原理

决策树学习常用的算法有ID3,C4.5,CART，下面结合这些算法分别叙述学习的特征选择、树的生成和剪枝过程。

特征选择

熵的定义

设X是一个取有限个值的离散随机变量，其概率分布为$P(X=x_i)=p_i,i=1,2,\cdots,n$
则随机变量X的熵的定义为：

$$H(X)=-\sum_{i=1}^{n}p_i\log p_i \qquad (5.1)$$

熵越大，随机变量的不确定性就越大。当随机变量只取两个值，例如1,0,即X的分布为

$$P(X=1)=p,\quad P(X=0)=1-p,\quad 0\leqslant p\leqslant 1$$

熵为

$$H(p)=-p\log_2 p-(1-p)\log_2(1-p) \tag{5.4}$$

这时，熵 $H(p)$ 随概率 p 变化的曲线如图 5.4 所示（单位为比特）.

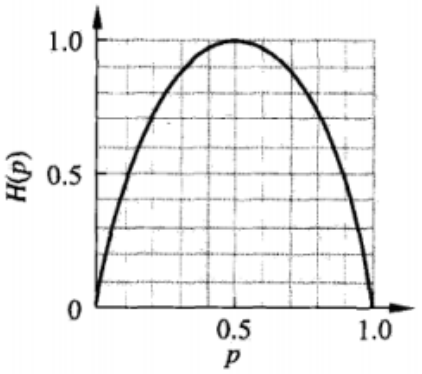

图 5.4　分布为贝努利分布时熵与概率的关系

当 $p=0$ 或 $p=1$ 时 $H(p)=0$，随机变量完全没有不确定性．当 $p=0.5$ 时，$H(p)=1$，熵取值最大，随机变量不确定性最大.

设有随机变量(X,Y)，其联合概率分布为

$P(X=x_i,Y=y_i)=p_{ij},i=1,2,\cdots n,j=1,2,\cdots,m$

条件熵$H(Y|X)$表示在已知随机变量X的条件下随机变量Y的不确定性。随机变量X给定的条件下

随机变量Y的条件熵$H(Y|X)$,

定义为X给定条件下Y的条件概率分布的熵对X的数学期望

$$H(Y\mid X)=\sum_{i=1}^{n}p_i H(Y\mid X=x_i) \tag{5.5}$$

这里，$p_i=P(X=x_i)$，$i=1,2,\cdots,n$.

信息增益

特征A对训练集D的信息增益g(D,A),定义为集合D的经验熵H(D)与H(D|A)的差，即

$$g(D,A)=H(D)-H(D\mid A) \tag{5.6}$$

也称为“互信息”

算法：

（1）计算数据集 D 的经验熵 $H(D)$

$$H(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|} \tag{5.7}$$

（2）计算特征 A 对数据集 D 的经验条件熵 $H(D|A)$

$$H(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|} \tag{5.8}$$

（3）计算信息增益

$$g(D,A)=H(D)-H(D|A) \tag{5.9}$$ ■

信息增益比

定义为信息增益g(D,A)与训练数据集D的经验熵之比：

$$g_R(D,A)=\frac{g(D,A)}{H(D)} \tag{5.10}$$

为什么这样选呢？因为信息增益大小会随着数据集变化。而这个公式相当于把这个变化量除以经验熵，
避免了这个问题。

树的生成

ID3

伪代码：

算法 5.2（ID3 算法）

输入：训练数据集 D，特征集 A，阈值 ε；

输出：决策树 T.

（1）若 D 中所有实例属于同一类 C_k，则 T 为单结点树，并将类 C_k 作为该结点的类标记，返回 T；

（2）若 $A=\varnothing$，则 T 为单结点树，并将 D 中实例数最大的类 C_k 作为该结点的类标记，返回 T；

（3）否则，按算法 5.1 计算 A 中各特征对 D 的信息增益，选择信息增益最大的特征 A_g；

（4）如果 A_g 的信息增益小于阈值 ε，则置 T 为单结点树，并将 D 中实例数最大的类 C_k 作为该结点的类标记，返回 T；

（5）否则，对 A_g 的每一可能值 a_i，依 $A_g=a_i$ 将 D 分割为若干非空子集 D_i，将 D_i 中实例数最大的类作为标记，构建子结点，由结点及其子结点构成树 T，返回 T；

（6）对第 i 个子结点，以 D_i 为训练集，以 $A-\{A_g\}$ 为特征集，递归地调用步 (1) ～步 (5)，得到子树 T_i，返回 T_i. ■

C4.5

C4.5和ID3的区别就是前者采用的是信息增益比