决策树流程

● 输入：

训练集D={$(x_1, y_1),(x_2, y_2),\ldots(x_m, y_m)$ }

属性集A={$a_1, a_2, \ldots a_d$}

● 生成TreeGenerate(D, A)的具体实现：

- 若D全属于同一类别C，则将该分支标记为叶结点，类别标记为C
- 若A=Ø（属性集为空）或者D在A上的所有属性的取值都相同，则将该分支标记为叶结点，类别标记为D中最多的类

① 从A中选出最优划分属性a_*

② for（a_*的每个属性值a_*^v）：

令D_v为D中属性a_*的取值为a_*^v的子集

if D_v为空（即该属性值不含任何样本）：

- 将该分支标记为叶结点，类别标记为D中最多的类——停止条件

else（非空）

生成TreeGenerate(D_v, A\{a_*})　——递归（对子集继续进行分支）

说明：决策树的生成是运用了递归，在递归式里有3个停止条件。

（1）信息熵

Ent(D) = - $\sum_{k=1}^{K} p_k log_2 p_k$　　k={1,2,…,K}——最终评判的分类类别

作用：度量样本集的不确定性（区分度）。值越大，区分度越弱。

范围：[0，$log_2 K$]

当p_i=1，其余 $p_j(j \neq i)$=0 时，Ent(D) =0

当$p_1 = p_2 = \ldots = p_K$=1/K时，Ent(D) =$log_2 K$

（2）信息增益

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

解释：使用属性a进行划分，能提升的区分度的增加程度。值越大，说明该属性越能区分样本。

特点：对可取值数目较多的属性有所偏好。

2、增益率

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

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

特点：对可取值数目较少的属性有所偏好。

3、基尼指数

（1）基尼值

Gini(D) = 1 - $\sum_{k=1}^{K}p_k^2$ ——从训练集D中任意取2个样本，2个样本类别不一致的概率。

基尼值越大，样本区分度越高。

（2）基尼指数

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