监督学习

模型从带标签的训练数据中学习。标签就是每个数据点对应的“正确答案”或“期望输出”。

像有一个老师（标签）在指导学生学习（模型训练）。老师提供问题和对应的标准答案，学生通过练习这些题来学习如何回答新问题

数据形式

输入数据：特征。例如，预测房价时的房屋面积、卧室数量、地理位置等。

输出数据：标签。例如，该房屋对应的实际价格（回归问题），或者该图片是“猫”还是“狗”（分类问题）

目标：学习一个从输入特征到输出标签的映射函数。模型的目标是学习这种关系，以便对新的、未见过的数据（只有特征）做出准确的预测（预测其标签）。

主要任务类型：

分类：预测离散的类别标签。

例子：垃圾邮件检测（垃圾邮件/正常邮件）、图像识别（猫/狗/汽车）、疾病诊断（患病/健康）。

回归： 预测连续的数值。

例子：房价预测、股票价格预测、销售额预测

无监督学习

2. 非监督学习

模型从无标签的数据中学习。数据只有输入特征，没有对应的“正确答案”。

就像让学生自己去探索一堆资料，发现其中的模式、结构或分组，没有老师提供标准答案。

数据形式

输入数据：特征。例如，描述客户行为的数据（购买频率、金额、浏览商品类别等）。

没有输出标签！

目标： 探索数据内部的固有结构、模式或关系。主要目标是：

发现隐藏的分组： 将相似的数据点聚集在一起。

降维： 减少数据特征的数量，同时保留重要信息，以便可视化或简化后续处理。

发现关联规则： 找出数据项之间经常同时出现的模式（常用于市场篮子分析）。

异常检测： 识别与大多数数据显著不同的点。

主要任务类型：

聚类： 将数据点分组，使得组内相似度高，组间相似度低。

例子： 客户细分、文档主题分组、基因序列分析。

降维： 将高维数据压缩到低维空间，同时保留尽可能多的信息。

例子： 数据可视化（将多维数据降到2D/3D以便绘图）、去除噪声和冗余特征。

关联规则学习： 发现数据中不同特征项之间同时出现的概率很高的规则。

例子： “购买尿布的顾客也经常购买啤酒” (经典的“啤酒与尿布”故事)。

异常检测： 识别与数据集主体显著偏离的罕见事件或观测值。

例子： 信用卡欺诈检测、网络入侵