第 1 章

CHAPTER 1

机器学习

机器学习（Machine Learning，ML）是一门多领域交叉学科，它涉及概率论、统计学、计算机科学等学科。机器学习的概念就是输入海量训练数据对模型进行训练，使模型掌握数据所蕴含的潜在规律，进而对新输入的数据进行准确的分类或预测。机器学习流程如图 1-1 所示。

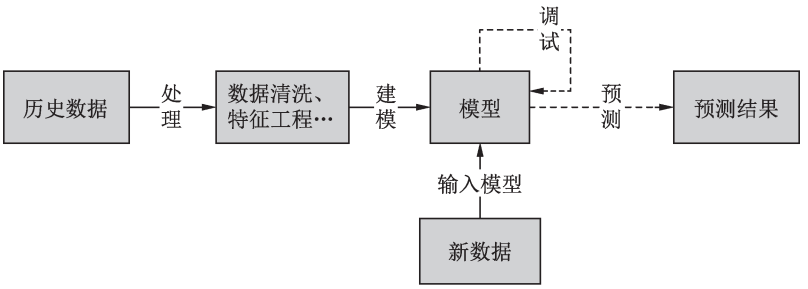

图 1-1　机器学习流程

1.1　机器学习的分类

根据数据类型的不同，对一个问题的建模有不同的方式。在机器学习或者人工智能领域，人们首先会考虑算法的学习方式，将算法按照学习方式分类，可以让人们在建模和算法选择时考虑能根据输入数据来选择最合适的算法获得最好的结果。图 1-2 为监督学习、无监督学习及强化学习的实际应用领域。

图 1-2　三种机器学习的应用领域

1.1.1 用监督学习预测未来

监督学习目标是从有标签的训练数据中学习模型，以便对未知或未来的数据作出预测。一个典型的监督学习流程如图 1-3 所示，先为机器学习算法对打过标签的训练数据提供拟合预测模型，然后用该模型对未打过标签的新数据进行预测。

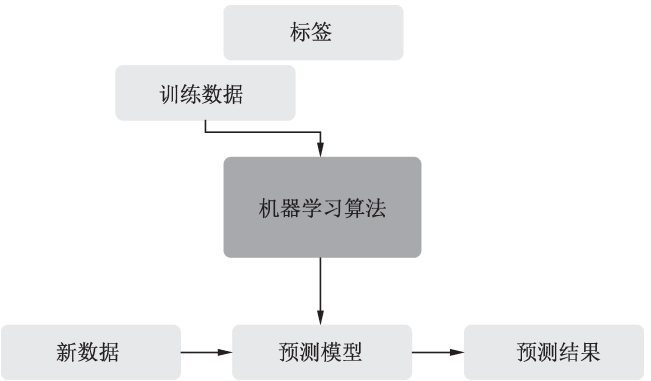

图 1-3 监督学习流程

以垃圾邮件过滤为例，可以采用监督学习算法在打过标签的电子邮件的语料库上训练模型，然后用该模型来预测新邮件是否属于垃圾邮件。带有离散分类标签的监督学习任务也被称为分类任务。监督学习的另一个子类被称为回归，其结果信号是连续的数值。

1. 分类

监督学习的一个分支是分类，分类的目的是根据过去的观测结果来预测新样本的分类标签。这些分类标签是离散的无序值。前面提到的邮件垃圾检测就是典型的二元分类任务，机器学习算法学习规则用于区分垃圾邮件和非垃圾邮件。

接下来将通过 30 个训练样本介绍二元分类任务（见图 1-4）的概念。30 个训练样本中有 15 个标签为负类（–），15 个标签为正类（+）。该数据集是二维的，这说明每个样本都与 x_1 和 x_2 的值相关，即可通过监督学习算法来学习一个规则：用一条虚线来表示决策边界，用于区分两类数据，并根据 x_1 和 x_2 的值为新数据分类。

值得注意的是，类标签集并不都是二元的，经过监督学习算法学习所获得的预测模型可以将训练数据集中出现过的任何维度的类标签分配给，还未打标签的新样本。手写字符识别是多元分类任务的典型实例。首先，收集包含字母表中所有字母的多个手写实例所形成的训练数据集。字母（A、B、C 等）代表要预测的不同的无序类别或类标签。然后，当用户通过输入设备提供新的手写字符时，预测模型能够以某一准确率将其识别为字母表中的正确字母。然而，该机器学习系统却无法正确地识别 0 到 9 的任何数字，因为它们并不是训练数据集中的一部分。

2. 回归

第二类监督学习是对连续结果的预测，也称为回归分析。回归分析包括一些预测（解释）变量和一个连续的响应变量（结果），用于寻找那些变量之间的关系，从而能够预测结果。

注意，机器学习领域的预测变量通常被称为“特征”，而响应变量通常被称为“目

标变量”。

图 1-5 为线性回归图，给定特征变量 x 和目标变量 y，对数据进行线性拟合，最小化样本点和拟合线之间的距离（常用距离为平均平方距离）。

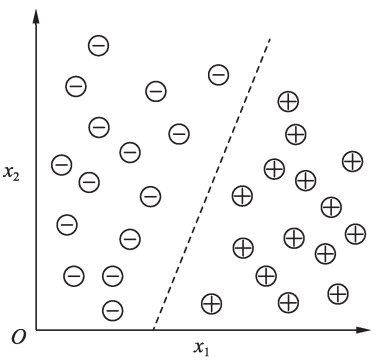

图 1-4　二元分类任务

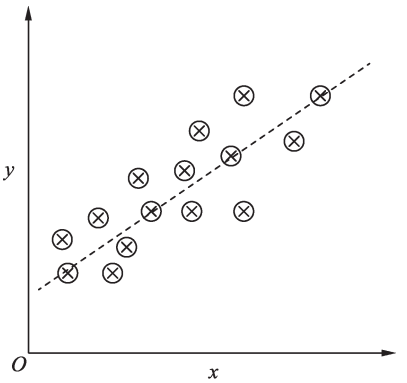

图 1-5　线性回归图

这时可以用从该数据中学习到的截距和斜率来预测新数据的目标变量。

1.1.2　用无监督学习发现隐藏结构

监督学习训练模型时，事先知道正确的答案。但是无监督学习处理的是无标签或结构未知的数据。用无监督学习技术，可以在没有已知结果变量或奖励函数的指导下，探索数据结构来提取有意义的信息。

1. 寻找子群

聚类是探索性的数据分析技术，可以在事先不了解成员关系的情况下，将信息分成有意义的子群（集群）。在分析过程中为出现的每个集群定义一组对象，集群的成员之间具有一定程度的相似性，但与其他集群中对象的差异性较大，这就是为什么聚类有时也被称为无监督分类。

聚类是一种构造信息和从数据中推导出有意义关系的有用技术，图 1-6 解释了如何应用聚类把无标签数据根据 x_1 和 x_2 的相似性分成三组。

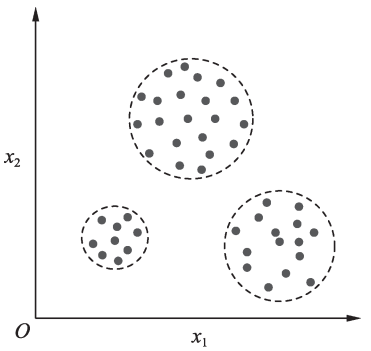

图 1-6　聚类分析法

2. 压缩数据

无监督学习的另一个常用子类是降维，我们经常要面对高维数据，然而，高维数据的每个观察通常都伴随着大量的测量数据，这对有限的存储空间和机器学习算法的计算性能提出了挑战。

无监督降维是特征预处理中一种常用的数据去噪方法，它不仅可以降低某些算法对预测性能的要求，还可以在保留大部分相关信息的同时将数据压缩到较小维数的子空间上。有时降维有利于数据的可视化，如为了通过二维散点图、三维散点图或直方图实现数据的可视化，可以把高维特征数据集映射到一维、二维或三维特征空间。图 1-7 展示了一个采用非线性降维将三维特征空间压缩成新的二维特征子空间的实例。

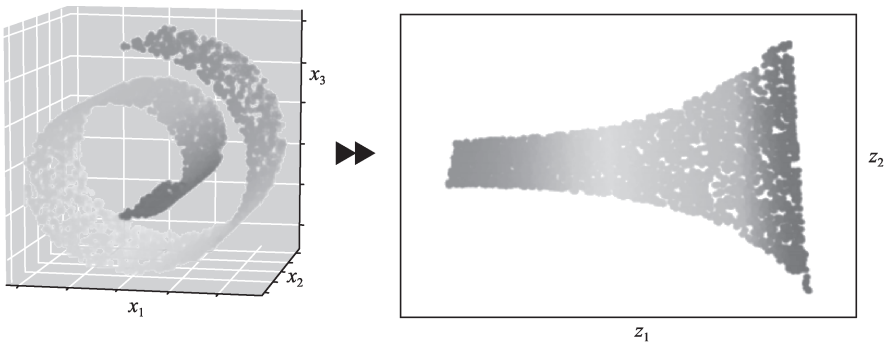

图 1-7　非线性降维效果

1.1.3　用强化学习解决交互问题

强化学习的目标是开发一个系统（智能体），通过与环境的交互来提高其性能。当前环境状态的信息通常包含奖励信号，可以把强化学习看作一个与监督学习相关的领域。但强化学习的反馈并非标定过的正确标签或数值，而是奖励函数对行动度量的结果。智能体可以与环境交互完成强化学习，并通过探索性的试错或深思熟虑的规划来最大化这种奖励。

强化学习的常见实例是国际象棋。智能体根据棋盘的状态或环境来决定一系列的行动，奖励定义为比赛的输或赢，如图 1-8 所示。

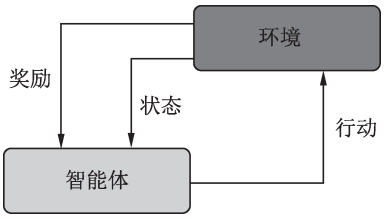

图 1-8　强化学习过程

强化学习有多种不同的子类，然而，一般模式是强化学习智能体试图通过与环境的一系列交互来最大化奖励。每种状态都可以与正或负的奖励相关联，奖励可以被定义为完成一个总目标，如赢棋或输棋。例如，国际象棋每走一步的结果都可以认为是环境的一个不同状态。为进一步探索国际象棋的实例，观察棋盘上与赢棋相关联的某些状况，如吃掉对手的棋子或威胁“皇后”。也注意棋盘上与输棋相关联的状态，如在接下来的回合中输给对手一个棋子。下棋只有到了结束时才会得到奖励（无论是正面的赢棋还是负面的输棋）。另外，最终的奖励取决于对手的表现，如对手可能牺牲了“皇后”，但最终赢棋了。

强化学习根据学习一系列的行动来最大化总体奖励，这些奖励可能即时获得，也可能延后获得。

1.1.4　分类和回归术语

分类和回归都包含很多专业术语，这些术语在机器学习领域都有确切的定义。

（1）样本（sample）或输入（input）：进入模型输出的数据点。

（2）目标（target）：真实值，对于外部数据源，理想情况下，模型应该能够预测出目标。

（3）预测（prediction）或输出（output）：从模型出来的结果。

（4）预测误差（prediction error）或损失值（loss value）：模型预测与目标之间的距离。

（5）标签（label）：分类问题中类别标注的具体例子。例如，如果 abcd 号图像被标注为包含类别“狗”，那么“狗”就是 abcd 号图像的标签。

（6）类别（class）：分类问题中供选择的一组标签。例如，对猫狗图像进行分类时，“狗”

和“猫”就是两个类别。

（7）真值（ground-truth）或标注（annotation）：数据集的所有目标，通常由人工收集。

（8）二分类（binary classification）：一种分类任务，每个输入样本都应被划分到两个互斥的类别中。

（9）多分类（multi classification）：一种分类任务，每个输入样本都应被划分到两个以上的类别中，如手写数字分类。

（10）多标签分类（multilabel classification）：一种分类任务，每个输入样本都可以分配多个标签。例如，如果一幅图像中可能既有猫又有狗，那么应该同时分配“猫”标签和“狗”标签。每幅图像的标签个数通常是可变的。

（11）标量回归（scalar regression）：目标是连续标量值的任务。预测房价就是一个很好的例子，不同的目标价格形成一个连续的空间。

（12）向量回归（vector regression）：目标是一组连续值（如一个连续向量）的任务。如果对多个值（如图像边界框的坐标）进行回归，那就是向量回归。

（13）小批量（mini-batch）或批量（batch）：模型同时处理的一小部分样本，样本数通常取 2 的幂次方（通常为 8 ~ 128），这样便于 GPU 上的内存分配。训练时，小批量用来为模型权重计算一次梯度下降更新。

1.2　选择正确的算法

有些问题非常具体，需要采取独特的方法。例如，如果使用推荐系统，这是一种常见的机器学习算法，解决的是非常具体的问题。而其他问题非常开放，则需要采用试错的方法去解决。监督学习、分类和回归都是非常开放的，它们可以用于异常检测，或者用来打造更通用的预测模型。有几个因素会影响机器学习选择正确的算法。可通过以下几方面缩小选择机器学习算法的范围。

1. 数据科学过程

在开始审视不同的机器学习算法前，必须对数据、面临的问题和局限有一个清晰的认识。

2. 了解对象数据

在决定使用哪种算法时，必须考虑数据的类型。一些算法只需要少量样本，另一些则需要大量样本，或某些算法只能处理特定类型的数据。例如，朴素贝叶斯算法与分类数据相得益彰，但对缺失数据完全不敏感。

因此，需要做到以下几点。

（1）了解数据。

首先，需要查看汇总统计信息和可视化，原因在于：

① 百分位数可以帮助确定大部分数据的范围；

② 平均值和中位数可以描述集中趋势；

③ 相关性可以指明紧密的关系。

其次，可视化数据，原因在于：

① 箱形图可以识别异常值；

② 密度图和直方图显示数据的分布；

③ 散点图可以描述双变量关系。

（2）清理数据。

清理数据的过程如下。

① 处理缺失数据。缺失数据对一些模型的影响很大，即便是能够处理缺失数据的模型，也可能受到影响（某些变量的缺失数据会导致糟糕的预测）。

② 选择如何处理异常值。异常值在多维数据中非常常见。一些模型对异常值的敏感性低于其他模型。通常，树模型对异常值的存在不太敏感。但是，回归模型或任何尝试使用等式的模型肯定会受到异常值的影响。异常值可能是数据收集不正确的结果，也可能是真实的极端值。

（3）增强数据。

增强数据可通过以下过程进行。

① 特征工程是从原始数据到建模可用数据的过程，这有几个目的：

- 使模型更易于解释（如分箱）；
- 抓取更复杂的关系（如神经网络）；
- 减少数据冗余和维度（如主成分分析）；
- 重新缩放变量（如标准化或正则化）。

② 不同的模型可能对特征工程有不同的要求。可根据输入数据或输出数据对问题进行归类。

第一步，按输入数据分类。

- 如果是标签数据，那就是监督学习问题；
- 如果是无标签数据，想找到结构，那就是非监督学习问题；
- 如果想通过与环境互动来优化一个目标函数，那就是强化学习问题。

第二步，按输出数据分类。

- 如果是一个数字，那就是回归问题；
- 如果是一个类，那就是分类问题；
- 如果是一组输入数据，那就是聚类问题；
- 如果想检测一个异常，那就是异常检测。

影响模型选择的因素包括：

- 模型是否符合商业目标；
- 模型需要多大程度的预处理；
- 模型的准确性；
- 模型的可解释性；
- 模型的速度，即建立模型需要多久、模型作出预测需要多久；
- 模型的扩展性。

影响算法选择的一个重要标准是模型的复杂性。一般来说，更复杂的模型：

- 需要更多的特征来学习和预测（如使用 10 个特征来预测一个目标）；
- 需要更复杂的特征工程（如使用多项式项、交互关系或主成分）；
- 需要更大的计算开销（如由 100 棵决策树组成的随机森林）。

此外，同一种机器学习算法会因为参数的数量或者对某些超参数的选择而变得更加复杂。例如：

- 一个回归模型可能拥有更多的特征或者多项式项和交互项；
- 一棵决策树可能拥有更大或更小的深度。

1.3　常用的机器学习算法

常用的机器学习算法主要包括以下几大类。

1. 线性回归

线性回归可能是最简单的机器学习算法。当想计算某个连续值时，可以使用回归算法，而分类算法的输出数据是类。所以，每当要预测一个正在进行的过程的某个未来值时，可以使用回归算法。但线性回归在特征冗余（也就是存在多重共线性）的情况下会不稳定。

线性回归的几个典型应用实例：预测特定产品在下个月的销量；预测血液酒精含量对身体协调性的影响；预测每月礼品卡销量和改善每年收入预期。

2. 逻辑回归

逻辑回归进行二元分类，所以输出数据是二元的。这种算法把非线性函数（Sigmoid）应用于特征的线性组合，所以它是一个非常小的神经网络实例。

逻辑回归提供了模型正则化的很多方法，与决策树和支持向量机相比，逻辑回归提供了出色的概率解释，能轻易地用新数据来更新模型。如果想建立一个概率框架，或者希望以后将更多的训练数据迅速整合到模型中，可以使用逻辑回归。

下面为逻辑回归的几个应用实例：预测客户流失；信用评分和欺诈检测；衡量营销活动的效果。

3. 决策树

人们很少使用单一的决策树，但与其他很多决策树结合起来，就能变成非常有效的算法，如随机森林和梯度提升树。

决策树的优点是：可以轻松地处理特征的交互关系，并且是非参数化的。其缺点是：不支持在线学习，所以在新样本到来时，必须重建决策树；容易过拟合，但随机森林和梯度提升树等集成方法可以克服这一缺点；占用很多内存（特征越多，决策树就可能越深、越大）。

决策树是帮助人们几个行动方案之间作出选择的出色工具，主要应用有：投资决策；客户流失；银行贷款违约人；自建与购买决策；销售线索资质。

4. *K*-Means（*K*- 均值）

聚类任务是并不知道任何标签，但目标是根据对象的特征赋予标签。聚类算法的一个实例是根据某些共同属性，将一大群用户分组。

如果在问题陈述中，存在“这是如何组织的”等疑问，或者要求将某物分组或聚焦特定的组，那么应该采用聚类算法。

K-Means 的最大缺点在于必须事先知道数据中将有多少个簇。因此，这可能需要进行很多的尝试，来“猜测”簇的最佳 *K* 值。

5. 主成分分析

主成分分析（Principal Component Analysis，PCA）能降维。有时，数据的特征很广泛，可能彼此高度相关，在数据量大的情况下，模型容易过拟合，这时可以使用 PCA。PCA 大受欢迎的一个关键在于除了样本的低维表示以外，它还提供了变量的同步低维表示。同步的样本和变量表示提供了以可视方式寻找一组样本的特征变量。

6. 支持向量机

支持向量机（Support Vector Machine，SVM）是一种监督学习方法，广泛用于模式识别和分类问题（前提是数据只有两类）。

SVM 的优点是精度高，对避免过拟合有很好的理论保障，而且只要有了适当的核函数，哪怕数据在基本特征空间中不是线性可分的，SVM 也能运行良好。在解决高维空间是常态的文本分类问题时，SVM 特别受欢迎。SVM 的缺点是消耗大量内存、难以解释和不易调参。

SVM 在现实中的几个应用：探测常见疾病（如糖尿病）患者；手写文字识别；文本分类——按话题划分新闻报道；股价预测。

7. 朴素贝叶斯

朴素贝叶斯是一种基于贝叶斯定理的分类方法，构建容易，对大数据集特别有用。该方法相对简单，分类效果却比某些高度复杂的分类方法要好。在 CPU 和内存资源是限制因素的情况下，朴素贝叶斯是很好的选择。

朴素贝叶斯算法十分简单，只需要做些算术运算即可。如果朴素贝叶斯关于条件独立的假设确实成立，那么朴素贝叶斯分类器将比逻辑回归等判别模型更快地收敛，因此需要的训练数据更少。即使假设不成立，朴素贝叶斯分类器在实践中仍然常常表现较好。如果需要的是快速简单且表现出色的方法，朴素贝叶斯将是不错的选择。其主要缺点是学习不了特征间的交互关系。

朴素贝叶斯在现实中的几个应用：情感分析和文本分类；推荐系统，如 Netflix 和亚马逊；把电子邮件标记为垃圾邮件或者非垃圾邮件；面部识别。

8. 随机森林算法

随机森林算法包含多棵决策树，它能解决拥有大数据集的回归和分类问题，还有助于从众多的输入变量中识别最重要的变量。但随机森林算法的学习速度可能很慢（取决于参数化），而且不可能迭代地改进生成模型。随机森林算法可用于预测制造业的零件故障、预测贷款违约人等。

9. 神经网络

神经网络包含神经元之间的连接权重。权重是平衡的，在学习数据点后继续学习数据点。所有权重被训练后，神经网络可以用来预测类或者量，如果发生了一个新的输入数据点的回归，用神经网络可以训练极其复杂的模型，再加上“深度方法”，即便是更加不可预测的模型也能被用来实现新的可能性。例如，利用深度神经网络，对象识别近期取得巨大进步。神经网络可以应用于非监督学习任务，如特征提取，深度学习还能从原始图像或语音中提取特征，不需要太多的人类干预。

但是，神经网络非常难以解释说明，参数化极其令人头疼，而且非常耗费资源和内存。

1.4 机器学习的应用领域

机器学习作为一种强大的人工智能技术，已经在多个领域得到广泛应用。以下为机器学习的几个主要应用领域。

1. 数据分析

机器学习在数据分析中的应用非常广泛，数据分析是一个三重过程，涉及从不同来源收集数据、以全面的方式呈现数据（即可视化）及应用机器学习算法来确保过程的效率和准确性。

2. 金融风险管理

机器学习在金融领域的应用越来越重要，特别是在风险管理方面。此外，机器学习和预测分析还被用于股票市场预测、市场研究和欺诈预防。由于网络欺诈活动大多是自动算法帮助完成的，机器学习和预测分析赋予了洞察欺诈活动的能力，并可以提供完整的犯罪情况。

3. 预测与推荐系统

机器学习在预测和推荐系统中也有广泛的应用，机器学习可以根据用户的偏好、意图和行为来做出相关内容的假设，实现服务个性化，从而提高用户对服务的参与度和整个用户体验的有效性和充实感。

4. 医疗健康

机器学习在医疗领域有着广泛的应用，包括医疗图像分析、疾病预测、药物发现等。例如，机器学习可以帮助识别无症状的左心室功能障碍，这是心力衰竭的前兆。

5. 自动驾驶

机器学习可以应用于自动驾驶技术，如图像识别和预测等。通过机器学习，自动驾驶车辆可以实现对周围环境的准确识别和理解，从而实现安全、高效的自动驾驶。

6. 工业生产

机器学习在工业生产中可以用于设备故障预测、优化生产流程、质量控制等。通过实时分析传感器数据，机器学习可以帮助企业实现设备预测性维护，提高生产效率并降低成本。

7. 决策支持与智能分析

机器学习在决策支持系统中可以帮助分析大量数据，辅助决策制定。基于数据的决策可以更加准确和有据可依。此外，机器学习还可以用于图像识别与计算机视觉，使计算机能够理解和解释图像，这在许多领域非常有价值。

以上只是机器学习部分应用领域的一小部分例子，随着技术的不断进步，机器学习在多个领域的应用将继续扩展和深化。通过机器学习，人们可以更好地理解和预测未来的趋势，为社会创造更大的效益。