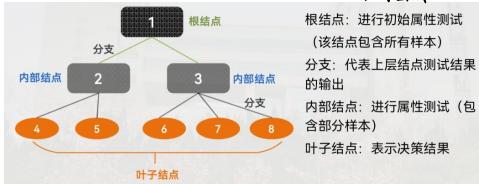


决策树：进行决策的树形结构



可解释性强、简单、高效

本质 \Rightarrow 从训练集中归纳出一组分类规则

算法实现

1. ID3 离散特征，多叉树

信息熵

关键：选择最优化划分属性 \Rightarrow 划分过程使得样本尽可能属同一类别 \Rightarrow 提高性能

使用信息熵度量纯度：集合D中第k类样本占比 p_k ，有信息熵 $Ent(D) = -\sum_{k=1}^{|D|} p_k \log_2 p_k$ ($p=0$ 时 $p \log_2 p = 0$)

信息增益：属性A对训练数据集D信息增益为集合D经验熵 $H(D)$ 与给定A条件的经验熵 $H(D|A)$ 之差

划分标准 $Gain(D, A) = H(D) - H(D|A)$

数据集D中有属性A上取值 A^i 的样本

$$H(D) = Ent(D) = -\sum_{k=1}^{|D|} p_k \log_2 p_k$$

$$H(D|A) = \frac{1}{|D|} \sum_{i=1}^{|D|} Ent(D^{(i)})$$

$$\sum H(D|A=A^i)$$

\rightarrow 按属性A分为类后，续子类内信息熵之和 (新信息熵)

核心思想：根据信息增益来选择划分属性

具体方法：

- 从根节点开始，计算所有属性的信息增益
- 选择信息增益最大的属性作为节点划分属性
- 由该属性的不同取值进行分支
- 重复以上步骤

优点：

- 假设空间包含所有的决策树，搜索空间完整
- 健壮性好，特征噪声影响较小
- 方法简单，理论清晰

缺点：

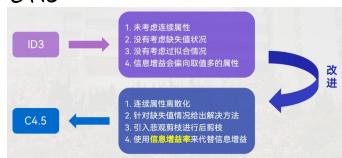
- 只考虑离散属性，没有考虑连续属性
- 对缺失值状况没有进行考虑
- 没有考虑过拟合情况
- 采用信息增益作为划分标准，但信息增益倾向于取值较多的属性。例如我们选择前面贷款的例子中的第一列ID3使用 A_5 来表示作为划分属性：

$$H(D|A_5 = i) = -(1 * \log_2 1 + 0 * \log_2 0) = 0 \quad (i = 1, 2, \dots, 5)$$

$$H(D|A_5) = \sum_{i=1}^5 \frac{1}{15} * H(D|A_5 = i) = 0.971 - 0 = 0.971$$

由此可以看出当选择ID3作为划分属性时，子结点的信息熵会直接降为0，此时的信息增益也远远大于其他属性。

2. C4.5



划分标准：信息增益率 $Gain_ratio(D, A) = \frac{Gain(D, A)}{I(A)}$

$$属性A的固有值 I(A) = -\sum_{i=1}^{|A|} p_i \log_2 p_i$$

* 属性A取值数目越多，I(A)越大，I(A)/值越大 \Rightarrow 倾向于取值少的属性
实际应用中采用启发式算法：先从候选特征中找到信息增益高于平均值的特征，再从中选择增益率高者

连续特征处理

- 连续离散化
- 将连续属性在数据集D上的取值按从小到大排列，记为 a_1, a_2, \dots, a_n
- 此步会产生 $n-1$ 个划分点（划分点取两侧值平均数）例如 $T_1 = \frac{a_1+a_2}{2}$ ，得到一个划分集合 $T_1 = \{x | a_1 \leq x \leq T_1\}$
- 将随着为选取取值来考察这些划分点（使用信息增益率进行划分）

缺失值处理 \Rightarrow 弃用浪费大量样本

属性值缺失情况下如何选择划分属性

$$取无缺失值样本子集 D' 来判断属性优劣 \quad Gain(D; A) = p \times Gain(D'; A)$$

给出划分属性，如何划分该属性缺失的样本

将样本以不同概率划分为所有属性中去属性A中取值比例分别为 p_1, p_2, \dots, p_n 将有缺失样本以权重 p_1, p_2, \dots, p_n 同时划分进三个分支

减轻过拟合剪枝策略 \Rightarrow 训练过程中为尽可能正确分类样本而导致划分过程不断重复，导致决策树分支过于拟合

剪枝策略：生成过程中剪枝

判断划分是否会使决策树泛化提升，若不能，就停止划分。

在已生成的决策树上，通过剪枝来减少不必要的分支，简化模型提高效率。

剪枝策略：生成过程中剪枝

后剪枝：生成一个完整决策树后再剪枝

训练时间开销小，有拟合风险

无缺失值样本占比

3. CART \Rightarrow 二分类归分割 \Rightarrow 二叉树（每属性均分割为两部分）

划分标准：基尼系数 $Gini(D) = \sum_{k=1}^{|D|} p_k \cdot p_k = 1 - \sum_{k=1}^{|D|} p_k^2$ 无对称运算，直观反映从数据集中随机抽取两个样本，其类别分布还不一致的概率
值越小，数据集纯度越高

$Gini_index(D, A) = \sum_{i=1}^{|D|} \frac{|D^{(i)}|}{|D|} Gini(D^{(i)})$ 选划分后基尼指数最小属性作最优划分属性 $A_i = \arg \min_{A_i \in A} Gini_index(D, A_i)$
对多分类指针：对特征进行二分 {青，中，老} \rightarrow {青，中，青老}，以 Gini 指数最小者为划分方式参与其它指标基尼指数的比较之中
 \downarrow 青，中，青老