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。

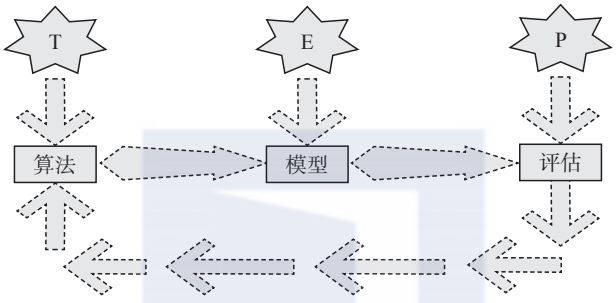

图 1-1　机器学习处理流程

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。

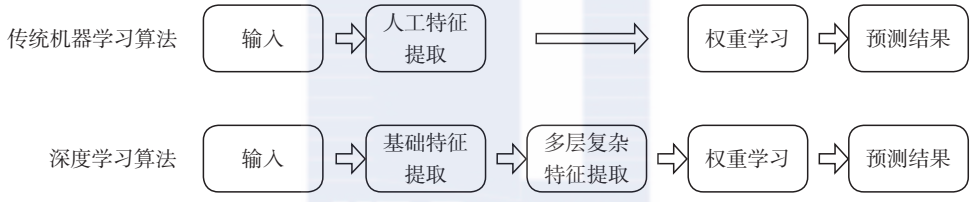

图 1-2　机器学习与深度学习流程对比

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。

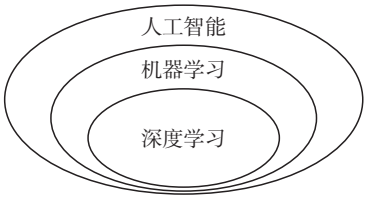

图 1-3　人工智能、机器学习与深度学习间的关系

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。

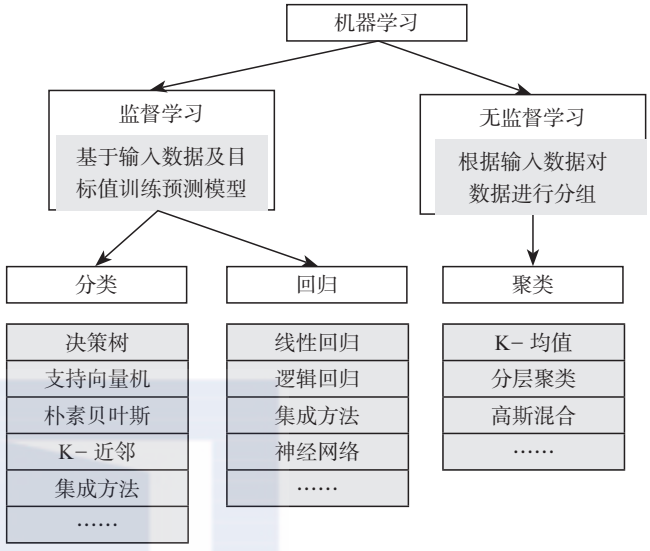

图 1-4　机器学习基本任务的关系

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。

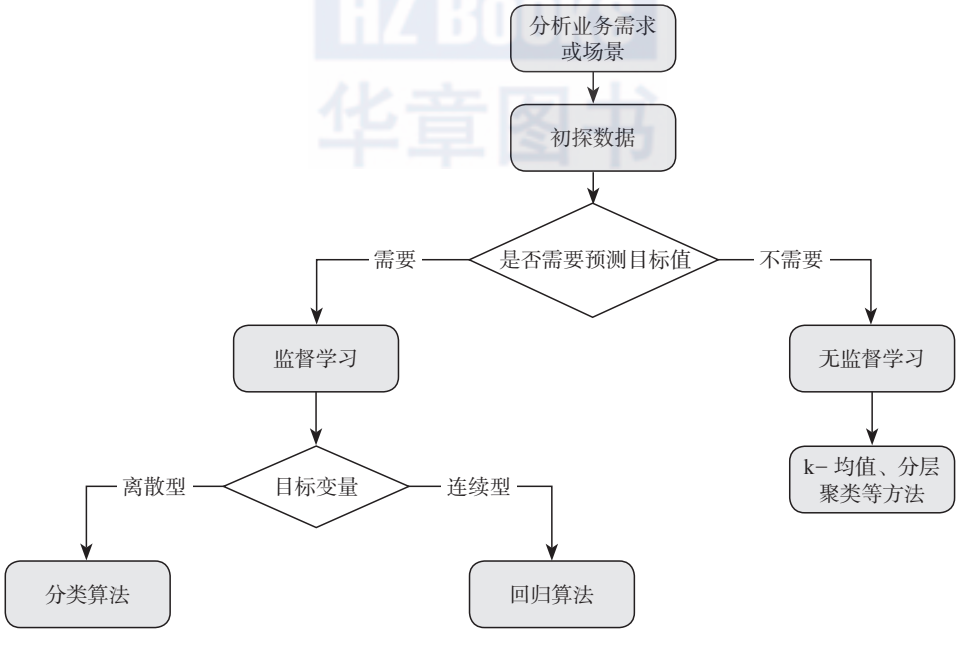

图 1-5　选择算法的一般步骤

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。