

# 人工智能

## 1 常见机器学习算法：有监督和无监督

- **监督学习**：机器在**标记数据**上训练，模型学习输入到输出的**映射**。用于**分类、回归、目标检测**等。
  - **分类**：逻辑回归，支持向量机（SVM），决策树，随机森林，卷积神经网络（CNN）
  - **回归**：线性回归，神经网络回归，xgboost（二者均可），LightGBM（二者均可）
- **无监督学习**：机器在**未标记的数据**上训练，模型**挖掘数据**的潜在结构。用于**聚类、降维、异常检测**等。
  - **聚类**：k-means，高斯混合模型（GMM）
  - **降维**：主成分分析（PCA），t-SNE，自编码器

## 2 回归和分类

回归算法是一种**有监督学习算法**，用来建立自变量X和观测变量Y之间的映射关系。

- 若观测变量是离散的，则称其为分类；若观测变量是连续的，则称其为回归。

## 3 k-means

1. 选择初始化的k个样本作为**初始聚类中心**
2. 对数据集中的每个样本，计算它到k个聚类中心的距离，**把它分到中心距离最小的类别**
3. 针对每个类别，重新计算它的聚类中心（即求该类所有样本的质心）
4. 重复2和3，直到终止条件（迭代次数、最小误差变化等）

优点：复杂度低，效果较好，易理解

缺点：容易陷入局部最优，k值需要人为设定，对初始聚类中心敏感

## 4 PCA（主成分分析）

PCA是一种**无监督学习算法**，常用于数据降维，通过线性投影将高维数据映射到低维空间，以用较少的数据维度，同时保留较多的原数据点的特性。还有随机森林、低方差滤波等方法用于数据降维。

对于m维样本集，低维空间d维

1. 先对所有样本去中心化（减去平均值），关于0对称
2. 计算样本的协方差矩阵  $XX^T$
3. 对协方差矩阵做特征值分解，提取出最大的d个特征值对应的特征向量。这些特征向量组成投影矩阵即输出

## 5 决策树

决策树是一棵由多个判断节点组成的树。使用模型预测时，根据输入参数在各个判断节点进行判断，直到叶节点为预测结果。需要剪枝以防止过拟合