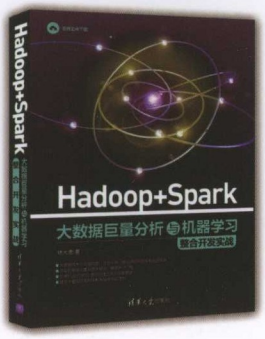
- 浅显易懂的原理介绍
- Step by Step 实机操作、范例程序详细解说
- 降低机器学习与大数据技术的学习门槛

Python+ Spark 2.0+Hadoop 机器学习与大数据实战

林大贵 著

轻松快速学会机器学习与大数据热门技术

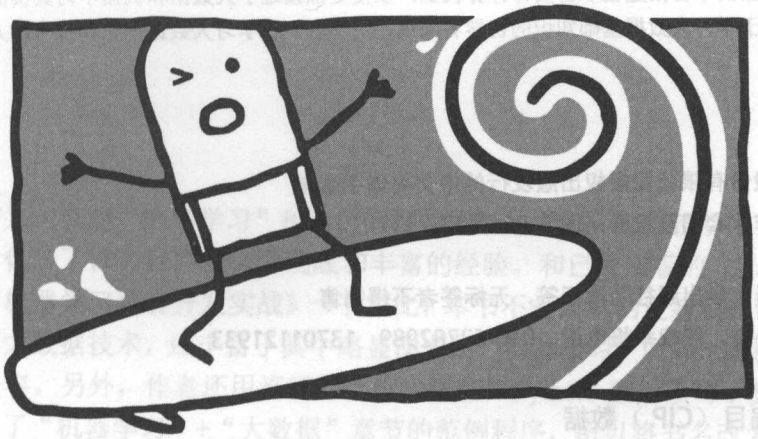
清華大學出版社



内 容 简 介

本书从浅显易懂的"大数据和机器学习"原理介绍和说明入手，讲述大数据和机器学习的基本概念，如分类、分析、训练、建模、预测、机器学习（推荐引擎）、机器学习（二元分类）、机器学习（多元分类）、机器学习（回归分析）和数据可视化应用。为降低读者学习大数据技术的门槛，书中提供了丰富的上机实践操作和范例程序详解，展示了如何在单台Windows系统上通过Virtual Box虚拟机安装多台Linux虚拟机，如何建立Hadoop集群，再建立Spark开发环境。书中介绍搭建的上机实践平台并不限制于单台实体计算机。对于有条件的公司和学校，参照书中介绍的搭建过程，同样可以将实践平台搭建在多台实体计算机上，以便更加接近于大数据和机器学习真实的运行环境。

本书非常适合于学习大数据基础知识的初学者阅读，更适合正在学习大数据理论和技术的人员作为上机实践用的教材。



Python+ Spark 2.0+Hadoop

机器学习与大数据实战

林大贵 著

清華大學出版社
北京

内 容 简 介

本书从浅显易懂的“大数据和机器学习”原理说明入手，讲述大数据和机器学习的基本概念，如分类、分析、训练、建模、预测、机器学习（推荐引擎）、机器学习（二元分类）、机器学习（多元分类）、机器学习（回归分析）和数据可视化应用等。书中不仅加入了新近的大数据技术，还丰富了“机器学习”内容。

为降低读者学习大数据技术的门槛，书中提供了丰富的上机实践操作和范例程序详解，展示了如何在单机Windows系统上通过Virtual Box虚拟机安装多机Linux虚拟机，如何建立Hadoop集群，再建立Spark开发环境。书中介绍搭建的上机实践平台并不限制于单台实体计算机。对于有条件的公司和学校，参照书中介绍的搭建过程，同样可以实现将自己的平台搭建在多台实体计算机上，以便更加接近于大数据和机器学习真实的运行环境。

本书非常适合于学习大数据基础知识的初学者阅读，更适合正在学习大数据理论和技术的人员作为上机实践用的教材。

图书在版编目（CIP）数据

Python+Spark 2.0+Hadoop 机器学习与大数据实战 / 林大贵著. —北京：清华大学出版社，2018（2018.4重印）
ISBN 978-7-302-49073-9

Ⅰ. ①P… Ⅱ. ①林… Ⅲ. ①软件工具－程序设计②数据处理软件 Ⅳ. ①TP311.561②TP274

中国版本图书馆 CIP 数据核字（2017）第 296017 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：33.75　　字　数：864 千字
版　　次：2018 年 1 月第 1 版　　印　次：2018 年 4 月第 2 次印刷
印　　数：3001 ~ 5000
定　　价：99.00 元

产品编号：073908-01

序

本书将方兴未艾的“机器学习”和热门的“大数据分析”技术与应用在一本书中融会贯通地娓娓道来，体现了作者深厚的技术功底和丰富的经验。和已经出版的《Hadoop+Spark 大数据巨量分析与机器学习整合开发实战》一书相比，本书不是简单的更新和升级，而是在原有的基础上增加了大数据技术，还丰富了其中略显薄弱的“机器学习”内容，增加了 4 章都和机器学习有关的内容。另外，作者还用流行的“胶水语言”Python 重新改写了另一本书中的范例程序，并添加了“机器学习”+“大数据”章节的范例程序，所以将书名改为“Python + Spark 2.0 + Hadoop 机器学习与大数据实战”，更加突出“机器学习”，并且强调范例程序是运用更加流行的 Python 语言来编写的。

在因特网、社交媒体、电子商务等交叉发展和呼应下，“网络”这个巨人已经拥有了难以计数的海量数据，虽有传统结构化的数据、半结构化的数据，但更多的是非结构化的数据。这些貌似杂乱无章、毫无意义的海量数据是一座等待发掘的巨大“金矿”。这些海量数据中蕴含着极为丰富的人类知识库，是一笔巨大的信息资产。随着云计算时代的来临，对这些原本很难收集整理的大数据进行及时甚至是实时分析和处理并加以有效利用就不再是“海市蜃楼”了。

与大数据相关的内容不外乎三方面：大数据理论，大数据分析和处理的技术（机器学习为核心技术），大数据的实践应用。在与大数据有关的出版物中，偏重于理论教学和技术介绍一类的比较多，而偏重于上机实践和自学的书比较少见。因此，本书非常适合“机器学习和大数据分析”的初学者和正在学习这个领域技术的人员作为学习和上机实践用的教材。

本书不是对原理进行纯理论的阐述，而是提供了丰富的上机实践操作和范例程序，从而降低了读者学习“机器学习和大数据分析”的门槛。对于需要直接上机实践的学习者而言，本书更像是一本学习实践和实战开发的上机手册。书中首先展示了如何在单台 Windows 系统上通过 Virtual Box 虚拟机安装多台 Linux 虚拟机，而后建立 Hadoop 集群，再建立 Spark 开发环境。搭建这个上机实践的平台并不限制于单台实体计算机，主要是考虑个人读者上机实践的实际条件和环境。对于有条件的公司和学校，参照这个搭建过程，同样可以将实践平台搭建在多台实体计算机上。另外，现在很多大专院校都开设了 Python 程序设计语言的课程，所以本书的所有范例程序都用 Python 语言重新改写了，非常接“地气”。

在搭建好“机器学习和大数据分析”上机实践的软硬件环境之后，就可以在各章节的学习中结合本书提供的范例程序逐一设置、修改、调试和运行，从中学到“机器学习和大数据分析”

实践应用中核心技术的真谛——对大数据进行高效的“智能加工”，萃取大数据中蕴含的“智慧和知识”，实现数据的“增值”，并最终将其应用于实际工作或者商业中。

对于企业在商业应用中的“机器学习和大数据分析”核心系统，需要运用商业公司的解决方案作为引擎。在中国市场活跃的国际和国内著名公司也提供了相当好的解决方案，比如Cloudera 对 Spark ml 提供完整的支持、星环科技基于 Spark 自主研发了机器学习产品 Discover。

大数据与云计算的关系密不可分，涉及众多关键技术，比如分布式处理、分布式数据库和云存储、虚拟化技术等，但是它们不是本书的重点，所以这里并未深入讲解。建议需要深入学习这方面内容的读者去寻找相关出版物，结合本书的实践来丰富和完善自己的大数据知识体系。

资深架构师 赵军

2017 年 11 月

前言

机器学习是近二十来年兴起的多领域学科，机器学习算法可从数据中建立模型，并利用模型对未知数据进行预测。机器学习技术不断进步，应用相当广泛，例如推荐引擎、定向广告、需求预测、垃圾邮件过滤、医学诊断、自然语言处理、搜索引擎、诈骗侦测、证券分析、视觉辨识、语音识别、手写识别等。

近年来 Google、Facebook、Microsoft、IBM 等大公司全力投入机器学习研究与应用。以 Google 为例，Google 已经将机器学习运用到垃圾邮件判断、自动回复、照片分类与搜索、翻译、语音识别等功能上。同时，各大主流 Hadoop 发行版公司加强了对机器学习的投入，比如 Cloudera 对 spark ml 的完整支持、星环科技基于 Spark 自主研发的机器学习产品 Discover。在不知不觉中，机器学习已经让日常生活更为便利。

为什么近年来机器学习变得如此热门，各大公司都争相投入？因为机器学习需要大量数据进行训练。大数据的兴起带来了大量的数据以及可存储大量数据的分布式存储技术，例如 Hadoop HDFS、NoSQL……还有分布式计算可进行大量运算，例如 Spark 基于内存的分布式计算框架/架构，可以大幅提升性能。

本书的主题是 Python+Spark+Hadoop 机器学习与大数据分析。使用 Python 开发 Spark 应用程序，具有多重优势：不仅可以享有 Python 语言特性所带来的好处，即程序代码简明、较易学习、高生产力等，再加上 Spark 基于内存的分布式计算框架/架构，还可以大幅提升性能，非常适合需要多次重复运算的机器学习算法，并且 Spark 还可以存取 Hadoop HDFS 分布式存储的大量数据。

本书希望能够用浅显易懂的原理介绍和说明以及上机实践操作、范例程序来降低机器学习与大数据技术的学习门槛，带领读者进入机器学习和大数据的领域。当然，整个机器学习与大数据的生态系统非常庞大，需要学习的东西很多。读者通过本书学习，对机器学习和数据有了基本的概念后就比较容易踏入这个领域了，以便深入研究其他的相关技术。

林大贵

本书章节与范例程序介绍

本书特色

提供了大量上机实践操作与范例程序。

➢ 上机实践操作

一般人可能会认为机器学习和大数据分析需要很多台机器的环境才能学习，实际上通过本书使用 Virtual Box 虚拟机的方法就能在自家的计算机上演练建立 Hadoop 集群以及 Python Spark 开发环境。同时，上机实践操作介绍了 Hadoop MapReduce 与 HDFS 的基本概念，以及 Spark RDD、DataFrame、Spark SQL 与 MapReduce 的基本概念。

➢ 范例程序

以实际范例程序来学习程序设计是最有效率的学习方式，因此本书使用实际的数据集，配合范例程序代码来介绍各种机器学习的算法，并示范如何获取数据、训练数据、建立模型、预测结果，由浅入深地介绍 Python Spark 机器学习。

本书章节内容及上机实践操作与范例程序介绍

➢ 基本概念介绍

章节	章节名称	说明
1	Python Spark 机器学习与 Hadoop 大数据	介绍机器学习、Spark 基本概念、Python 开发 Spark 机器学习与大数据应用、Spark ML Pipeline 机器学习流程、大数据定义、Hadoop 基本概念、HDFS、MapReduce 等基本原理

➢ Hadoop 的安装

章节	章节名称	说明
2	Virtual Box 虚拟机软件的安装	**上机实践操作** 安装 Virtual Box 虚拟机，让你可以在 Windows 系统上安装多台 Linux 虚拟机
3	Ubuntu Linux 操作系统的安装	**上机实践操作** 在 Virtual Box 虚拟机上安装 Ubuntu Linux 操作系统

（续表）

章节	章节名称	说明
4	Hadoop Single Node Cluster 的安装	**上机实践操作** 在 Ubuntu Linux 的操作系统上安装单台机器的 Hadoop Single Node Cluster
5	Hadoop Multi Node Cluster 的安装	**上机实践操作** 在 Ubuntu Linux 的操作系统上安装多台机器 Hadoop Multi Node Cluster，并介绍 Hadoop Resource-Manager 与 NameNode HDFS Web 界面

➢ Hadoop 的基本功能

章节	章节名称	说明
6	Hadoop HDFS 命令	**上机实践操作** 示范如何使用 HDFS 命令，并介绍 Hadoop HDFS Web 界面
7	Hadoop MapReduce	WordCount.java 范例程序 介绍 Hadoop MapReduce 原理，示范如何使用 Hadoop MapReduce 计算文章内的每一个单词（或字）出现的次数

➢ Spark 的基本功能介绍

章节	章节名称	说明
8	Python Spark 的安装与介绍	**上机实践操作** 示范如何安装 Python Spark，并在 pyspark“终端”程序界面中在本地以 Hadoop YARN-client 或 Spark Stand Alone 模式来运行 Python Spark 程序
9	在 IPython Notebook 运行 Python Spark 程序	**上机实践操作** 示范如何安装 Aanconda Python 软件包，在 IPython Notebook 中在本地以 Hadoop YARN-client 或 Spark Stand Alone 模式来运行 Python Spark 程序
10	Python Spark RDD	**上机实践操作** Spark 基本功能 RDD（Resilient Distributed Dataset，弹性分布式数据集）的基本运算
11	Python Sparkr 的集成开发环境	**上机实践操作** 示范如何安装 eclipse+pyDev 集成开发环境来运行 Python Spark 程序 WordCount.py 范例程序 示范如何在 eclipse 中在本地以 Hadoop YARN-client 或 Spark Stand Alone 模式运行 Python Spark 程序

➢ 以 RDD 为基础的 Spark MLlib 机器学习

章节	章节名称	说明
12	Python Spark 创建推荐引擎	IPython Notebook 范例程序 通过 IPython Notebook 交互式界面，示范如何使用 Spark MLlib 命令建立电影的推荐引擎 RecommendTrain.py 范例程序 示范如何提取数据、训练并建立模型、存储模型 Recommend.py 范例程序 示范如何载入模型、使用模型推荐用户或电影
13	Python Spark MLlib 决策树二元分类	RunDecisionTreeBinary.py 范例程序 示范如何使用决策树二元分类分析 StumbleUpon 数据集，预测哪些网页是暂时性或可以长久存在的，并且找出最佳参数组合，提高预测准确度
14	Python Spark MLlib 逻辑回归二元分类	RunLogisticRegressionWithSGDBinary.py 范例程序 示范如何使用逻辑回归二元分类分析 StumbleUpon 数据集，预测哪些网页是暂时性或可以长久存在的，并找出最佳参数组合，提高预测准确度
15	Python Spark MLlib 支持向量机 SVM 二元分类	RunSVMWithSGDBinary.py 范例程序 示范如何使用支持向量机 SVM 二元分类分析 StumbleUpon 数据集，预测哪些网页是暂时性或可以长久存在的，并找出最佳参数组合，提高预测准确度
16	Python Spark MLlib 朴素贝叶斯二元分类	RunNaiveBayesBinary.py 范例程序 示范如何使用朴素贝叶斯二元分类分析 StumbleUpon 数据集，预测哪些网页是暂时性或可以长久存在的，并找出最佳参数组合，提高预测准确度
17	Python Spark MLlib 决策树多元分类	RunDecisionTreeMulti.py 范例程序 示范如何使用决策树多元分类分析 Covertype 数据集（森林覆盖植被），根据不同的土地条件可以预测该地的植被，并找出最佳参数组合，提高预测准确度
18	Python Spark MLlib 决策树回归分析	RunDecisionTreeRegression.py 范例程序 示范如何使用决策树回归分析 Bike Sharing（共享单车）数据集。根据天气假日条件，可以预测每一个小时租借的数量，并找出最佳参数组合，提高预测准确度

➢ **以 DataFrame 为基础的 Spark ML Pipeline 机器学习流程**

章节	章节名称	说明
19	Python Spark SQL、DataFrame、RDD 数据统计与可视化	IPython Notebook 范例程序 通过 IPython Notebook 交互式界面介绍并比较 Spark 数据的处理方式：DataFrame vs Spark SQL vs RDD，并且使用 Pandas 与 matplotlib 绘图
20	Spark ML Pipeline 机器学习流程二元分类	IPython Notebook 范例程序 以“StumbleUpon”数据集示范如何使用 Spark ML Pipeline 机器学习流程二元分类，预测网页是暂时性的还是长久存在的，并且使用训练验证与交叉验证找出最佳模型，提高预测准确度，最后介绍如何使用随机森林 RandomForestClassifier 分类算法进一步提高准确率
21	Spark ML Pipeline 机器学习流程多元分类	IPython Notebook 范例程序 以“森林覆盖植被”多元分类数据集示范如何使用 Spark ML Pipeline 机器学习流程多元分类，预测 Cover Type 森林覆盖分类，并且使用训练验证与交叉验证找出最佳模型，提高预测准确度
22	Spark ML Pipeline 机器学习流程回归分析	IPython Notebook 范例程序 以“Bike Sharing”数据集示范如何使用 Spark ML Pipeline 机器学习流程回归分析，预测每一小时租借总数量，并且使用训练验证与交叉验证找出最佳模型，提高预测准确度，最后介绍使用 GBT（Gradient-Boosted Trees，梯度提升决策树）进一步提高预测准确度

本书范例程序下载与安装说明

可参考附录 A 中有关本书范例程序下载与安装的说明。本书范例程序主要分为两个部分：

范例	说明
IPython Notebook 范例	第 9、10、12、13、19、20、21、22 章。按照第 9 章的说明来安装 anaconda 及其设置后才能使用这些范例程序
eclipse 范例	第 11~18 章。按照第 11 章的说明完成 eclipse 的安装与全部设置后才能执行这些范例程序

本书上机实践操作命令的整理

本书第 2 章到第 10 章使用了很多 Linux、spark-shell、SparkSQL 等命令。不过很多命令都很长，只要有一个字母打错就无法运行，这样会增加挫折感。因此我们在博客文章中整理了各个章节使用的命令，可参考如下网页：

http://www.weibo.com/hadoopsparkbook

安装或练习命令时，你可以复制博客文章中的命令，然后粘贴到“终端”程序中。这样既可以节省打字的时间，又不用担心打错字母（无法在 VirtualBox 虚拟机的 Ubuntu“终端”程序中执行复制/粘贴操作时，可参考第 3.9 节的说明设置好 VirtualBox 的共享剪贴板）。

读者服务与社区交流

在网络时代，购买本书的读者不仅可以获得本书的内容，还能通过网络社区获得更多的信息。

➢ 本书的博客

网址：http://blog.sina.com.cn/hadoopsparkbook。

我们将一些需要排列整齐、系统化的信息放在了博客文章中，还会随时更新，内容包括：

- 本书上机实践操作命令的整理。
- 本书内容或程序代码的勘误。
- 分享最新的 Hadoop 或 Spark 信息。

➢ 本书的微博

网址：http://www.weibo.com/hadoopsparkbook。

我们建立了本书的 Facebook 粉丝团，欢迎读者们加入。粉丝团会不定期贴文，分享最新的 Hadoop 或 Spark 信息，大家可以随时提问并参与交流。

➢ 百度网盘

网址：**http://pan.baidu.com/s/1i4AzAk9**（注意区分数字和英文字母大小写）

如果下载有问题，请发送电子邮件至 booksaga@126.com，邮件主题设置为“求 Python+Spark 2.0+Hadoop 机器学习与大数据实战范例程序”。

目 录