HZ BOOKS
华章IT

[PACKT] PUBLISHING

以分布式和可扩展的方式处理和分析大数据，包含精心设计的抽取过程；
构建和使用回归模型预测航班延误，并编写复杂的Spark流水线。

大数据技术丛书

Mastering Machine Learning with Spark 2.x

Spark机器学习

核心技术与实践

[美] 亚历克斯·特列斯（Alex Tellez）
马克斯·帕普拉（Max Pumperla） 著
迈克尔·马洛赫拉瓦（Michal Malohlava）

邵赛赛 阳卫清 唐明洁 译

机械工业出版社
China Machine Press

大数据技术丛书

Mastering Machine Learning with Spark 2.x

Spark机器学习

核心技术与实践

[美] 亚历克斯·特列斯（Alex Tellez）
马克斯·帕普拉（Max Pumperla） 著
迈克尔·马洛赫拉瓦（Michal Malohlava）

邵赛赛 阳卫清 唐明洁 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Spark 机器学习：核心技术与实践 /（美）亚历克斯·特列斯（Alex Tellez），（美）马克斯·帕普拉（Max Pumperla），（美）迈克尔·马洛赫拉瓦（Michal Malohlava）著；邵赛赛，阳卫清，唐明洁译 . —北京：机械工业出版社，2018.5
（大数据技术丛书）
书名原文：Mastering Machine Learning with Spark 2.x

ISBN 978-7-111-59846-6

I. S… II. ① 亚… ② 马… ③ 迈… ④ 邵… ⑤ 阳… ⑥ 唐… III. 机器学习 IV. TP181

中国版本图书馆 CIP 数据核字（2018）第 090318 号

本书版权登记号：图字 01-2017-7522

Spark 机器学习：核心技术与实践

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：缪 杰 责任校对：殷 虹
印 刷：三河市宏图印务有限公司 版 次：2018 年 6 月第 1 版第 1 次印刷
开 本：186mm×240mm 1/16 印 张：15.25
书 号：ISBN 978-7-111-59846-6 定 价：69.00 元

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 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

The Translator's Words 译 者 序

近些年来，随着计算力的提升、集群规模的扩大以及大数据基础软件的普及，人们对于海量数据的存储和处理能力得到了极大的提升，大数据的处理要求不再仅限于简单的数据清理、聚合和转换，更希望能从海量数据中发现模式，预测规律。而人工智能、深度学习热潮的兴起，更是把大数据和机器学习推向了更高的高度。

在这样的背景下，传统的基于单机、小数据的机器学习算法库变得不再合适，人们迫切希望有一个分布式的、可伸缩的、高性能的机器学习库来处理海量数据。为此大量基于分布式框架的机器学习库应运而生，而 Spark MLlib 作为其中最为耀眼的明星受到了极大的关注。为什么 Spark MLlib 会受到如此青睐呢?

首先，Spark MLlib 是基于 Spark 通用计算引擎所构建的，Spark 计算引擎因其更为优异的架构和实现而拥有很好的性能。因此构建在其上的 MLlib 也具备很好的性能表现。

其次，Spark MLlib 充分吸收和借鉴了当今主流的机器学习 API 的设计，它的 API 与 Python scikit-learn 极为相似，因此传统的机器学习工程师可以极低的学习成本将已有经验迁移到 Spark 上来。

再者，Spark 提供了一个完整的生态圈，囊括了各种类型的计算模式。用户可以轻松地使用 Spark 来实现完整的数据处理和机器学习流水线，而无须切换不同的系统。

当然，Spark 和 Spark MLlib 还有很多其他的优点，这就是它为什么能够成为现今主流大数据机器学习库的原因所在。

本书作为 Spark MLlib 的书籍从实践的角度介绍了各种机器学习算法，其中第 2 章和第 3 章介绍了分类算法，第 4 ～ 6 章介绍了自然语言处理的相关技术，第 7 章介绍了图算法相关的知识，第 8 章则集合了前面所有的算法介绍了一个端到端的机器学习处理实例。

本书与其他常见机器学习书籍的最大不同点在于，本书并没有对算法原理进行详细的介绍，同时也没有过多涉及 Spark 和 Spark MLlib 实现 /API。本书希望读者对算法基础和 Spark

使用有一个基本的了解。相反，本书每一章对于每一种算法都辅以一个具体而实际的例子，从利用二分类算法来分类希格斯玻色子到使用自然语言处理技术来区分电影评论中的正负面情绪。更为难能可贵的是 1 本书同时介绍了模型部署和模型评分的实际场景和问题，而不仅仅局限于模型训练。最后本书以 Lending Club 借贷预测的实际例子来展示从数据整理到建模再到部署的整个过程。

算法和技术的更新日新月异，译者并不期待书中所介绍的东西不会过时，但是译者更希望读者能从本书中了解到数据科学和机器学习的方法论，这将会更有帮助。

本书由本人和两位同事阳卫清、唐明洁共同翻译而成。在翻译期间译者付出了大量的时间和精力。同时，感谢家人的支持和帮助，没有她们在背后的支持，本书的翻译是不可能实现的。

译者水平有限，翻译之中难免疏漏，有任何的意见和建议请及时反馈，以便译者及时修正和改进。为此译者在 GitHub 上创建了一个网页 https://github.com/jerryshao/Mastering-Machine-Learning-with-Spark，读者可以通过创建 issues 来对任何问题进行反馈。

邵赛赛

2018 年 1 月 12 日于上海

About the Authors 关于作者

Alex Tellez 是一名终身的数据黑客 / 爱好者，对数据科学及其在商业问题上的应用充满了激情。他在多个行业拥有丰富的经验，包括银行业、医疗保健、在线约会、人力资源和在线游戏。Alex 还在各种人工智能 / 机器学习会议上进行过多次演讲，同时也在大学讲授关于神经网络的课程。闲暇时间，Alex 喜欢和家人在一起，骑自行车，并利用机器学习来满足他对法国葡萄酒的好奇心！

首先，我要感谢 Michal 与我一起编写本书。同样作为资深的机器学习（Machine Learning，以下简称 ML）爱好者、自行车爱好者、跑者和父亲，在一年来共同努力的过程中，我们对彼此有了更深的了解。换句话说，没有 Michal 的支持和鼓励，本书是不可能完成的。

接下来，我要感谢我的妈妈、爸爸和哥哥 Andres，从我出生第一天直到现在的每一步，你们都陪伴在我的周围。毋庸置疑，我的哥哥仍会是我的英雄，是我永远仰望的人，是我的指路灯。当然，还要感谢我美丽的妻子 Denise 和女儿 Miya，在每个夜晚和周末给予我写作上的关心和支持。我无法描述你们对我而言意味着多少，你们是我保持持续创作的灵感和动力。对我的女儿 Miya，我的希望是，有一天当你拿起这本书时，会意识到你的老爸并不像看起来那么傻。

最后，我也要感谢你——读者，感谢你对这个令人兴奋的领域以及难以置信的技术感兴趣。无论你是一名经验丰富的 ML 专家，还是希望立足的新人，你都会找到适合自己的内容，我希望你能像 Michal 和我一样，从本书中获得很多。

Max Pumperla 是一名数据科学家和工程师，专注于深度学习及其应用。他目前在 Skymind 担任深度学习工程师，并且是 aetros.com 的联合创始人。Max 是几个 Python 软件包的作者和维护者，包括 elephas，一个使用 Spark 的分布式深度学习库。他的开源足迹包括对许多流行的机器学习库的贡献，如 keras、deeplearning4j 和 hyperopt。他拥有汉堡大学的代数几何博士学位。

Michal Malohlava 是 Sparkling Water 的创建者、极客和开发者，Java、Linux、编程语言爱好者，拥有 10 年以上的软件开发经验。他于 2012 年在布拉格的查尔斯大学获得博士学位，并在普渡大学攻读博士后。

在学习期间，他关注利用模型驱动方法和领域特定语言构建分布式、嵌入式、实时和模块化系统，参与了各种系统的设计和开发，包括 SOFA 和分形组件系统以及 jPapabench 控制系统。

现在，他的主要兴趣是大数据计算。他参与了高级大数据计算平台 H2O 的开发，并将其嵌入到 Spark 引擎中作为 Sparkling Water 项目发布。

我要感谢我的妻子 Claire，感谢她对于我的爱和鼓励。

Preface 前　言

大数据是几年前我们开始探索用 Spark 进行机器学习时的初衷。我们希望建立的机器学习程序能够充分利用大量数据训练模型，但一开始这并不容易。Spark 仍在演进阶段，还没有包含强大的机器学习库，而且我们也在试图弄清楚建立一个机器学习程序到底意味着什么。

慢慢地，我们开始探索 Spark 生态系统的各个角落，追随它的演进。对我们来说，最关键的是需要一个强大的机器学习库，能够提供像 R 和 Python 库那样的功能。这对我们来说比较容易，因为当时我们正积极参与 H2O 机器学习库和它的一个叫作 Sparkling Water 的分支的开发，这个分支能够让 Spark 应用程序使用 H2O 库。然而，模型训练只是机器学习的冰山一角，我们还不得不弄清楚如何把 Sparkling Water 连接到 Spark RDD、DataFrame 以及 DataSet，怎样用 Spark 连接和读取不同的数据源，以及怎样把模型导出到其他的应用程序加以使用。

在这个过程中，Spark 自身也在演进。Spark 最初是一个纯粹的 Scala 项目，后来开始提供 Python 接口，之后提供 R 接口。Spark API 也在这个漫长的过程中从提供底层的 RDD 接口发展到高阶的 DataSet 接口（一组类 SQL 的接口）。而且，Spark 也采纳了源自 Python scikit-learn 库的机器学习流水线的概念。所有这些改进使得 Spark 成为一个非常好的数据转换和处理工具。

基于这些经验，我们决定撰写本书，同世界分享我们得到的知识，意图很简单：用示例来展示建立 Spark 机器学习应用的方方面面，不仅展示如何使用最新的 Spark 功能，而且也展示 Spark 底层接口。我们所发现的关于 Spark、机器学习应用开发流程和源代码组织方面很多小的技巧和捷径也会在本书中分享给读者，让大家免于犯同样的错误。

本书的示例使用 Scala 作为主要的实现语言。使用 Python 还是 Scala 是一个艰难的抉择，但是最终 Scala 胜出。使用 Scala 有两个主要的原因：它提供了最为完整的 Spark 接口，而且得益于 JVM 带来的性能优势，在生产环境中部署的大部分应用都使用 Scala。最后，本书的示例源代码都可以在网上下载。

希望你能够享受本书带来的阅读乐趣，并且希望它能够帮助你遨游 Spark 的世界，帮助你开发机器学习应用。

本书主要内容

第 1 章带领读者进入机器学习和大数据的世界，介绍它们的历史，以及包括 Apache Spark 和 H2O 在内的当代工具。

第 2 章专注于二项模型的训练和评估。

第 3 章尝试根据健身房中人体传感器所收集的数据推测人的活动。

第 4 章介绍使用 Spark 处理自然语言问题，展示其对电影评论进行情感分析的能力。

第 5 章详细讨论了当代自然语言处理技术。

第 6 章介绍频繁模式挖掘的基础知识，Spark MLlib 中相关的三个算法，以及把算法部署为 Spark Streaming 应用。

第 7 章介绍图和图分析的基本概念，解释 Spark GraphX 的核心功能，以及一些图算法，如 PageRank。

第 8 章把之前章节介绍的技巧组合为一个完整的示例，包括数据处理、模型搜索和训练，以及把模型部署为一个 Spark Streaming 应用。

所需的环境

本书提供的代码示例基于 Apache Spark 2.1 及 Scala API，使用 Sparkling Water 库来访问 H2O 机器学习库。在每一章中，我们会展示如何使用 spark-shell 启动 Spark，以及如何下载运行代码所需要的数据。

简而言之，运行本书提供的代码所需的基础环境包括：

- Java 8
- Spark 2.1

面向的读者

如果你是一名开发者，有着机器学习和统计背景，但是受限于现有的、慢速的、基于小规模数据的机器学习工具，这本书适合你！在本书中，你会使用 Spark 创建可扩展的机器学习应用，用来支撑现代数据驱动商业。我们假定你已经了解机器学习的基本概念和算法，能

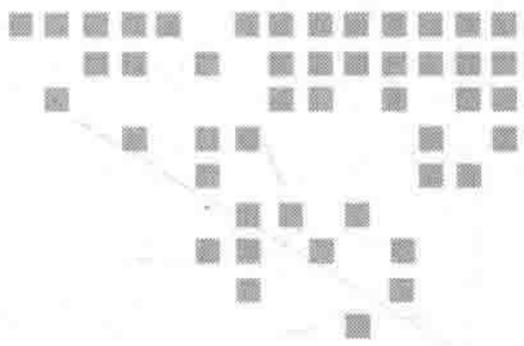
够运行 Spark（在集群或者本地运行），而且对 Spark 的各种基本库有一些基础的了解。

下载示例代码和彩图

本书的代码包在 GitHub（https://GitHub.com/PacktPublishing/Mastering-Machine-Learning-with-Spark-2.x）上也可以找到。来自其他书籍的代码和视频也可以在 https://github.com/PacktPublishing/ 上找到。去看看吧！

本书提供一个 PDF 文件，它以彩图的格式包含了本书中所使用到的屏幕截图和图表。这些彩图可以帮助读者更好地理解输出中发生的变化。该 PDF 文件可以从 https://www.packtpub.com/sites/default/files/downloads/MasteringMachineLearningwithSpark2.x_ColorImages.pdf 下载到。

目　录 *Contents*



第1章 Chapter 1

大规模机器学习和 Spark 入门

"信息是 21 世纪的石油，而数据分析则是内燃机。"

——Peter Sondergaard，Gartner Research

据估计，到 2018 年全世界的公司在大数据有关的项目上花费将达到 1140 亿美元，相比 2013 年大约增长 300%（https://www.capgemini-consulting.com/resource-file-access/resource/pdf/big_data_pov_03-02-15.pdf）。支出增长很大的一部分归因于大量数据的创建和通过使用 Hadoop 这样的分布式文件系统带来的存储数据的能力。

然而，收集数据只是战役的一半；另一半涉及数据的提取、转换，以及利用现代计算机的计算能力使用各种数学方法处理数据，以更深入地理解数据和数据的模式，从中提取有用的信息以做出相关决策。过去数年里，增长的计算能力，容易得到的可伸缩云服务（如亚马逊 AWS、微软 Azure 以及 Heroku），以及许多有助于管理、控制和扩展的基础架构，帮助应用构建的工具和开发库的出现，大大加速了整个数据处理流的发展。计算能力的增长也使得处理更大规模的数据和使用之前无法应用的算法成为可能。各种计算密集的统计和机器学习算法开始被用来从数据中提取有用的信息。

第一个被广泛采用的技术之一是 Hadoop。Hadoop 通过把中间结果存储在硬盘上来支持 MapReduce 计算，可是仍然缺乏合适的大数据工具来进行信息提取。不过，Hadoop 只是一个开始，随着计算机内存越来越大，新的内存计算框架开始出现，这些框架也开始对数据分析和建模提供基本的支持，比如 SystemML、Spark 的 SparkML，以及 Flink 的 FlinkML。这些框架也仅仅是大数据这个持续演进的生态系统的冰山一角，因为数据量总是在增长，所以要求着新的大数据算法和处理方法。比方说，**物联网**（IoT）代表了一个新的领域：来自各种数据源（如家庭安全系统、Alexa Echo、生命监视器）的海量流数据不仅带来了数据

挖掘的无限潜能，也对新的数据处理和建模方法提出了要求。

在本章中，我们将从头开始探讨如下主题：

- 数据科学家的基本工作任务
- 分布式系统中的大数据计算
- 大数据生态系统
- Spark 及其机器学习支持

1.1 数据科学

对数据科学这个词给出一个一致认可的定义，就像品尝红酒然后在朋友之间分辨其口感一样，每个人都不一样。每个人都会有他自己的定义，没有谁的更准确。然而就其本质而言，数据科学是向数据提问的艺术：问聪明的问题，得到聪明的有用的回答。不幸的是，反过来也是成立的：问笨的问题将得到糟糕的回答！因此，认真构造你的问题是从数据中提取出有价值洞见的关键。因为这个原因，公司现在都雇佣数据科学家来帮助构建这些问题。图 1-1 是谷歌趋势（Google Trends）给出的大数据和数据科学的搜索次数及变化趋势。

图 1-1　大数据和数据科学持续增长的谷歌趋势

1.2 数据科学家：21 世纪最炫酷的职业

要给出一个典型数据科学家的古板肖像是很容易的：T 恤衫、宽松的运动裤、边框厚厚的眼镜，在 IntelliJ 里调试大段的代码……大致如此。不过，除了外表之外，数据科学家的特征还有哪些？图 1-2 是我们最喜欢的海报之一，它描述了这个角色。

现代数据科学家

数据科学家——21 世纪最炫酷的职业，需要各学科的复合技能，包括数学、统计学、计算机科学、沟通力以及商业技能。找到一位数据科学家是不容易的，找到一位伯乐也同样困难。为此这里提供了一个表格来描述什么是真正的数据科学家。

数学和统计学

- ☆机器学习
- ☆统计建模
- ☆实验设计
- ☆贝叶斯推断
- ☆监督学习：决策树、随机森林、逻辑回归
- ☆非监督学习：聚类、降维
- ☆优化：梯度下降及其变种

编程和数据库

- ☆计算机科学基础
- ☆脚本语言，如 Python
- ☆统计计算包，如 R
- ☆数据库：SQL 和 NoSQL
- ☆关系代数
- ☆并行数据库和并行查询处理
- ☆MapReduce 概念
- ☆Hadoop 和 Hive/Pig
- ☆自定义 reducers
- ☆xaaS 经验，如 AWS

领域知识和软技能

- ☆对于业务的激情
- ☆对于数据的好奇
- ☆无权威影响力
- ☆骇客精神
- ☆问题解决者
- ☆有策略、前瞻性和创造力，有创新性和合作精神

沟通力和视野

- ☆能与资深经理合作
- ☆讲故事的能力
- ☆将数据视野转换为决策和行动
- ☆可视化艺术设计
- ☆R 包如 ggplot 或是 lattice
- ☆任何可视化工具的知识，如 Flare、D3.js、Tableau

图 1-2　什么是数据科学家

数学、统计学和计算机科学的一般知识都列出了，但是常常在职业数据科学家中见到的一个困难是对业务问题的理解。这又回到向数据提问的问题上来，这一点怎么强调也不为过：一个数据科学家只有理解了业务问题，理解了数据的局限性，才能够向数据问出更多聪明的问题；没有这些根本的理解，即使是最聪明的算法也无法在一个脆弱的基础上得出可靠的结论。

1.2.1　数据科学家的一天

对你们中间的一些人来说，这个事实可能会让你惊讶——数据科学家不只是喝着浓缩咖啡忙于阅读学术论文、研究新工具和建模直到凌晨；事实上，这些只占数据科学家相当少的一部分时间。数据科学家将一天中的绝大部分时间都花在各种会议里，以更好地理解业务问题，分析数据的局限性（打起精神，这本书会给你无数特征工程（feature engineering）和特征提取（feature extraction）的任务），以及如何能够把结果最好地呈现给非数据科学家们。这是个繁杂的过程，最好的数据科学家也需要能够乐在其中，因为这样才能更好地理解需求和衡量成功。事实上，要从头到尾描述这个过程，我们完全能够写一本新书！

那么，向数据提问到底涉及哪些方面呢？有时，只是把数据存到关系型数据库里，运行一些 SQL 查询语句："购买这个产品的几百万消费者同时购买哪些其他产品？找出最多的前三种。"有时问题会更复杂，如："分析一个电影评论是正面的还是负面的？"这本书主要关注后一种复杂问题。对这些复杂问题的回答是大数据项目对业务真正产生影响的地方，同时我们也看到新技术的快速涌现，让这些问题的回答更加简单，功能加更丰富。

一些试图帮助回答数据问题的最流行的开源框架包括 R、Python、Julia 和 Octave，所有这些框架在处理小规模（小于 100 GB）数据集时表现良好。在这里我们对数据规模（大数据与小数据）给出一个清晰的界定。在工作中一个总的原则是这样的：

如果你能够用 Excel 来打开你的数据集，那么你处理的是小规模数据。

1.2.2 大数据处理

如果要处理的数据集太大放不进单个计算机的内存，那么必须分布到一个集群的多个节点上，该怎么处理？比方说，难道不能仅仅通过修改扩展一下原来的 R 代码，来适应多个计算节点的情况？要是事情能这么简单就好了。把单机算法扩展到多台机器的困难原因有很多。设想如下一个简单的例子，一个文件包含了一系列名字：

```
B
D
X
A
D
A
```

我们想算出文件中每个名字出现的次数。如果这个文件能够装进单机，你可以通过 Unix 命令组合 sort 和 uniq 来解决：

```
bash> sort file | uniq -c
```

输出如下：

```
2 A
1 B
1 D
1 X
```

然而，如果文件很大需要分布到多台机器上，就有必要采用一种稍微不同的策略，比如拆分文件，使每个分片都能够装入内存，再对每个分片分别进行计算，最后将结果汇总。因此，即使是简单的任务，像这个统计名字出现次数的例子，在分布式环境中也会变得复杂。

1.2.3 分布式环境下的机器学习算法

机器学习算法把简单任务组合为复杂模式，在分布式环境下变得更为复杂。以一个简