# 决策树

W.J.Z

**摘要：**决策树学习通常包括特征选择、决策树生成和决策树剪枝，其内部节点表示一个特征或属性，叶子节点表示一个类，学习损失函数通常为正则化的极大似然公式，决策树生成只考虑局部最优，剪枝则考虑全局最优。

## 1 ID3 and C4.5

在信息论与概率统计中，熵是表示随机变量不确定性的度量。

$$P\left(p\right)=-\sum_{i=1}^{n}p_i\log p_i \tag{1}$$

定义 $0\log 0=0$, 对数以 2 为底或以 e 为底。信息增益公式为：

$$g\left(D,A\right)=H\left(D\right)-H\left(D|A\right) \tag{2}$$

数据集 $D$ 的经验熵为 $H(D)$ 为：

$$H\left(D\right)=-\sum_{k=1}^{K}\frac{|C_k|}{D}\log_2\frac{|C_k|}{D} \tag{3}$$

特征 $A$ 对数据集 $D$ 的经验条件熵为：

$$H\left(D|A\right)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H\left(D_i\right)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|} \tag{4}$$

信息增益为：

$$g(D,A)=H(D)-H(D|A) \tag{5}$$

数据集 $D$ 关于特征 A 的值的熵比为：

$$g\left(D,A\right)=\frac{g\left(D,A\right)}{H_AD} \tag{6}$$

其中 $H_A(D) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|}$，$n$ 为特征 A 取值的个数。

**Algorithm 1:** ID3 算法

**Input:** 训练数据集 D，特征集 A，阈值 $\varepsilon$

**Output:** 决策树 T

1 从根节点开始，对节点计算所有可能的特征的信息增益，选择信息增益最大的特征作为节点的特征，由该特征的不同取值建立子节点；再对子节点递归的调用以上方法，构建决策树；直到所有特征的信息增益均很小或者没有特征可以选择为止。

C4.5 决策树剪枝通过极小化决策树整体的损失函数或代价函数来实现。设树 $T$ 的叶子节点个数为 $|T|$,$t$ 是树 $T$ 的叶节点，该叶节点有 $N_t$ 个样本点，其中 k 类的样本点有 $N_{tk}$ 个，$H_t(T)$ 为叶节点 t 上的经验熵。

$$C_\alpha = \sum_{t=1}^{|T|} N_t H_t(T) + \alpha |T| \tag{7}$$

其中经验熵为：

$$H_t(T) = -\sum_{k} \frac{N_{tk}}{N_t} \log \frac{N_{tk}}{N_t} \tag{8}$$

$C(T)$ 表示模型对训练数据的预测误差，$|T|$ 表示模型负责度，$\alpha \geq 0$ 控制两者之间的影响，较大的 $\alpha$ 促使选择较简单的模型，较小的 $\alpha$ 促使选择教复杂的模型。

# 2 CART 算法

CART 决策树同样由特征选择、树的生成及剪枝组成，既可以用于分类也可以用于回归。CART 假设决策树是二叉树，内部节点特征的取值为' 是' 和' 否', 做左分支是取值为' 是' 的分支，右节点是取值为' 否' 的分支。

## 2.1 回归

一个回归树对应着将输入空间的一个划分以及在划分后的单元上的输出值。假设已将输入空间划分为 M 个单元 $R_1, R_2, \ldots R_M$, 并且在每个单元

$R_m$ 上有一个固定的输出值 $c_m$，于是回归树模型可表示为：

$$f(x) = \sum_{m=1}^{M} C_m I(x \epsilon R_m) \tag{9}$$

对空间进行划分，选择第 $j$ 个变量 $x^j$ 和它的取值 $s$，作为切分变量和切分点并定义两个区域：

$$R_1(j,s) = \left\{x|x^j \leq s\right\} 和 R_2(j,s) = \left\{x|x^j > s\right\} \tag{10}$$

寻找最佳切分变量 $j$ 和最佳切分点 $s$：

$$\underset{j,s}{min}\left[\underset{c1}{min}\sum_{x_i \epsilon R_1(j,s)}(y_i - c_1) + c_2 \sum_{x_i \epsilon R_2(j,s)}(y_i - c_2)^2\right] \tag{11}$$

最优切分点为：

$$ave(y_i|x_i \epsilon R_1(j,s)) 和 \quad ave(y_i|x_i \epsilon R_2(j,s)) \tag{12}$$

## 2.2 分类

概率分布基尼指数定义为：

$$Gini(p) = \sum_{k=1}^{K} p_k(1 - p_k) \tag{13}$$

d 对应二分类问题，概率分布基尼指数为：

$$Gini(p) = 2p(1-p) \tag{14}$$

对于给定的样本集合 D，基尼指数为：

$$Gini(D) = 1 - \sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2 \tag{15}$$

在特征 A 的条件下，集合 D 的基尼指数定义为：

$$Gini(D,A) = \frac{|D_1|}{D}Gini(D_1) + \frac{|D_2|}{D}Gini(D_2) \tag{16}$$