决策树

基本流程

输入: 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
属性集 $A = \{a_1, a_2, \ldots, a_d\}$.
过程: 函数 TreeGenerate(D, A)
生成结点 node;
if D 中样本全属于同一类别 C **then**
  将 node 标记为 C 类叶结点; **return**
end if
if $A = \varnothing$ **OR** D 中样本在 A 上取值相同 **then**
  将 node 标记为叶结点, 其类别标记为 D 中样本数最多的类; **return**
end if
从 A 中选择最优划分属性 a_*;
for a_* 的每一个值 a_*^v **do**
  为 node 生成一个分支; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;
  if D_v 为空 **then**
    将分支结点标记为叶结点, 其类别标记为 D 中样本最多的类; **return**
  else
    以 TreeGenerate(D_v, $A \setminus \{a_*\}$)为分支结点
  end if
end for
输出: 以 node 为根结点的一棵决策树

划分过程

决策树学习的关键是如何选择最优划分属性，希望分支结点所包含的样本尽可能属于同一类别，即结点的“纯度”越来越高。

信息增益

假定当前样本集合D中第k类样本所占比例为$p_k(k=1,2,\cdots,|y|)$，则D的信息熵定义为

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

$Ent(D)$的值越小，则D的纯度最高。

假定离散属性a有V个可能的取值$a^1,a^2,\cdots,a^V$,属性a对样本集D进行划分的信息增益

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

增益率

属性a的可能取值数目越多（即V越大），a的固有值越大，固有值表示为

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

增益率定义为

$$Gain_ratio(D,a)=\frac{Gain(D,a)}{IV(a)}$$

基尼系数

数据集D的纯度可用基尼值来度量

$$Gini(D)=\sum_{k=1}^{|y|}\sum_{k'\neq k}p_kp_{k'}=1-\sum_{|y|}^{k=1}p_k^2$$

基尼指数

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

剪枝处理

通过 `剪枝` 降低过拟合。
两种策略： `预剪枝` 和 `后剪枝` 。

连续与缺失值

连续值处理

二分法最连续属性进行处理。

假定连续属性a在样本集D上出现了n个不同的取值，将这些值从小到大排序，记为$\{a^1, a^2, \cdots, a^n\}$。把区间$[a^i, a^{i+1})$的中位点$\frac{a^i+a^{i+1}}{2}$作为候选划分点，得到信息增益

$$Gain(D,a) = \max_{t \in T_a} Gain(D,a,t) = \max_{t \in T_a} Ent(D) - \sum_{\lambda \in \{-,+\}} \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda)$$

其中$Gain(D,a,t)$是样本集D基于划分点t二分后的信息增益。

缺失值处理

给定训练集D和属性a，令$\widetilde{D}$表示D中在属性a上没有缺失值的样本子集。假定属性a有V个可取值$\{a^1, a^2, \cdots, a^V\}$，令$\widetilde{D}^v$表示在$\widetilde{D}$中属性$a$上取值为$a^v$的样本子集，$\widetilde{D}_k$表示$\widetilde{D}$中属于第$k$类$k(k=1,2,\cdots,|y|)$的样本子集，则显然有$\widetilde{D} = \cup_{k=1}^{|y|} \widetilde{D}_k$，$\widetilde{D} = \cup_{v=1}^{|V|} \widetilde{D}^v$
为样本x赋予权重w_x，定义

$$\rho = \frac{\sum_{x \in \widetilde{D}} w_x}{\sum_{x \in D} w_x}$$

$$\tilde{p}_k = \frac{\sum_{x \in \widetilde{D}_k} w_x}{\sum_{x \in \widetilde{D}} w_x} \qquad (1 \le k \le |y|)$$

$$\tilde{r}_v = \frac{\sum_{x \in \widetilde{D}^v} w_x}{\sum_{x \in \widetilde{D}} w_x} \qquad (1 \le v \le V)$$

对属性a，ρ表示无缺失值样本所占的比例，$\tilde{p}_k$表示无缺失值样本中第k类所占的比例，$\tilde{r}_v$则表示无缺失值样本中在属性a上取值a^v的样本所占的比例。
可将信息增益推广为

$$Gain(D,a) = \rho \times Gain(\widetilde{D}, a) = \rho \times (Ent(\widetilde{D}) - \sum_{v=1}^{V} \tilde{r}_v Ent(\widetilde{D}^v))$$

其中，

$$Ent(\widetilde{D}) = -\sum_{k=1}^{|y|} \tilde{p}_k \log_2 \tilde{p}_k$$

多变量决策树

在多变量决策树的学习过程中，不是为了每个非叶结点寻找一个最优划分属性，而是试图简历一个合适的线性分类器。