ID3

使用熵增益计算特征的价值（不纯度指标）

ID3倾向于选择标签多的列

连续数据的离散化

C4.5

使用熵增益率计算特征价值

现在我们假设将训练元组D按属性A进行划分，则A对D划分的期望信息为：

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} info(D_j)$$

而信息增益即为两者的差值：

$$gain(A) = info(D) - info_A(D)$$

$$\text{增益率} = \frac{info_A(D)}{gain(A)}$$

增益率越大越好

CART(分类和回归树)

gini指标

$$gini(play) = 1 - \sum_{i=1}^{n} p_i^2$$

$$p_1 = \frac{9}{14}$$
$$p_2 = \frac{5}{14}$$
$$1 - p_1^2 - p_2^2$$

$$sunny\text{的条件}gini\text{值}$$

$$p_{sunny-yes} = \frac{2}{5}$$

$$p_{sunny_no} = \frac{3}{5}$$

$$p_{sunny-result} = \frac{5}{14}(1 - p^2_{sunny-yes} - p^2_{sunny-no})$$

gini指标越小越好

CART 分类树的生长：

- **核心思想：** 每次选择能够使数据集基尼指数下降最多的特征及其划分点进行划分。
- **基尼指数：** 对于包含 K 个类别的样本集 D，假设某个节点中属于类别 k 的样本占比为 pk，则该节点的基尼指数 Gini(p) 定义为： $Gini(p) = \sum p_k = \sum p_k(1 - p_k) = 1 - \sum p_k^2$ 基尼指数越小，数据集的纯度越高。
- **划分过程：**

1. 遍历所有可能的特征 A 以及该特征的所有可能的划分点 s。
2. 对于每个特征 A 的每个划分点 s，将数据集 D 划分为两个子集 D1 和 D2。
3. 计算划分后的基尼指数： $Gini_index(D, A) = \frac{|D1|}{|D|}Gini(D1) + \frac{|D2|}{|D|}Gini(D2)$
4. 选择使得划分后基尼指数最小的特征和划分点作为最优划分。

$$\text{有太阳和没有太阳}$$

$$p_{\text{有太阳且出去玩}} = \frac{2}{5}$$

$$p_{\text{有太阳不出去玩}} = \frac{3}{5}$$

$$gini_{\text{有太阳}} = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2$$

$$p_{\text{没有太阳且出去玩}} = \frac{7}{9}$$

$$p_{\text{没有太阳不出去玩}} = \frac{2}{9}$$

$$gini_{\text{没有太阳}} = 1 - (\frac{7}{9})^2 - (\frac{2}{9})^2$$

$$gini(\text{出去玩, 有太阳}) = \frac{5}{14}gini_{\text{有太阳}} + \frac{9}{14}gini_{\text{没有太阳}}$$

$$\text{按照热不热进行划分}p_{\text{很热且出去玩}} = 0.5$$

$$p_{\text{很热且不出去玩}} = 0.5$$

$$gini_{\text{很热}} = 1 - 0.5^2 - 0.2^2$$

$$p_{\text{很热的补集中出去玩}} = 0.7$$

$$p_{\text{很热的补集不出去玩}} = 0.3$$

$$gini_{\text{很热的补集}} = 1 - 0.7^2 + 0.2^2$$

$$gini(\text{出去玩, 很热}) = \frac{4}{14}gini_{\text{很热}} + \frac{10}{14}gini_{\text{很热的补集}}$$

剪枝

预剪枝

后剪值