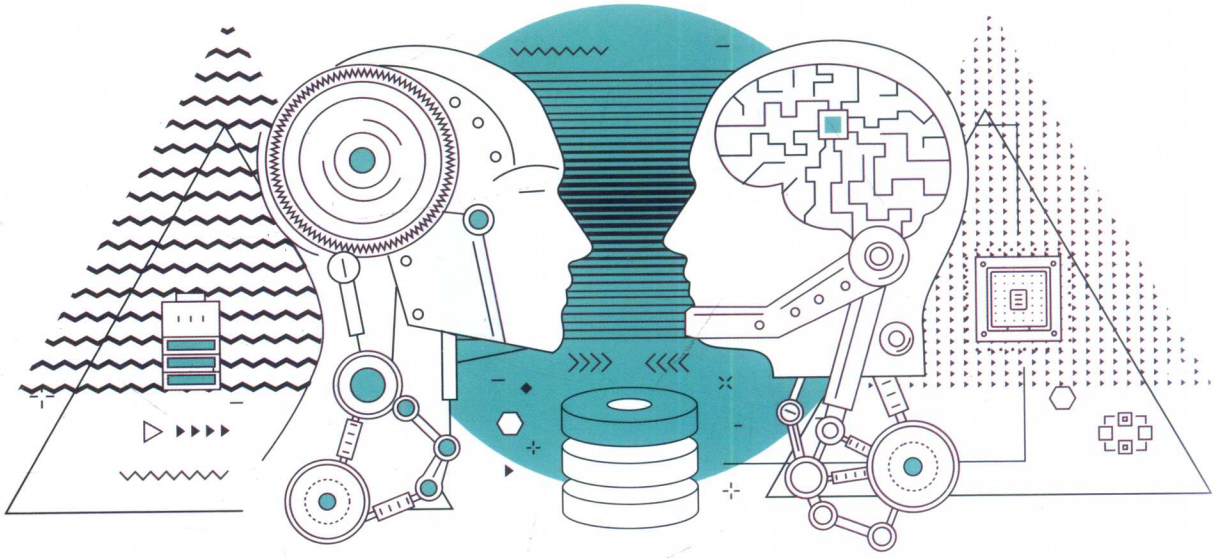
HZ BOOKS
华章科技

系统讲解Spark机器学习的技术、原理、算法和组件，以及构建Spark机器学习系统的方法、流程、标准和规范

Inside Explore the Spark Machine Learning

深度实践 Spark机器学习

吴茂贵 郁明敏 朱凤元 张粤磊◎等著

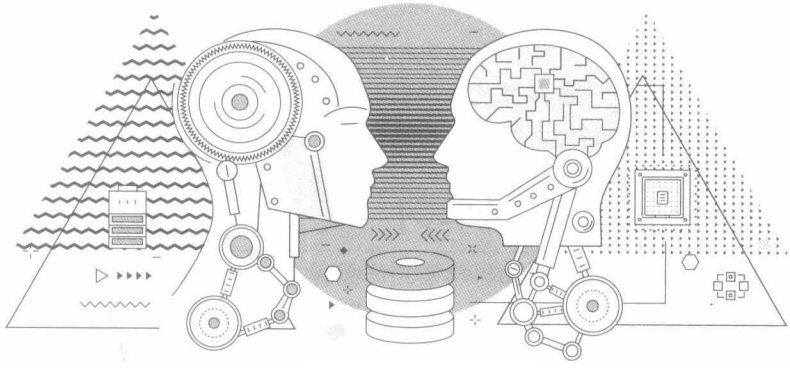


机械工业出版社
China Machine Press

Inside Explore the Spark Machine Learning

深度实践 Spark机器学习

吴茂贵 郁明敏 朱凤元 张粤磊 杨本法 著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

深度实践 Spark 机器学习 / 吴茂贵等著 . —北京：机械工业出版社，2018.1

ISBN 978-7-111-58995-2

I. 深… II. 吴… III. 数据处理软件 - 机器学习 IV. TP274

中国版本图书馆 CIP 数据核字（2018）第 015240 号

深度实践 Spark 机器学习

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：李　艺　　责任校对：殷　虹

印　　刷：北京市兆成印刷有限责任公司　　版　　次：2018 年 2 月第 1 版第 1 次印刷

开　　本：186mm × 240mm　1/16　　印　　张：15.25

书　　号：ISBN 978-7-111-58995-2　　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

Preface 前　言

为什么写这本书

大数据、人工智能正在改变或颠覆各行各业，包括我们的生活。大数据、人工智能方面的人才已经供不应求，但作为人工智能的核心——机器学习，因涉及的知识和技能比较多，除了需要具备一定的数学基础、相关业务知识外，还要求有比较全面的技术储备，如操作系统、数据库、开发语言、数据分析工具、大数据计算平台等，无形中提高了机器学习的门槛。如何降低机器学习的门槛，让更多有志于机器学习、人工智能的人能更方便或顺畅地使用、驾驭机器学习？

很多企业也正在考虑和处理这方面的问题，本书也希望借 Spark 技术在这方面做一些介绍或总结。

如何使原本复杂、专业性强的工作或操作简单化？封装是一个有效方法。封装降低了我们操作照相机的难度、降低了我们维护各种现代设备的成本，同时也提升了我们使用这些设备的效率。除封装外，过程的标准化、流程化同样是目前现代企业用于提升生产效率，降低成本，提高质量的有效方法。

硬件如此，软件行业同样如此。目前很多机器学习的开发语言或平台，正在这些方面加大力度，比如：对特征转换、特征选择、数据清理、数据划分、模型评估及优化等算法的封装；对机器学习过程的进行流程化、标准化、规范化；给大家比较熟悉的语言或工具提供 API 等方法或措施，以简化机器学习中间过程，缩短整个开发周期，使我们能更从容地应对市场的变化。Spark 在这方面可谓后来居上，尤其是最近发布的版本，明显加大了这方面的力度，我们可以从以下几个方面看出这种趋势：

1）Spark 机器学习的 API，正在由基于 RDD 过渡到基于 Dataset 或 DataFrame，基于 RDD 的 API 在 Spark2.2 后处于维护阶段，Spark3.0 后将停止使用（来自 Spark 官网）；

2）建议大家使用 Spark ML，尤其是它的 Pipeline；

3）增加大量特征选择、特征转换、模型选择和优化等算法；

4）丰富、增强 Spark 与 Java、Python、R 的 API，使其更通用。

SKLearn、Spark 等机器学习平台或工具在这方面都处于领先的地位，我们也希望借助本书，把 Spark 在这方面的有关内容介绍给大家，使大家可以少走些弯路。

此外，Spark 目前主要涉及常用机器学习算法，缺乏对一般神经网络的支持，更不用说深度学习了，这好像也是目前 Spark 的一个不足。不过好消息是：雅虎把深度学习框架 TensorFlow 与 Spark 整合在一起，而且开源了这些代码。为弥补广大 Spark 爱好者的上述缺憾，本书介绍了 TensorFlowOnSpark，其中包括深度学习框架 TensorFlow 的基础知识及使用卷积神经网络、循环神经网络等的一些实际案例。

另外，我们提供了与本书环境完全一致的免费云操作环境，这样一来是希望节约您的宝贵时间，二来是希望能通过真正的实战，给您不一样的体验和收获！总之，我们希望能使更多有志于大数据、人工智能的朋友加入这个充满生机、前景广阔的行业中来。

本书特色

本书最大特点就是注重实战！或许有读者会问，能从哪几个方面体现出来？

1）介绍了目前关于机器学习的新趋势，并分析了如何使用 Pipeline 使机器学习过程流程化。

2）简介了机器学习的一般框架 Spark、深度学习框架 Tensorflow 及把两者整合在一起的框架 TensorflowOnSpark。

3）提供可操作、便执行及具有实战性的项目及其详细代码。

4）提供与书完全一致的云操作环境，而且这个环境可以随时随地使用实操环境，登录地址为 http://www.feiguyun.com/spark/support。

5）除了代码外，还附有一些必要的架构或原理说明，便于大家能从一个更高的角度来理解把握相关问题。

总之，希望你通过阅读本书，不但可以了解很多内容或代码，更可以亲自运行或调试这些代码，从而带来新的体验和收获！

读者对象

- 对大数据、机器学习感兴趣的广大在校、在职人员。
- 对 Spark 机器学习有一定基础，欲进一步提高开发效率的人员。
- 熟悉 Python、R 等工具，希望进一步拓展到 Spark 机器学习的人。

❑ 对深度学习框架 TensorFlow 及其拓展感兴趣的读者。

如何阅读本书

本书正文共 14 章，从内容结构来看，可以分为四部分。

第一部分为第 1 ～ 7 章，主要介绍了机器学习的一些基本概念，包括如何构建一个 Spark 机器学习系统，Spark ML 主要特点，Spark ML 中流水线（Pipeline），ML 中大量特征选取、特征转换、特征选择等函数或方法，同时简单介绍了 Spark MLlib 的一些基础知识。

第二部分为第 8 ～ 12 章，主要以实例为主，具体说明如何使用 Spark ML 中 Pipeline 的 Stage，以及如何把这 Stage 组合到流水线上，最后通过评估指标，优化模型。

第三部分即第 13 章，与之前的批量处理不同，这一章主要以在线数据或流式数据为主，介绍 Spark 的流式计算框架 Spark Streaming。

第四部分即第 14 章，为深度学习框架，主要包括 TensorFlow 的基础知识及它与 Spark 的整合框架 TensorFlowOnSpark。

此外，书中的附录部分还提供了线性代数、概率统计及 Scala 的基础知识，以帮助读者更好地掌握机器学习的相关内容。

勘误和支持

除封面署名外，参加本书编写、环境搭建的人还有杨本法、张魁、刘未昕等、杨本法负责第 12 章 Spark R 的编写，张魁、刘未昕负责后台环境的搭建和维护。由于笔者水平有限，加之编写时间仓促，书中难免出现错误或不准确的地方。恳请读者批评指正，你可以通过访问 http://www.feiguyun.com 留下宝贵意见。也可以通过微信（wumg3000）或 QQ(1715408972) 给我们反馈。非常感谢你的支持和帮助。

致谢

在本书编写过程中，得到很多在校老师和同学的支持！感谢上海大学机电工程与自动化学院的王佳寅老师及黄文成、杨中源、熊奇等同学，上海理工管理学院的张帆老师，上海师大数理学院的田红炯、李昭祥老师，华师大的王旭同学，博世王冬，飞谷云小伙伴等提供的支持和帮助。

感谢机械工业出版社的杨福川、李艺老师给予本书的大力支持和帮助。

感谢参与本书编写的其他作者及提供支持的家人们，谢谢你们！

目　录 Contents

第 1 章 Chapter 1

了解机器学习

大数据、人工智能是目前大家谈论比较多的话题，它们的应用也越来越广泛，与我们的生活关系也越来越密切，影响也越来越深远，其中很多已进入寻常百姓家，如无人机、网约车、自动导航、智能家电、电商推荐、人机对话机器人等。

大数据是人工智能的基础，而使大数据转变为知识或生产力，离不开机器学习（Machine Learning），可以说机器学习是人工智能的核心，是使机器具有类似人的智能的根本途径。

本章主要介绍与机器学习有关的概念，机器学习与大数据、人工智能间的关系，机器学习常用架构及算法等，具体如下：

- 机器学习的定义
- 大数据与机器学习
- 机器学习与人工智能、深度学习
- 机器学习的基本任务
- 如何选择合适算法
- Spark 在机器学习方面的优势

1.1 机器学习的定义

机器学习是什么？是否有统一或标准定义？目前好像没有，即使在机器学习的专业领域，也没有一个被广泛认可的定义。在维基百科上对机器学习有以下几种定义：

（1）机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如

何在经验学习中改善具体算法的性能。

（2）机器学习是对能通过经验自动改进的计算机算法的研究。

（3）机器学习是用数据或以往的经验来优化计算机程序的性能标准。

一种经常引用的英文定义是：A computer program is said to learn from experience (E) with respect to some class of tasks (T) and performance(P) measure, if its performance at tasks in T, as measured by P, improves with experience E。

可以看出机器学习强调三个关键词：算法、经验、性能，其处理过程如图 1-1 所示。

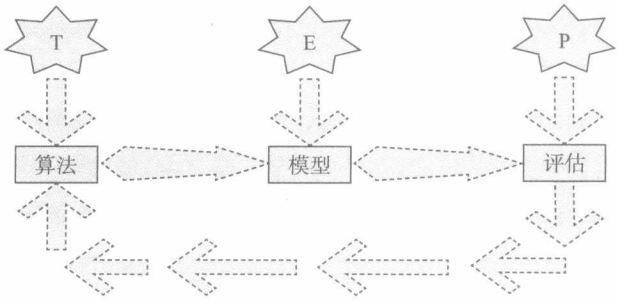


图 1-1 机器学习处理流程

图 1-1 表明机器学习是使数据通过算法构建出模型，然后对模型性能进行评估，评估后的指标如果达到要求就用这个模型测试新数据，如果达不到要求就要调整算法重新建立模型，再次进行评估，如此循环往复，最终获得满意结果。

1.2 大数据与机器学习

我们已进入大数据时代，产生数据的能力迅速增长，如互联网、移动互联网、物联网、成千上万的传感器、穿戴设备、GPS 等都会产生大量数据，存储数据、处理数据等能力也得到了几何级数的提升，如利用 Hadoop、Spark 技术为我们存储、处理大数据提供有效方法。

数据就是信息，就是依据，其背后隐含了大量不易被我们感官识别的信息、知识、规律等，如何揭示这些信息、规则、趋势，正成为当下能给企业带来高回报的热点。

而机器学习的任务，就是要在大数据量的基础上，发掘其中蕴含的有用信息。其处理的数据越多，机器学习就越能体现出优势，以前很多用机器学习解决不了或处理不好的问题，通过大数据可以得到很好解决，性能也会大幅提升，如语言识别、图像识别、天气预测等。

1.3 机器学习、人工智能及深度学习

“人工智能”和“机器学习”这两个科技术语如今广为流传，已成为当下的热词，然而，

它们有何区别？又有哪些相同或相似的地方？虽然人工智能和机器学习高度相关，但却并不尽相同。

人工智能是计算机科学的一个分支，目的是开发一种拥有智能行为的机器，目前很多大公司都在努力开发这种机器学习技术，努力让计算机学会人类的行为模式，以便推动很多人眼中的下一场技术革命——让机器像人类一样“思考”。

过去 10 年，机器学习为我们带来了无人驾驶汽车、实用的语音识别、有效的网络搜索等。接下来它将如何改变我们的生活？在哪些领域最先发力？让我们拭目以待。

有一点需要注意，对很多机器学习来说，特征提取不是一件简单的事情。在一些复杂问题上，要想通过人工的方式设计有效的特征集合，往往要花费很多的时间和精力。

作为机器学习的一个分支，深度学习解决的核心问题之一就是自动将简单的特征组合成更加复杂的特征，并利用这些组合特征解决问题。它除了可以学习特征和任务之间的关联以外，还能自动从简单特征中提取更加复杂的特征。图 1-2 展示了深度学习和传统机器学习在流程上的差异。深度学习算法可以从数据中学习更加复杂的特征表达，使得最后一步权重学习变得更加简单且有效。

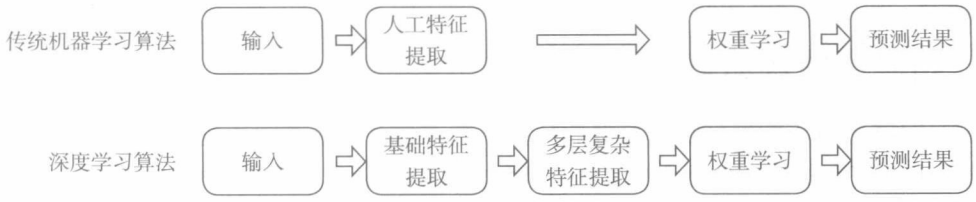


图 1-2　机器学习与深度学习流程对比

前面我们分别介绍了机器学习、人工智能及深度学习，那么它们的关系如何？

人工智能、机器学习和深度学习是紧密相关的几个领域。图 1-3 说明了它们之间的大致关系。人工智能是一类非常广泛的问题，机器学习是解决这类问题的一个重要手段，深度学习则是机器学习的一个分支。在很多人工智能问题上，深度学习的方法突破了传统机器学习方法的瓶颈，推动了人工智能领域的快速发展。

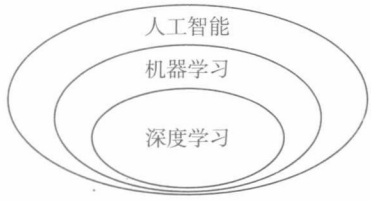


图 1-3　人工智能、机器学习与深度学习间的关系

1.4　机器学习的基本任务

机器学习基于数据，并以此获取新知识、新技能。它的任务有很多，分类是其基本任务之一。所谓分类，就是将新数据划分到合适的类别中，一般用于类别型的目标特征。如果目标特征为连续型，则往往采用回归方法。回归是对新目标特征进行预测，是机器学习中使用非常广泛的方法之一。

分类和回归，都是先根据标签值或目标值建立模型或规则，然后利用这些带有目标值的数据形成的模型或规则，对新数据进行识别或预测。这两种方法都属于监督学习。与监督学习相对的是无监督学习，无监督学习不指定目标值或预先无法知道目标值，它可以把相似或相近的数据划分到相同的组里，聚类就是解决这一类问题的方法之一。

除了监督学习、无监督学习这两种最常见的任务外，还有半监督学习、强化学习等，这里我们就不展开了，图 1-4 展示了这些基本任务间的关系。

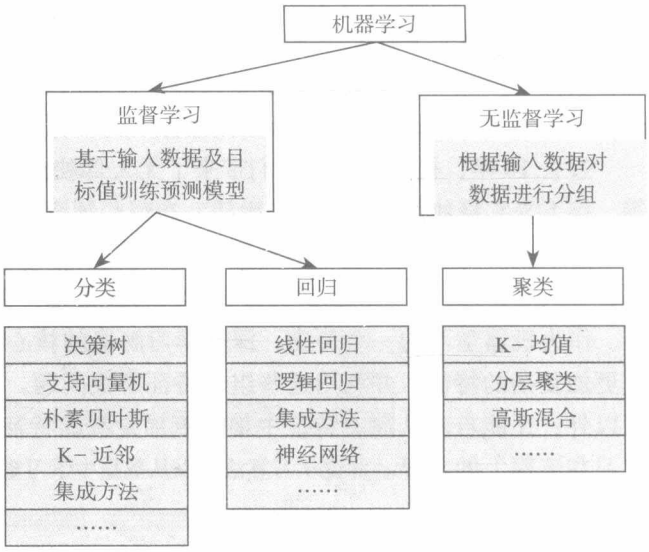


图 1-4 机器学习基本任务的关系

1.5 如何选择合适算法

当我们接到一个数据分析或挖掘的任务或需求时，如果希望用机器学习来处理，首先要做的是根据任务或需求选择合适算法，选择算法的一般步骤如图 1-5 所示。

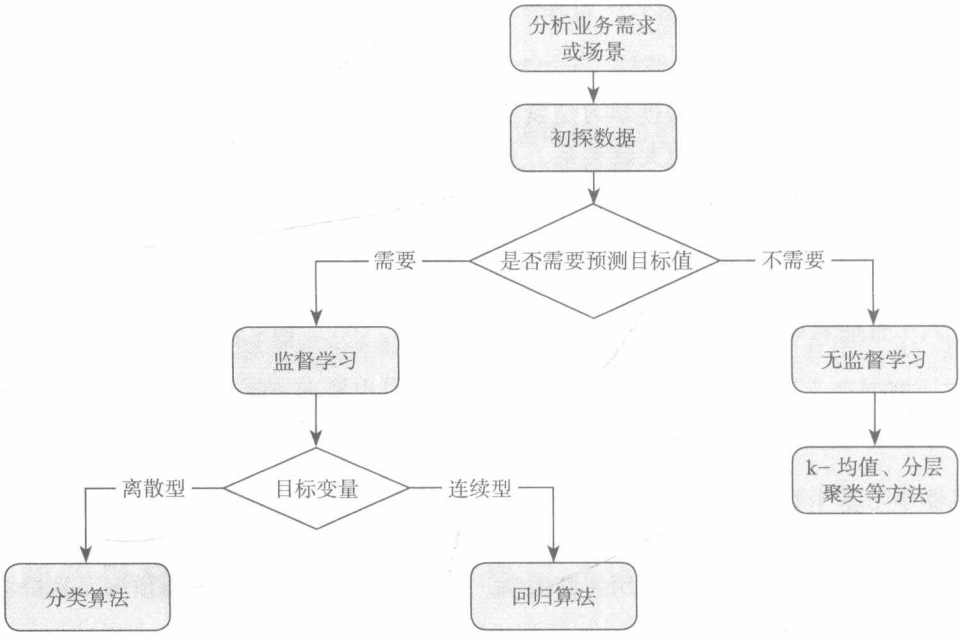


图 1-5 选择算法的一般步骤

充分了解数据及其特性，有助于我们更有效地选择机器学习算法。采用以上步骤在一定程度上可以缩小算法的选择范围，使我们少走些弯路，但在具体选择哪种算法方面，一般并不存在最好的算法或者可以给出最好结果的算法。在实际做项目的过程中，这个过程往往需要多次尝试，有时还要尝试不同算法。不过先用一种简单熟悉的方法，然后，在这个基础上不断优化，时常能收获意想不到的效果。

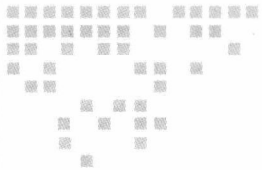
1.6　Spark 在机器学习方面的优势

在大数据基础上进行机器学习，需要处理全量数据并进行大量的迭代计算，这要求机器学习平台具备强大的处理能力。Spark 与 Hadoop 兼容，它立足于内存计算，天然适用于迭代式计算。Spark 是一个大数据计算平台，其具体有以下优势：

- 完整的大数据生态系统：大家熟悉的 SQL 式操作组件 Spark SQL，功能强大、性能优良的机器学习库 Spark MLlib，用于图像处理的 SparkGraphx 及用于流式处理的 SparkStreaming 等。
- 高性能的大数据计算平台：因为数据被加载到集群主机的分布式内存中，所以数据可以被快速转换迭代，并缓存后续的频繁访问需求。基于内存运算，Spark 可以比 Hadoop 快 100 倍，在磁盘中运算也比 Hadoop 快 10 倍左右。
- 与 Hadoop、Hive、HBase 等无缝连接：Spark 可以直接访问 Hadoop、Hive、HBase 等的数据，同时也可使用 Hadoop 的资源管理器。
- 易用、通用、好用：Spark 编程非常高效、简洁，支持多种语言的 API，如 Scala、Java、Python、R、SQL 等，同时提供类似于 Shell 的交互式开发环境 REPL。

1.7　小结

本章简单介绍了机器学习与大数据、人工智能的关系，同时也介绍了机器学习的一些基本任务和如何选择合适算法等问题。在选择机器学习平台时，我们着重介绍了 Spark 这样一个大数据平台的集大成者，它有很多优势，而且得到了很多企业的青睐。Spark 是本书的主要介绍对象，下一章我们将介绍如何构建一个 Spark 机器学习系统。



Chapter 2 第 2 章

构建 Spark 机器学习系统

构建机器学习系统的方法，根据业务需求和使用工具的不同，可能会有些区别，不过主要流程差别不大，基本包括数据抽取、数据探索、数据处理、建立模型、训练模型、评估模型、优化模型、部署模型等阶段。在构建系统前，我们需要考虑系统的扩展性，与其他系统的整合，系统升级及处理方式等。本章我们主要介绍基于 Spark 机器学习的架构设计或系统构建的一般步骤，以及需要注意的一些问题。

构建 Spark 机器学习系统的一般步骤如下：

- 介绍系统架构
- 启动集群
- 加载数据
- 探索数据
- 数据预处理
- 构建模型
- 模型评估
- 模型优化
- 模型保存

2.1 机器学习系统架构

Spark 发展非常快，到我们着手编写本书时，Spark 已升级为 2.1 版。自 2.0 以后，Spark 大大增强了数据流水线的内容。数据流水线的思路与 SKLearn 非常相似，我想这种思

路或许是未来的一个趋势，使机器学习的流程标准化、规范化、流程化，将很多原来需要自己编写的代码封装成可直接调用的模块或函数，模型评估、调优这些任务也可实现了更高的封装，大大降低机器学习的门槛。

Spark 机器学习系统的架构图如图 2-1 所示，其中数据探索与预处理、训练及测试算法或建模阶段可以组装成流水线方式，模型评估及优化阶段可以采用自动化方式。

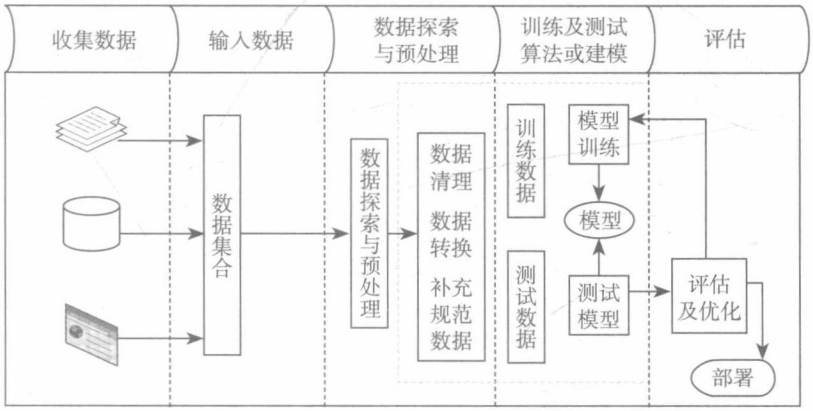


图 2-1　Spark 机器学习系统的架构图

2.2　启动集群

对于 Spark 集群的安装配置，这里不做详细介绍，对 Spark 集群的安装配置感兴趣的读者，可参考由我们编写的《自己动手做大数据系统》。

常见的 Spark 运行方式有本地模式、集群模式。本地模式所有的处理都运行在同一个 JVM 中，而后者，可以运行在不同节点上。具体运行模式如表 2-1 所示。

表 2-1　Spark 运行模式

运行模式	含义
local	使用单线程在本机上运行 Spark 任务
local[K]	使用 *K* 个工作线程在本机上运行 Spark。*K* 值最好小于等于 CPU 的核数
local[*]	使用和 CPU 核数相同的线程数在本机上运行 Spark
spark://host:port	Standalone 模式运行，host 是集群中 Master 节点机器名，port 是端口号，默认是 7077
mesos://host:port	连接到 Mesos 集群运行任务，port 端口号默认是 5055
yarn-client	以客户端方式连接到 YARN 集群
yarn-cluster	以集群方式连接到 YARN 集群

本书主要以 Spark Standalone（独立模式）为例，如果想以其他模式运行，只要改动对应参数即可。

Spark 支持 Scala 或 Python 的 REPL（Read-Eval-Print-Loop，交互式 shell）来进行交互