1 决策树

1.1 熵

在信息论与概率统计中, 熵 (Entropy) 是表示随机变量不确定性的度量. 设 X 是一个取有限个值的离散随机变量, 其概率分布为

$$P(X = x_i) = p_i, i = 1, 2, ..., n$$

则随机变量 X 的熵定义为

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

设有随机变量 (X,Y), 其联合概率分布为

$$P(X = x_i, Y = y_i) = P_{ij}, i = 1, 2, ..., n; j = 1, 2, ..., m$$

条件熵 $H(Y|X)$ 表示在已经随机变量 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 (conditional entropy)$H(Y|X)$, 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

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

这里,$p_i = P(X = x_i), i = 1, 2, ..., n$.

1.2 信息增益 (ID3)

特征 A 对训练数据集 D 的信息增益 $g(D, A)$, 定义为集合 D 的经验熵 $H(D)$ 与特征 A 给定条件下 D 的条件经验熵 $H(D|A)$ 之差, 即

$$g(D, A) = H(D) - H(D|A)$$

1.3 信息增益比

以信息增益作为划分训练数据集的特征, 存在偏向于选择取值较多的特征的问题. 使用信息增益比 (information gain ratio) 可以对这一问题进行校

正. 特征 A 对训练集数据集 D 的信息增益比 $g_R(D|A)$ 定义为其信息增益 $g(D|A)$ 与训练数据集 D 关于特征 A 的值的熵 $H_A(D)$ 之比, 即

$$g_R(D,A)=\frac{g(D,A)}{H_A(D)}$$

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

1.4 决策树的剪枝

决策树的剪纸往往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,$k=1,2,...,K$,$H_t(T)$ 为叶节点 t 上的经验熵,$\alpha\leq 0$ 为参数, 则决策树学习的损失函数可以定义为

$$C_\alpha(T)=\sum_{t=1}^{|T|}N_tH_t(T)+\alpha|T|$$

其中经验熵为

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

1.5 基尼指数

分类问题中, 假设有 K 个类, 样本点属于第 k 类的概率为 p_k, 则概率分布的基尼指数定义为

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$$

1.6 CART 剪枝算法

输入: CART 算法生成的决策树 T_0;
输出: 最优决策树 T_α.
(1) 设 $k=0,T=T_0$.
(2) 设 $\alpha=+\infty$.
(3) 自上而下地对各内部节点 t 计算 $C(T_t)$,$|T_t|$ 以及

$$g(t)=\frac{C(t)-C(T_t)}{|T_t|-1},\alpha=\min(\alpha,g(t))$$

这里,T_t 表示以 t 为根结点的子树,$C(T_t)$ 是对训练数据的预测误差,$|T_t$ 是 T_t 的叶结点个数.

(4) 对 $g(t) = \alpha$ 的内部结点 t 进行剪枝, 并对叶结点 t 以多数表决法决定其类, 得到树 T.

(5) 设 $k = k + 1, \alpha_k = \alpha, T_k = T$.

(6) 如果 T_k 不是由根结点及两个叶结点构成的树, 则回到步骤 (3); 否则令 $T_k = T_n$.

(7) 采用交叉验证法在子树序列 $T_0, T_1, ..., T_n$ 中选取最优子树 T_α.