# 1. 绪论

- **人工智能、机器学习、深度学习的关系**
    - 人工智能（AI）：让机器表现出类似人类智能的能力，例如理解语言、识别图像、做出决策。
    - 机器学习（ML）：AI的一个分支，教计算机通过数据和经验自动改进。
    - 深度学习（DL）：ML的一个子集，使用多层神经网络来学习复杂模式。
- **机器学习的应用场景及无法解决的问题**
    - 应用场景：推荐系统（如电影推荐）、自动驾驶、语音识别（如Siri）、图像识别（如人脸识别）。
    - 无法解决的问题：道德判断、创造性思维、需要广泛知识的复杂决策。
- **机器学习的基本步骤**
    - 数据搜集：收集大量相关的数据。
    - 数据清洗：处理数据中的缺失值和错误数据。
    - 特征工程：从数据中提取重要的特征。
    - 数据建模：选择合适的算法并训练模型。
    - 模型评估：使用测试数据评估模型的性能。
    - 模型部署：将模型应用到实际环境中。

# 2. 机器学习的类型

- **监督学习**
    - 分类：将输入数据分到预定义的类别中。例如垃圾邮件分类（垃圾邮件或非垃圾邮件）。
    - 回归：预测连续值。例如预测房价。
- **无监督学习**
    - 聚类：将数据分成若干组。例如客户分群，根据购物习惯将客户分组。
    - 降维：降低数据的维度，使其更易于分析。例如主成分分析（PCA）。

# 3. 线性回归

- **概念及应用场景**
    - 线性回归：通过线性方程建模变量之间的关系，常用于预测。例如根据房子的面积预测价格。
- **最小二乘法**
    - 最小二乘法：找到最小化预测值和实际值之间平方误差的线性方程。
- **梯度下降法**
    - 批量梯度下降：一次使用整个数据集更新模型参数。
    - 随机梯度下降：每次使用一个样本更新模型参数。
    - 小批量梯度下降：每次使用一部分样本更新模型参数。
- **数据归一化/标准化的重要性和方法**
    - 归一化：将数据缩放到特定范围（如0到1），使各特征在相同尺度上。
    - 标准化：将数据调整为均值为0、标准差为1的分布，消除特征值的量纲影响。
- **过拟合和欠拟合的定义与解决方法**

    - 过拟合：模型在训练集上表现好，但在测试集上表现差。解决方法：正则化、减少模型复杂度、增加训练数据。
    - 欠拟合：模型在训练集和测试集上都表现差。解决方法：增加模型复杂度、选择更合适的特征。
- **回归的评价指标**
    - MSE（均方误差）：预测误差的平方和的平均值。
    - RMSE（均方根误差）：MSE的平方根。
    - MAE（平均绝对误差）：误差的绝对值的平均值。

## 4. 模型评估与选择

- **泛化误差与经验误差**
    - 泛化误差：模型在新数据上的误差。
    - 经验误差：模型在训练数据上的误差。
- **评估方法**
    - 交叉验证：将数据集分成多个子集，多次训练和验证以评估模型性能。
- **性能度量**
    - 准确率：预测正确的比例。
    - 精确率：预测为正样本中实际为正的比例。
    - 召回率：实际正样本中预测为正的比例。
    - F1分数：精确率和召回率的调和平均数。
    - AUC-ROC：衡量分类器性能的曲线下面积。

## 5. 对数几率回归

- **概念及应用场景**
    - 对数几率回归：用于二分类问题，通过对数几率函数将线性回归的输出映射到0和1之间的概率值。例如二分类的信用卡欺诈检测。

## 6. 决策树

- **原理及特点**
    - 决策树：使用树状结构进行决策，通过属性分割数据。每个节点表示一个属性，每个分支表示一个属性的值，每个叶子节点表示一个类别。
    - 优点：简单易懂、处理非线性数据。
    - 缺点：容易过拟合、对数据噪声敏感。
- **经典模型**
    - ID3算法：使用信息增益选择分裂属性。
    - C4.5算法：改进ID3，使用信息增益比，处理连续属性和缺失值。
    - CART算法：使用基尼系数或均方误差选择分裂属性，生成二叉树。

## 7. 神经网络

- **基本概念及应用场景**
    - 神经网络：模仿生物神经网络，通过神经元和连接进行数据处理。
    - 应用场景：图像识别、语音识别、自然语言处理等。
- **常用结构和训练方法**
    - 常见结构：前馈神经网络、卷积神经网络（CNN）、循环神经网络（RNN）。
    - 训练方法：反向传播、梯度下降等。

## 8. 支持向量机

- **基本概念及应用场景**
    - 支持向量机：通过寻找最大间隔的超平面将数据分类。
    - 应用场景：文本分类、图像分类等。
- **核函数的作用**
    - 核函数：将低维数据映射到高维空间，以便找到非线性分割。例如径向基函数（RBF）核、线性核、多项式核。

## 9. 贝叶斯分类器

- **贝叶斯决策论**
    - 贝叶斯决策论：根据贝叶斯定理进行分类决策，选择后验概率最大的类。
- **先验概率与后验概率**
    - 先验概率：分类前对类别的初始估计。
    - 后验概率：根据新数据更新后的概率。
- **贝叶斯定理及其应用**
    - 贝叶斯定理：

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

    - 应用：垃圾邮件过滤、医学诊断等。

## 10. 聚类

- **性能度量**
    - 外部指标：依赖于外部信息（如分类标签）。
    - 内部指标：仅依赖于数据本身（如轮廓系数、簇间距离）。
- **距离计算**
    - 连续属性：欧几里得距离、曼哈顿距离。
    - 离散属性：汉明距离、Jaccard距离。
    - 有序属性：Spearman距离。
    - 无序属性：使用二值化处理。
- **密度聚类**
    - 定义：根据数据点的密度进行聚类。
    - 经典算法：DBSCAN（基于密度的空间聚类算法）。

## 11. 降维

- **概述及维数灾难**
    - 降维：减少数据的维度，保留主要信息。
    - 维数灾难：高维数据导致的计算复杂性和过拟合问题。
- **常见降维技术**
    - 主成分分析（PCA）：通过线性变换减少维度，找出数据中最重要的方向。
    - 线性判别分析（LDA）：根据类别信息进行降维，最大化类间方差和最小化类内方差。
    - t-SNE：非线性降维方法，适用于数据可视化。

# 可能考点

**监督学习与无监督学习**

- **分类与回归**：理解分类和回归的基本概念和区别，掌握常见的分类和回归算法。
- **聚类与降维**：理解聚类和降维的基本概念和区别，掌握常见的聚类和降维算法。

**线性回归**

- **最小二乘法与梯度下降法的比较**：理解最小二乘法和梯度下降法的基本原理和应用场景，掌握两者的优缺点。
- **数据归一化和标准化的必要性**：理解数据归一化和标准化的基本概念和方法，掌握其在模型训练中的作用。
- **过拟合和欠拟合及其解决方法**：理解过拟合和欠拟合的基本概念和解决方法，掌握正则化、增加训练数据等常用方法。

**决策树**

- **ID3、C4.5、CART算法的特点和区别**：理解ID3、C4.5和CART算法的基本原理和应用场景，掌握三者的优缺点和区别。
- **决策树的优缺点**：理解决策树的基本原理和优缺点，掌握其在实际应用中的局限性。

**贝叶斯分类器**

- **先验概率和后验概率的理解**：理解先验概率和后验概率的基本概念和区别，掌握贝叶斯定理的应用场景。
- **贝叶斯定理的应用**：掌握贝叶斯定理的基本公式和应用场景，理解其在垃圾邮件过滤、医学诊断等领域的应用。

**支持向量机**

- **基本概念和核函数的作用**：理解支持向量机的基本原理和应用场景，掌握核函数的作用和常见类型。

**模型评估与选择**

- **交叉验证方法**：理解交叉验证的基本原理和应用场景，掌握常见的交叉验证方法（如k折交叉验证）。
- **性能度量指标**：掌握准确率、精确率、召回率、F1分数、AUC-ROC等常见性能度量指标的计算方法和应用场景。

**降维**

- **维数灾难及解决方法**：理解维数灾难的基本概念和解决方法，掌握常见的降维技术（如PCA、LDA、t-SNE）。

**聚类**

- **聚类的性能度量**：掌握外部指标和内部指标的基本概念和计算方法，理解其在聚类性能评估中的作用。
- **DBSCAN算法**：理解DBSCAN算法的基本原理和应用场景，掌握其优缺点和常见应用。