**决策树的基础**

**基本原理**：将预测变量空间划分为若干简单的区域，每个区域的预测值是区域内训练数据响应变量的均值（回归）或最常见类别（分类）。

**构建方式**：基于分割规则（如Xj < tk）递归地划分预测变量空间，生成类似倒置树的结构。

**回归树的构建流程**

**分割预测空间**：

将预测变量的值域分割为若干矩形区域（R1, R2,...）。

目标是最小化残差平方和（RSS）。

**递归二分法（Recursive Binary Splitting）**：

**贪心策略**：每次选择最佳分割点，使RSS最大化减少。

按需继续细分子区域，直至达到停止标准（如每个区域含有的观测值少于指定阈值）。

**预测**：测试数据点的预测值为其所属区域训练数据响应变量的均值。

**修剪决策树（Pruning）**

**原因**：防止过拟合。

**方法**：代价复杂度修剪（Cost Complexity Pruning）。

定义一个调节参数 α，控制子树复杂度与拟合程度之间的权衡。

使用交叉验证（Cross-validation）选择最佳 α 值。

**分类树**

**与回归树的不同点**：

目标是预测定性响应变量。

分类标准替代RSS：

**分类错误率（Classification Error Rate）**：



**Gini 指数**：



衡量节点纯度。

**交叉熵（Cross-Entropy）**：

A black text on a white background

Description automatically generated

类似于Gini指数。

通常用Gini指数或交叉熵来生成树，用分类错误率来修剪树。

**优点与不足**

**优点**：

易于解释和可视化。

无需大量数据预处理（如标准化）。

**缺点**：

容易过拟合（Overfitting），需要修剪。

对数据噪声敏感。

**决策树**

决策树（Decision Trees）对数据的变化敏感，因此具有低偏差（low-bias）但高方差（high-variance）。

树模型往往需要通过其他方法（例如 Bagging、Random Forest 和 Boosting）来降低方差。

**Bagging（Bootstrap Aggregating）**

**基本原理**：

利用自助法（Bootstrap）生成多个训练数据集，并对每个数据集训练单独的模型。

将这些模型的预测结果进行平均（回归问题）或通过投票机制得出最终分类（分类问题）。

**优点**：

显著降低模型的方差。

适用于回归和分类问题。

**评估方法**：

使用测试集计算误差（例如 MSE 或分类错误率）。

利用袋外数据（OOB Data）估计误差。

**随机森林（Random Forest）**

**与 Bagging 的区别**：

随机森林通过在每次分裂节点时仅使用一部分随机选择的特征来减少模型间的相关性，从而进一步降低方差。

**算法**：

从训练数据中抽取多个自助样本。

对每个节点，随机选择 m 个特征（通常 m≈pm ≈ \sqrt{p}，其中 p 是总特征数），基于这些特征进行分裂。

**优势**：

比 Bagging 模型有更好的性能。

提供内置的特征重要性评估。

**Boosting**

**基本原理**：

提高弱学习器（例如浅树模型）的性能。

模型以序列方式训练，每个模型试图修正前一个模型的错误。

**算法**：

初始残差设为真实值。

每棵树拟合当前的残差，并通过学习速率（Shrinkage Parameter, λ）调整模型。

最终模型是所有子模型的加权和。

**调参**：

树的数量 B：过多可能导致过拟合。

学习速率 λ：较小的 λ通常需要更多树。

每棵树的分裂深度 d：控制模型的复杂性，常用 d=1（即单分裂）。

**优点**：

高效处理复杂的分类和回归问题。

在正确选择参数的情况下，通常优于随机森林。

**关键比较**

**Bagging** 和 **Random Forest**：

Bagging 通过简单平均降低方差；Random Forest 增加了特征随机性，进一步减少模型相关性。

**Boosting** 与其他两种方法：

Boosting 更关注模型偏差的降低，强调模型的序列式优化。

与 Bagging 和 Random Forest 相比，

Boosting 可能更容易过拟合，但调整得当时效果更佳。

**Log Loss**

**A black and white symbol

Description automatically generated**

**A close up of black text

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A math equations on a white background

Description automatically generated**

****