功能强大的、高级别的库组成，这些库可以无缝地应用到同一个应用程序中。目前这些库包括 SparkSQL、Spark Streaming、MLlib 及 GraphX，如图 1-1 所示。

Spark Core 是一个基本引擎，用于大规模并行和分布式数据处理。Spark 引入了弹性分布式数据集（RDD，Resilient Distributed Dataset）。RDD 是一个不可变的、容错的、分布式对象集合，我们可以并行地操作这个集合，并且 RDD 提供了丰富的数据操作接口。

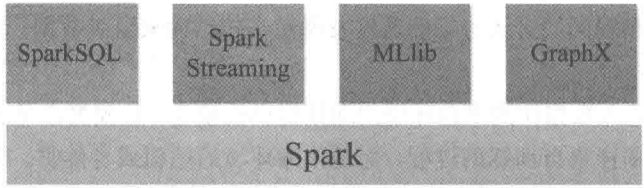


图 1-1　Spark 组件

SparkSQL 是 Spark 的一个组件，它支持我们通过 SQL 或者 Hive 查询语言来查询数据。

Spark Streaming 支持对流数据的实时处理，例如产品环境 Web 服务器的日志文件（例如 Apache Flume 和 HDFS/S3），Spark Streaming 会接收日志数据，然后将其分为不同的批次，接下来 Spark 引擎来处理这些批次，并根据批次中的结果，生成最终的流。

MLlib 是一个机器学习库，它提供了各种各样的算法，这些算法用来在集群上针对分类、回归、聚类、协同过滤等。

GraphX 是一个图计算库，用来处理图，执行基于图的并行操作。

1.3　Spark MLlib 介绍

MLlib 是 Spark 中可扩展的机器学习库，它由一系列机器学习算法和实用程序组成，包括分类、回归、聚类、协同过滤、降维，还包括一些底层的优化方法，如图 1-2 所示。

1）依赖

MLlib 的底层实现采用数值计算库 Breeze 和基础线性代数库 BLAS。

2）优化计算

MLlib 目前支持随机梯度下降法、少内存拟牛顿法、最小二乘法等。

3）回归

MLlib 目前支持线性回归、逻辑回归、岭回归、保序回归和与之相关的 L1 和 L2 正则化的变体。MLlib 中回归算法的优化计算方法采用随机梯度下降法。

4）分类

MLlib 目前支持贝叶斯分类、决策树分类、线性 SVM 和逻辑回归，同时也包括 L1 和 L2 正则化的变体。优化计算方法也采用随机梯度下降法。