# 决策树

廖星宇

2017 年 6 月 13 日

## 1 基本流程

决策树 (decision tree) 遵循 “分而治之” (divide-and-conquer) 策略，目的为了产生一棵泛化能力强的决策树。

三类情况会递归返回：

1. 当前节点所有样本属于同一类；
2. 当前属性集为空，无法划分，标记为当前结点最多的类；
3. 当前节点样本为空，标记为父节点最多的类

## 2 划分选择

### 2.1 信息增益

“信息熵” (information entropy) 定位为

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

其中 $p_k$ 表示样本集合中 D 中第 k 类样本所占的比例，Ent(D) 的值越小，D 的纯度越高。

“信息增益” (information gain) 根据不同分支节点的样本数不同赋予不同的权重 $|D^v|/|D|$，即样本数越多的分支结点的影响越大，定义为

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{2}$$

信息增益越大，说明使用属性 a 划分获得的“纯度提升”越大，所以选择属性可以通过下面的公式来得到

$$a_* = \arg\max_{a \in A} \quad Gain(D, a) \tag{3}$$

## 2.2 增益率

信息增益对取值数目较多的属性有所偏好，为了减少这种偏好带来的影响，使用“增益率”(gain ratio) 来选择最优划分属性。

$$Gain\ ratio(D, a) = \frac{Gain(D, a)}{IV(a)} \tag{4}$$

其中

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} log_2 \frac{|D^2|}{|D|} \tag{5}$$

需要注意的是，增益率准则对可取数目较少的属性有所偏好。

## 2.3 基尼指数

CART 决策树使用“基尼指数”(Gini index) 来选择划分属性，数据集 D 的纯度可用基尼值来度量

$$\begin{aligned} Gini(D) &= \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} \\ &= 1 - \sum_{k=1}^{|y|} p_k^2 \end{aligned} \tag{6}$$

Gini(D) 反映了数据集中随机两个样本标定不一致的概率，所以 Gini(D) 越小，表示数据集 D 的纯度越高。

属性 a 的基尼指数定义为

$$Gini\ index(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v) \tag{7}$$

所以我们可以依据使得划分后基尼指数最小的属性作为最优划分属性，即

$$a_* = \arg\min_{a \in A} Gini\ index(D, a) \tag{8}$$