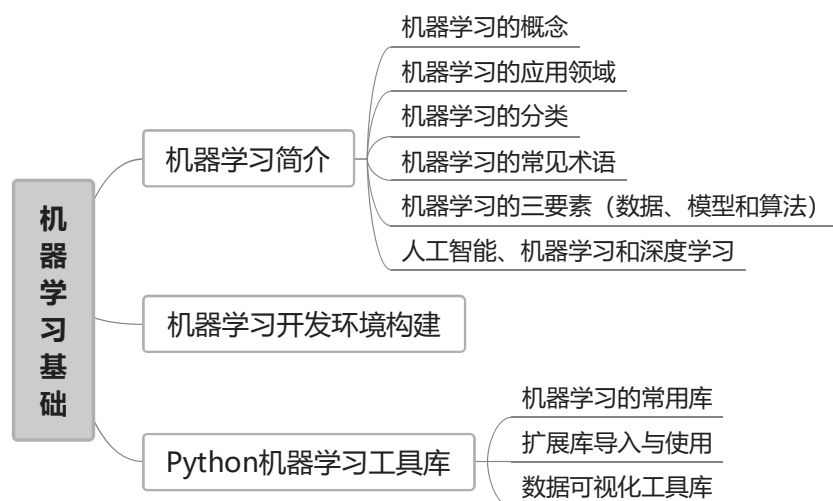


# 第 1 章 机器学习基础



## 本章导读

机器学习是当今科技领域备受瞩目的一个分支，它的广泛应用涉及各领域，从自然语言处理到图像识别，从医疗诊断到金融风险管理。本章将带领读者深入了解机器学习的基本概念、应用领域、分类和常见术语，并讨论人工智能、机器学习和深度学习之间的关系。

## 本章要点

- 📍 机器学习概述。
- 📍 机器学习的应用领域。
- 📍 机器学习的分类。
- 📍 机器学习的三要素。
- 📍 人工智能、机器学习和深度学习。
- 📍 环境构建和 Python 工具库。

## 1.1 机器学习简介

机器学习是人工智能的分支，它通过数据学习模式和规律使计算机系统自主作出决策或预测，广泛应用于自然语言处理、图像识别、医疗、金融等领域，通过不断适应新数据提高性能，成为解决复杂问题和提高效率的强大工具。



机器学习的概念  
和应用领域

### 1.1.1 机器学习的概念

对机器学习的主流定义有下述三种。

定义1：机器学习是人工智能的一个研究方向，其主要研究对象是人工智能算法，研究重点是在经验中学习提高具体算法的性能。

定义2：机器学习研究能通过经验自动提高自身的算法。

定义3：机器学习是数据或以往的经验作为提高算法性能标准的过程。

对机器学习更具体的解释是，计算机程序通过学习将无序的数据转化为有用的信息，使程序自行解决实际问题。该学习过程通常不需要人类对计算机程序下达指令，由程序独立完成，其关键是建立一个正确的模型，建模过程就是机器的“学习”。

### 1.1.2 机器学习的应用领域

机器学习在不同领域发挥作用，能够处理与分析不同数据，包括文本数据、语音数据、图像数据和视频数据。机器学习的应用领域涵盖文本分类、语音识别、图像分类和视频识别等。下面是常见应用举例。

#### 1. 文本数据

文本数据也称字符串数据，如英文字母、汉字、不作为数值使用的数字和其他可输入的字符。超文本是文本数据的另一种形式，包含标题、作者、超链接、摘要和内容等信息。文本数据的应用场景包含垃圾邮件检测、信用卡欺诈检测和电子商务决策等。

(1) 垃圾邮件检测。根据邮箱中的邮件识别垃圾邮件和非垃圾邮件，可以实现邮件的归类。

(2) 信用卡欺诈检测。根据用户一个月内的信用卡交易识别符合该用户操作习惯和不符合操作习惯的交易，可以发现欺诈交易。

(3) 电子商务决策。根据用户的购物记录、搜索记录和冗长的收藏清单识别其真正感兴趣并愿意购买的产品，可以为用户提供建议，鼓励用户消费。

#### 2. 语音数据

语音数据是指通过语音记录的数据或通过云传输的声音信息，也称声音文件。语音数据的应用场景包含语音识别、语音合成、语音交互、机器翻译、声纹识别等。

(1) 语音识别。语音识别是指机器通过识别和理解过程把语音信号转换为相应的文本或命令的技术，例如从一个用户的话语中确定其提出的具体要求，可以自动填充用户需求或者理解说话人要表述的意思或情感。

(2) 语音合成。语音合成是指通过机械的、电子的方法产生人造语音技术，例如将外部输入的文字信息转换为可以听懂的流利的汉语口语输出。

(3) 语音交互。语音交互是指基于语音输入的新一代交互模式，通过说话得到反馈结果，典型应用场景是语音助手（如百度公司推出的度秘）。

(4) 机器翻译。机器翻译又称自动翻译，是利用计算机将一种自然语言（源语言）转换为另一种自然语言（目标语言）的过程，如有道词典等翻译软件。

(5) 声纹识别。声纹识别是指把声信号转换成电信号，再用计算机识别，也称说话人模式。声纹识别可用于说话人辨认和说话人确认。例如，缩小刑侦范围时需要辨认技术，银行交易时需要确认技术。

### 3. 图像数据

图像识别是机器学习领域的核心研究方向，它的应用场景包括文字识别、指纹识别、人脸识别、形状识别等。

(1) 文字识别。文字识别是利用计算机自动识别字符的技术，是模式识别应用的一个重要领域，一般包含文字信息的采集、信息的分析与处理、信息的分类判别等部分。

(2) 指纹识别。指纹识别的原理是通过比较不同指纹的细节特征点来识别，涉及图像处理模式识别、计算机视觉、数学形态学、小波分析等学科。

(3) 人脸识别。人脸识别是基于人的脸部特征信息进行身份识别的一种生物识别技术，是指用摄像机或摄像头采集含有人脸的图像，并自动在图像中检测和跟踪人脸，进而对检测到的人脸进行识别的一系列相关技术。例如手机的人脸解锁功能，通过面部识别解锁手机。

(4) 形状识别。形状识别是模式识别的重要方向，广泛应用于图像分析、机器视觉和目标识别等领域。例如，医学领域中的病变形状识别，通过分析 X 射线或磁共振成像 (Magnetic Resonance Imaging, MRI) 图像中的形状特征帮助医生诊断疾病。

### 4. 视频数据

视频可以看作特定场景下的连续图像，视频比图像数据的维度高、信息量多、处理难度大。视频的应用场景包含智能监控和计算机视觉等领域。

(1) 智能监控。智能监控用于将视频转换成图像并处理，首先提取视频中的运动物体，然后跟踪提取的运动物体。其中涉及监控视频的去模糊、去雾、夜视增强、视频浓缩等步骤。

(2) 计算机视觉。计算机视觉是利用摄像机和计算机模仿人类视觉，实现对目标的分割、分类识别、检测、跟踪、判别、决策等功能的人工智能技术。它的研究目标是使计算机具有通过二维图像认知三维环境的能力，在基本图像处理的基础上进一步进行图像识别、图像 (视频) 理解和场景重构。

## 1.1.3 机器学习的分类

机器学习的分类方式有很多种，常见的有按任务类型分类和按学习方式分类。

### 1. 按任务类型分类

(1) 分类问题。分类是机器学习中的常见任务，它涉及将数据样本分配到已知类别，常见应用有垃圾邮件过滤、图像识别、手写体识别等。

(2) 回归问题。回归任务是根据已知的特征和属性预测一个连续变量的值，常见应用有房价预测、股票价格预测、销售量预测等。

(3) 聚类问题。聚类任务涉及将数据样本分为不同的组，使得同一组内的数据点相似度高，不同组之间的相似度低，常见应用有市场细分、用户分群、图像分割等。

(4) 异常检测。异常检测是识别和检测数据中的异常模式或离群值，常见应用有信用卡欺诈检测、网络入侵检测、设备故障检测等。

(5) 降维问题。降维是指采用某种映射方法将原高维空间中的数据点映射到低维空间。为什么要降维呢？可能是原始高维空间中包含冗余信息或噪声，需要通过降维将其消除；也可能是某些数据集特征维度过大，训练过程比较困难，需要通过降维来减少特征量。在降维中应用的模型有主成分分析 (Principal Component Analysis, PCA)、线性判别分



机器学习的分类

析 (Linear Discriminant Analysis, LDA) 等。

## 2. 按学习方式分类

(1) 有监督学习 (Supervised Learning)。有监督学习, 简称监督学习, 是指基于一组带有结果标注的样本训练模型对新的未知结果的样本作出预测。通俗地讲就是利用训练数据学习得到一个将输入映射到输出的关系映射函数, 然后将关系映射函数使用在新实例上, 得到新实例的预测结果。例如, 某商品以往的销售数据可以用来训练商品的销售模型。该模型可以用来预测该商品的销售走势。常见的监督学习任务有分类和回归。

- 分类: 当模型被用于预测样本所属类别时就是一个分类问题, 例如区别某张给定图片中的对象是猫还是狗。
- 回归: 当所有预测的样本结果为连续数据时就是一个回归问题, 例如预测某股票未来一周的市场价格。

(2) 无监督学习 (Unsupervised Learning)。无监督学习是指训练样本的结果信息是没有被标注的, 训练集的结果标签是未知的, 我们的目标是通过学习这些没有标记的训练样本揭示数据的内在规律, 发现隐藏在数据之下的内在模式, 为进一步的数据处理提供基础。无监督在学习任务中比较常用的有聚类和降维。

- 聚类: 用于将数据集中的对象划分成具有相似特征的组 (簇), 以便将相似的对象分配到相同的簇中。聚类算法的目标是在不需要事先标记数据的情况下发现数据集中的隐藏结构和模式。例如在市场研究中, 聚类分析可以用来细分消费者, 将具有类似购买偏好和行为模式的消费者分配到同一簇中。
- 降维: 通过减少数据的特征维度, 从而在保持数据关键信息的同时减少数据的复杂性和存储空间。在机器学习和数据分析中, 降维可以用于处理高维数据, 从而降低问题的复杂度和提高算法的效率。例如通过将原始数据映射到一个具有较低维度的空间, 可以保留数据的关键特征, 同时去除冗余信息。

(3) 半监督学习 (Semi-supervised Learning)。半监督学习是介于无监督学习和有监督学习之间的一种学习方式, 它利用有标签和无标签的数据训练及预测。在半监督学习中, 通常只有小部分数据是带有标签的, 而大部分数据是没有标签的。半监督学习的目标是利用无标签数据的信息帮助提高模型的性能和泛化能力, 常用于图像分类和异常检测。

- 图像分类: 在图像分类任务中, 通过半监督学习可以使用有标签图像和大量无标签图像来训练分类模型。通过有标签图像, 模型可以学习类别特定的知识; 通过无标签图像, 模型可以使用无监督学习算法来学习数据中的结构和模式, 从而提高分类模型的性能。
- 异常检测: 在异常检测任务中, 半监督学习可以利用有标签的正常样本和大量无标签样本训练异常检测模型。通过有标签的正常样本模型可以学习到正常样本的分布特征; 通过无标签样本模型可以使用无监督学习算法来学习数据的整体分布, 从而更好地发现异常样本。

(4) 强化学习 (Reinforcement Learning)。强化学习旨在通过与环境的交互学习如何作出最优决策, 以最大化预期的累积奖励。在强化学习中, 有一个智能体 (Agent) 通过与环境进行连续的交互来学习, 智能体可观察环境的状态并采取行动来影响环境。在每个时间步, 智能体都会收到一个奖励信号, 表示其行为的优劣。强化学习的目标是通过学习找到最优策略, 使得智能体在长期中获得最大的累积奖励。强化学习常用于游戏和机器人

控制、机器人路径规划等方面。

- **游戏和机器人控制**：强化学习在游戏和机器人控制中有广泛应用。例如，可以使用强化学习算法训练智能体在不同游戏中自动参与游戏，如围棋、象棋、扑克等。此外，强化学习还可以用于训练机器人进行复杂任务，如自主导航、抓取和操纵物体等。
- **机器人路径规划**：强化学习可以用于训练机器人在复杂环境中规划最优路径。通过与环境交互，机器人可以学习避开障碍物、选择最优路径以及完成任务的方法，如清洁、仓储等。

(5) **迁移学习 (Transfer Learning)**。迁移学习用于将学习到的知识迁移到新的领域，以提高学习效率和模型性能。在传统的机器学习中，通常假设训练集和测试集是从同一个分布中采样得到的。然而，在现实世界中，不同的任务和领域之间往往存在一定的差异。迁移学习的目标是通过利用源域（已经存在或已经学习的任务和领域）的知识改善目标域（目标任务和领域）的学习效果，尤其是在目标域数据较少或标注困难的情况下。迁移学习在各领域均有应用，在图像识别和机器翻译领域尤为突出。

- **图像识别**：在图像识别任务中，迁移学习可以将在大规模图像数据集上训练好的卷积神经网络的知识迁移到其他图像分类任务上。迁移学习可以减少在目标任务上的训练时间和数据需求，并提高分类性能。
- **机器翻译**：在机器翻译任务中，迁移学习可以使用在大规模平行语料库上预训练好的翻译模型，将其知识迁移到特定领域或低资源语言对的翻译任务中，从而提高翻译质量和效率。

#### 1.1.4 机器学习的常见术语

机器学习是一种人工智能技术，涉及多种基本术语，见表 1-1。

表 1-1 机器学习的基本术语

术语	定义	数学描述	示例
数据集	数据样本的集合	$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ ，其中 $n$ 为样本量	500 个哈尔滨市房屋的面积、楼层、位置、朝向以及部分房价信息的数据集
样本	数据中的一条具体记录	$(x_i, y_i), i \in n$	一个房屋的数据
特征	用于描述数据的输入变量	$x_i = \{t_1, t_2, \dots, t_m\}$ ，也是一个向量	面积 ( $t_1$ )、楼层 ( $t_2$ )、位置 ( $t_3$ )、朝向 ( $t_4$ )
标签	要预测的真实事物或结果，也称目标	$\{y_1, y_2, \dots, y_n\}$	房价
有标签样本	有特征标签，用于训练模型	$(x_i, y_i), i \in n$	200 个哈尔滨市房屋的面积、楼层、位置、朝向以及房价信息
无标签样本	有特征，无标签	$(x_j, y_j), j \in n$	100 个哈尔滨市房屋的面积、楼层、位置、朝向，但无房价信息
模型	能够将样本映射到预测标签	$f(x)$ 函数	能够通过面积、楼层、位置、朝向等特征确定房价的函数



机器学习的常见术语  
和三要素

续表

术语	定义	数学描述	示例
模型中的参数	模型中的参数确定了机器学习的具体模型	$f(x)$ 函数的参数	如 $f(x) = 5x + 2$ ，其中 5 和 2 是该模型的参数
模型的映射结果	通过模型映射出无标签样本的标签	$\{y'_1, y'_2, \dots, y'_n\}$	100 个被预测出来的房价
机器学习	通过学习样本数据发现规律得到模型参数，从而得到预测的目标模型	确定 $F(x)$ 及其参数的过程	确定房价预测函数和具体参数过程

下面介绍最重要的三个术语：特征、标签和模型。

### 1. 特征

特征是基于学习的输入，原始的特征描述数据的属性。它是有维度的，特征的维度是指特征的数目（不是数据里面样本的数目）。不同数据中的数据特征的维度不同。

- 少：可以少到只有一个特征，也就是一维特征数据，比如房价标签仅取决于面积特征。
- 多：可以多到几万、几十万，比如一个 100 像素 × 100 像素的 RGB 彩色图像输入，每个像素都可以视为一个特征，也就是 1 万维再乘以 RGB 三个颜色通道，那么该图像数据的特征维度可以达到 3 万维。

### 2. 标签

标签是机器学习要输出的结果，也是试图预测的目标。表 1-1 中的标签是房价，实际上机器学习要解决什么问题呢？标签是什么？比如未来的股票价格、图片中的内容（猫和狗）、文本翻译的结果、音频输入的内容等。

下面是一个有标签的数据格式：

$$(x_1, x_2, x_3 : y)$$

标签有时是随样本一起的，有时是机器推断出来的，称为预测标签  $y'$ 。比较  $y$  和  $y'$  的差异就是在评价机器学习模型的效果。不是所有的样本都有标签，在无监督学习中，所有样本都没有标签。

### 3. 模型

模型是将样本映射到预测标签  $y'$ ，其实模型就是函数，是预测的工具。函数由模型的内部参数定义，而这些内部参数通过从数据中学习规律得到。

在机器学习中先确定模型的类型（如线性回归模型、逻辑回归模型、神经网络模型等），也可以说是算法，再确定模型的参数。如果选择线性回归模型，那么  $f(x) = 5x + 2$  中的 5 和 2 就是它的参数，而神经网络有神经网络的参数。模型和参数都确定后，机器学习的模型也就确定了。

#### 1.1.5 机器学习的三要素（数据、模型和算法）

人工智能的本质是一个函数（数学模型），我们给机器提供自己目前已有的数据，机器从这些数据中找一个最能拟合这些数据的函数，当需要预测新的数据时，机器可以通过这个函数预测新数据及对应的结果。