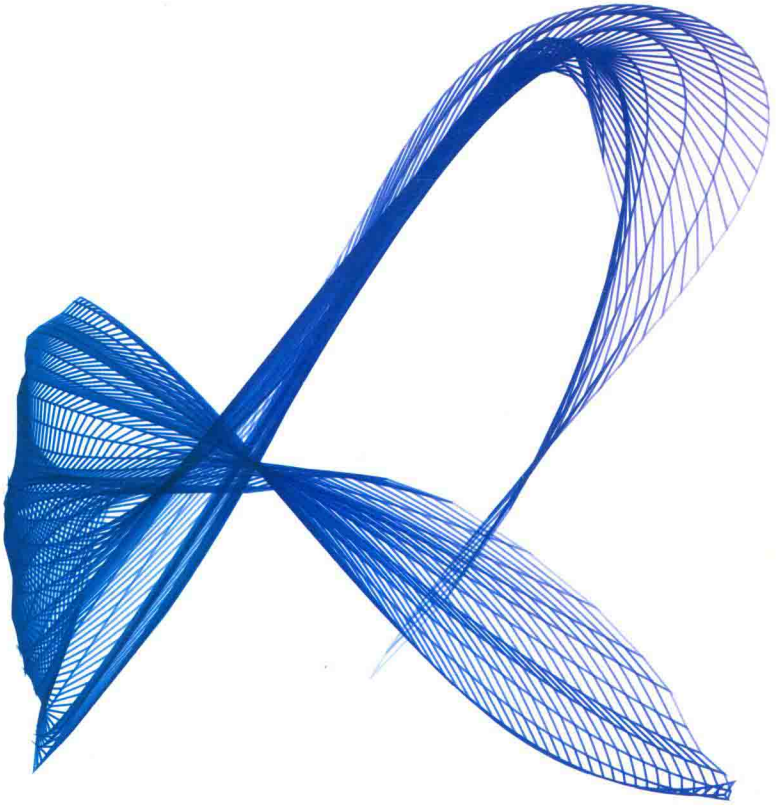
HZ BOOKS 华章IT

科大讯飞大数据专家团队撰写，不囿于Spark机器学习库，突出工程化思维与实践

6大算法模型构建，5大场景（异常检测、用户画像、点击率预估、企业征信、智慧交通）应用，从内涵认知到实践技能，全面提升

大数据技术丛书



Spark机器学习进阶实战

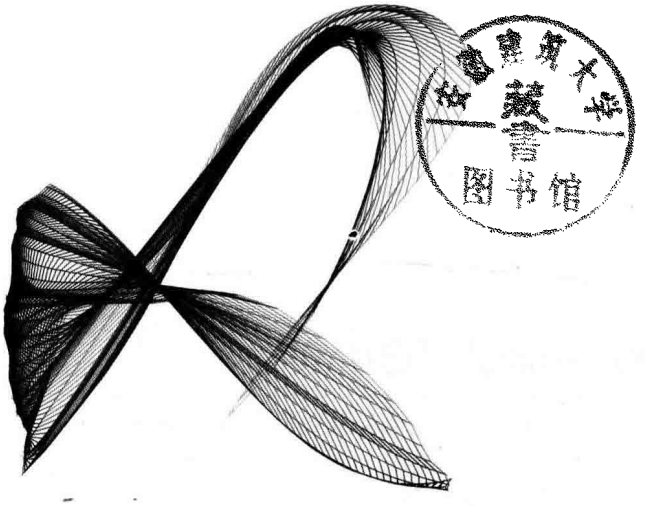
马海平 于俊 吕昕 向海◎著

机械工业出版社
China Machine Press

大数据技术丛书

Spark机器学习进阶实战

马海平 于俊 吕昕 向海◎著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Spark 机器学习进阶实战 / 马海平等著 . —北京：机械工业出版社，2018.9
（大数据技术丛书）

ISBN 978-7-111-60810-3

I. S… II. 马… III. 数据处理软件 IV. TP274

中国版本图书馆 CIP 数据核字（2018）第 201129 号

Spark 机器学习进阶实战

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：高婧雅　　责任校对：李秋荣
印　　刷：北京市兆成印刷有限责任公司　　版　　次：2018 年 9 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：14
书　　号：ISBN 978-7-111-60810-3　　定　　价：59.00 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

作者简介

马海平

科大讯飞大数据研究院研究主管，中国科学与技术大学计算机技术博士，专注数据挖掘和人工智能算法的研究，及其在计算广告和个性化教育等方向的落地应用。

于俊

科大讯飞大数据专家，专注大数据和人工智能应用方案设计、基于Spark的大数据分析和价值挖掘，在大数据算法工程化实现方面具有丰富经验。

吕昕

科大讯飞大数据专家，专注大数据和人工智能技术在消费者业务中的应用、基于Spark的大数据分析和算法建模，在用户画像、内容推荐和精准营销领域有丰富的实践。

向海

邂智科技算法负责人，前科大讯飞大数据专家。专注Spark机器学习在智能客服中的应用，在NLP与对话机器人应用方面有丰富经验。

内容简介

科大讯飞大数据专家团队撰写，不囿于Spark机器学习库，突出算法的工程化思维与实践。从基础引出算法，从算法实践到场景应用，层层推进，分享笔者的一些想法和见解，铺展开更为深入、全面的思路。

6大机器学习模型构建

分类：刻画事物特征的类标识，有效预测未知数据的归类情况。

聚类：根据相似程度生成对象集合，同集合相似，不同集合相异。

回归：找出数据规律和趋势，预测数据未来变化。

关联规则：挖掘关联关系，辅助商业决策。

协同过滤：刻画用户相似兴趣，实现偏好预测。

降维：有效地消除无关和冗余特征，提升模型精度。

5大典型应用场景

异常检测：有效解决入侵检测、欺诈检测、社交假新闻等问题。

用户画像：高度精炼用户的特征标识，为产品与决策提供数据支持和事实依据。

点击率预估：预估点击概率，计算点击收益，选出收益最高的策略。

企业征信：提供信用信息服务，洞察企业信用风险。

智慧交通：实现交通数据的价值，提供解决城市交通问题的思路。

同时，本书从《道德经》和《庄子》精选名言，并结合大数据机器学习相关内容，对名言加以讲解，引导大家以老庄的思想来认识大数据的内涵。

华章IT
HZBOOKS | Information Technology

Preface 前　言

上善若水，水善利万物而不争。

数据一如水，无色无味，非方非圆，以百态存于自然，于自然无违也。绵绵密密，微则无声，巨则汹涌；与人无争却又容纳万物。生活离不开水，同样离不开数据，我们被数据包围，在数据中生活，体会着数据量爆炸式增长带来的幸福和挑战。

本书从《道德经》和《庄子》精选名言，并结合大数据机器学习相关内容，对名言加以讲解，引导大家以老庄的思想认识大数据的内涵，使用机器学习进行大数据价值挖掘，探求老子道之路和庄子智慧之路。

为什么要写这本书

2014 年春天，曾经和公司大数据团队小伙伴一起聚焦研究大数据，为了解决国内资料匮乏、学习门槛较高的问题，着手编写《Spark 核心技术与高级应用》[⊖]一书，并于 2016 年 1 月出版，取得了较好的反响，得到很多朋友的支持。

近年来，随着收集、存储和分析的数据量呈爆炸式增长，大规模的数据分析和数据价值挖掘能力已经成为影响企业生死存亡的关键，越来越多的企业必须面对这残酷而美好的挑战。基于大数据的机器学习有效解决了大数据带来的数据分析和数据挖掘瓶颈。

如何让更多的大数据从业人员更轻松地使用机器学习算法进行大数据价值挖掘，通过简单的学习建立大数据环境下的机器学习工程化思维，在不必深究算法细节的前提下，实现大数据分类、聚类、回归、协同过滤、关联规则、降维等算法，并使用这些算法解决实际业务场景的问题。2016 年秋天，在机械工业出版社高婧雅编辑的指导下，怀着一颗附庸风雅之心，

⊖ 该书已由机械工业出版社出版，书号为 ISBN 978-7-111-52354-3。——编辑注

我决定和小伙伴们一起朝着新的目标努力。

本书的写作过程中，Spark 版本也在不断变化，秉承大道至简的原则，我们一方面尽量按照新的版本进行统筹，另外一方面尽量做到和版本解耦，希望能抛砖引玉，以个人的一些想法和见解，为读者拓展出更深入、更全面的思路。

本书只是一个开始，如何使用机器学习算法从海量数据中挖掘出更多的价值，还需要无数的大数据从业人员前赴后继，突破漫漫雄关，共同创造美好的大数据机器学习时代。

本书特色

本书介绍大数据机器学习的算法和实践，同时对传统文化进行了一次缅怀，吸收传统文化的精华，精选了《道德经》和《庄子》部分名言，实现大数据和哲学思想的有效统一。结合老子的“无为”和庄子的“天人合一”思想，引导读者以辩证法思考方式认识大数据机器学习的内涵。

从技术层面上，本书一方面基于 Spark 现有的机器学习库讲解，另一方面尽量做到和现有 Spark 版本中的机器学习库解耦，突出对大数据机器学习的宏观理解，并给出典型算法的工程化实现，使更多的人轻松使用机器学习进行大数据价值挖掘，从而建立大数据机器学习工程化思维，在不必深究算法细节的前提下有效解决实际问题。本书更加强调在实际场景中的应用，并有针对性地给出了综合应用场景。

从适合读者阅读和掌握知识的结构安排上讲，本书分“基础篇”“算法篇”“综合应用篇”三个维度层层推进，便于读者在深入理解基础上根据相应的解决思路找到适合自己的方案。

本书使用的机器学习算法和应用场景都是实际业务的抽象，并基于具体业务进行实现。作为本书的延续，接下来我们会聚焦应用实践并提供更深层次的拓展，专注知识图谱的技术与应用，以及 Bot 技术与构建实战，期待相关图书能和读者尽早见面。

读者对象

（1）对大数据感兴趣的读者

伴随着大数据时代的到来，很多工作都和大数据息息相关，无论是传统行业、互联网行业，还是移动互联网行业，都必须要了解大数据，通过大数据发现自身的价值。对这部分读者来说，本书的内容能够帮助他们加深对大数据 / 机器学习及其演进趋势的理解，通过本书可以了解机器学习相关算法，以及 Spark 机器学习应用场景和存在价值，如果希望更深层次地

掌握 Spark 机器学习相关知识，本书可以作为一个很好的开始。

（2）从事大数据机器学习算法的研究人员

本书基于分类、聚类、回归、关联规则、协同过滤、降维等算法，结合异常检测、用户画像、广告点击率预估、企业征信大数据、智慧交通大数据等场景，系统地讲解了 Spark 机器学习相关知识，对从事大数据算法的研究人员来说，能够身临其境地体验各种场景，了解各类算法在不同场景下的优缺点，减少自己的研究成本。本书对生产环境中遇到的算法建模、数据挖掘等问题有很好的借鉴作用。

（3）大数据工程开发人员

大数据工程开发人员可以从本书中获取需要的机器学习算法工程化知识。对大数据工程开发人员来说，掌握并快速对算法进行工程化，是很重要的技能，本书为填补算法工程开发人员与算法研究人员之间的鸿沟、高效工作提供了更多可能。

（4）大数据架构设计人员

基于大数据的采集、存储、清洗、实时计算、统计分析、数据挖掘等是大数据架构师必备技能。他们需要对 Spark 机器学习进行了解，才能在架构设计中综合考虑各种因素，构建稳定高效的大数据架构。

如何阅读本书

本书分为三篇，共计 13 章内容。

基础篇（第 1 和 2 章），对机器学习进行概述讲解，并通过 Spark 机器学习进行数据分析。

算法篇（第 3 ～ 8 章），针对分类、聚类、回归、关联规则、协同过滤、降维等算法进行详细讲解，并进行算法建模应用实现。

综合应用篇（第 9 ～ 13 章），综合异常检测、用户画像，引出广告点击率预估，并对企业征信大数据、智慧交通大数据等场景进行实践，详细讲解基于 Spark 的大数据机器学习综合应用。

勘误和支持

由于笔者的水平有限，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。如果你有更多的宝贵意见，可以通过大数据技术交流 QQ 群 435263033，或者邮箱 datadance@163.com 联系我们，期待能够得到大家的真挚反馈，在大数据和人工智能征程中互勉共进。

致谢

感谢亲爱的搭档马海平、吕昕、向海三位大数据专家以及谭昶博士，在本书写作遇到困难的时候，我们一直互相鼓励，牺牲休息时间，坚持不放弃。

感谢大数据团队的张志勇、张龙、陈爱华、杨柳、俞祥祥、王庆庆、牛鑫、谢榭、李雅洁，以及廖攀、覃雪辉等小伙伴，你们为本书的修改贡献了宝贵的智慧，你们的参与使本书更上一层楼。

本书使用了部分互联网测试数据，包括：Stanford 的 gowalla 数据、360 的应用市场数据、UCI 的鸢尾花卉数据和裙子销售数据、数据堂的豆瓣电影评分数据、Digit 数据集、新闻 App 的用户行为数据、某运营商手机信令数据、某地图路况的道路拥堵指数数据，在这里进行特别感谢。

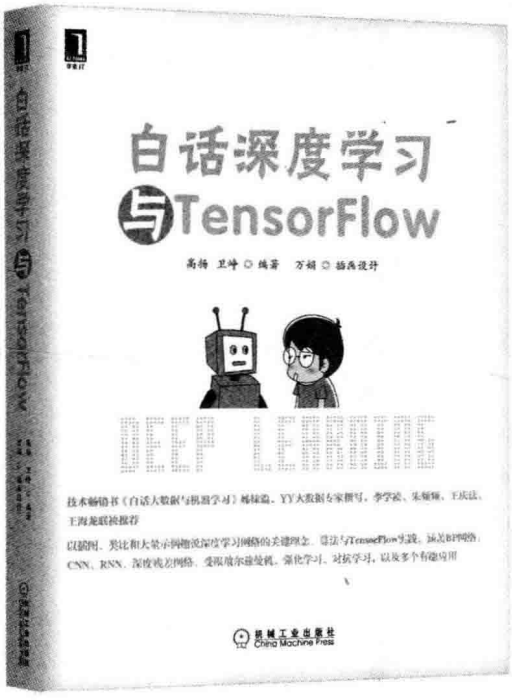
最后特别祝福本书写作期间出生的马海平家的二宝和向海家的二宝，你们的出生代表了大数据机器学习有了新的传承，也让我们的努力变得更有意义。

谨以此书献给大数据团队的小伙伴，以及众多热爱大数据机器学习技术的朋友！

于俊

2018 年 8 月

推荐阅读



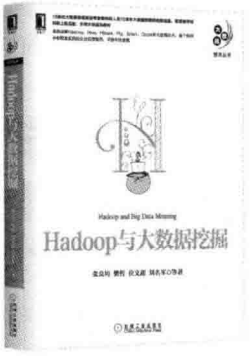
白话深度学习与TensorFlow

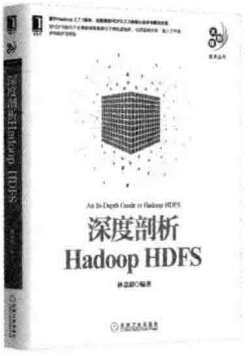
书号：978-7-111-57457-6 作者：高扬 定价：69.00元

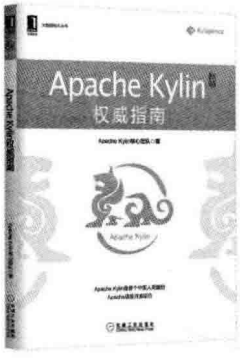
《白话大数据与机器学习》姊妹篇，YY大数据专家多年实践沉淀之作，简单易懂，轻松入门

推荐阅读


大数据治理与安全
从理论到开源实践


Hadoop与大数据挖掘


深度剖析
Hadoop HDFS


Apache Kylin
权威指南


Ceph源码分析


Kafka源码解析与实战


循序渐进学Spark


分布式实时处理系统
原理、架构与实现


大数据架构商业之路
从业务需求到技术方案

Contents 目　录

第三篇 综合应用篇