算法学习-决策树

1.什么是决策树

决策树（Decision Tree）是一类常见的机器学习算法。比如，我们可以从给定的数据集学习得到一个学习模型。通过这个模型，我们可以对新添加示例进行分类。这个的一个过程可以称作"决策"或者"判定"过程。决策树是基于树结构来进行决策的。

对于一个问题进行决策时，我们常常会进行一系列的判断或者"子决策"。一般的，一颗决策树包含一个根节点，若干个内部节点和若干个子叶节点。叶节点对应决策结果，其他每个节点则对应一个属性测试。每个节点包含的样本集合根据属性测试的结果被划分到子结点中；根节点包括样本全集。从根节点到每个叶节点的路径对应了一个判定测试序列。决策树的目的时产生一棵泛化能力强，处理未见示例能力强的决策树。其基本流程遵循简单且直观的"分而治之"（Divide-and-conquer）的策略。

2.决策树的基本流程

决策树的基本流程可以用以下流程概括

输入: 训练集D =$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$;

属性集A = $\{a_1, a_2, \ldots, a_d\}$.

过程： 函数TreeGenerate(D,A)

1:生成节点node;

2.**if** D中样本属性全属于同一类别C **then**

3: 将node标记为C类叶节点 ;**return**

4:**end if**

5:**if** A = ∅ **or** D中样本在A上取值相同 **then**

6: 将node标记为叶节点，其类别标记为D中样本最多的类；**return**

7:**end if**

8:从A中选择最优划分属性a_*;

9:**for** a_*中每一个值a_*^v **do**

10: 为node生成一个分支；令D_v表示D在a_*^v上取值的样本子集;

11: **if** D_v为空 **then**

12: 将分支节点标记为叶节点，其类别标记为D中样本最多的类；**return**

13: **else**

14: 以TreeGenerate(D_v,$A \setminus a_*$)为分支节点

15: **end if**

16:**end for**

输出： 以node 为根节点的一颗决策树

3.决策树的建立

3.1 划分选择

3.1.1信息增益

首先是信息熵的定义，信息熵是衡量信息不确定性的指标。设样本集合D 中第K 类样本所占的比例为 $p_k(k=1,2,...|\gamma|)$,则样本 D 的信息熵定义为

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

$Ent(D)$ 的值越小，则 D 的纯度越高.

若离散属性 a 有 V 个取值 $\{a^1, a^2, ... a^v\}$.我么可以使用 a 来对样本集进行划分，则可得到 V 个分支节点。第v个分支节点包含了 D 中所有属性 a 上取值为 a^v 的样本，我们记为 D^v .根据信息熵的定义可以求出D^v 的信息熵，然后根据 D^v在整个样本集中的权重，我们可以求出属性a 对样本集的信息增益。

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

通常情况下，信息增益越大，说明使用属性a 获得的“纯度提升越大”。这样我们可以用信息增益来进行决策树的划分属性的选择。著名的 $ID3$ 决策树学习算法就是以信息增益为准则进行划分属性选择。

3.1.2信息增益率

实际上，信息增益率会对可取值较多的属性有所偏好。为了避免这种偏好可能产生的影响，我们可以使用增益率来进行划分选择。著名的$C4.5$ 决策树算法就是采用的增益率。增益率定义为：

$$Gain_{ratio}(D,a) = \frac{Gain(D,a)}{IV(a)} \tag{3}$$

其中

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|} \tag{4}$$

$IV(a)$ 称之为 a 的 ”固有值“，若a 属性的可取值越多，则$IV(a)$ 值通常越大。

需要注意的是增益率对属性可取值数量少的属性有所偏好。因此我们在划分选择时，先从划分属性中选择信息增益高于平均值的，然后在从中选择信息增益率高的。

3.1.3 基尼系数

$CART$ 决策树使用基尼系数来选择划分属性。数据集D 的纯度可以使用基尼值来进行度量，

$$
\begin{aligned}
Gini(D) &= \sum_{k=1}^{|\gamma|} \sum_{k' \neq k} p_k p'_k \\
&= 1 - \sum_{k=1}^{|\gamma|} p_k^2
\end{aligned} \tag{5}
$$

$Gini(D)$ 反映了从数据集 D 中随机取两个样本，其类别标记不一致的概率，因此$Gini(D)$越小，数据纯度越高。

若将(5)式变为与(2)式相同形式，则

$$
Gini_{index}(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v) \tag{6}
$$

若根据基尼系数来进行划分选择，我们可以选择使得基尼系数最小的那个属性作为最有划分选择。即：

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

3.2 剪枝处理

3.3 连续值与缺失值处理