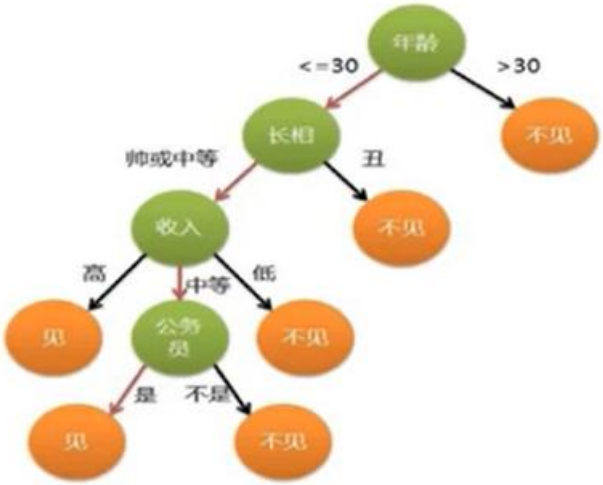
决策树算法

1、定义

例如有人给我们介绍新的对象的时候，我们就要一个个特点去判断，于是这种判断的过程就可以画成一棵树，例如根据特点依次判断：



如上，决策的形式以树的形式进行示意和编码，就形成了决策树。

2、划分选择

一般的原则是，希望通过不断划分节点，使得一个分支节点包含的数据尽可能的属于同一个类别，即“纯度“越来越高。

3、信息增益准则

我们先对一个节点的纯度进行定义，我们将其称之为信息熵。

$$\mathrm{E}nt(D) = -\sum_{k=1}^{|\gamma|} p_k \log_2(p_k)$$

由于 p_k 都属于[0,1]，Ent(D)必定为正值，值越大说明纯度越低，**值越小说明纯度越高**。

银行贷款数据

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

信息熵计算（信息熵的值是根据目标值是求信息熵的）：

类别(是否给贷款)：否表示不贷款（6 个样本），是表示贷款（9 个样本）；

$$\mathrm{Ent}(D) = -\frac{6}{15}\log_2\frac{6}{15} + (-\frac{9}{15}\log_2\frac{9}{15}) = 0.9709505944546686$$

4、信息增益(ID3 算法)

在定义了信息熵之后，对信息增益进行定义，假设选取属性 a 有 V 个取值，$\{a^1, a^2 ..... a^V\}$，按照决策树的规则，D 将被划分为 V 个不同的节点数据集，D^v 代表其中第 v 个节点：

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

由此，我们得到了一种选择划分属性的方法，计算以每个属性进行划分子节点得到的信息增益，选择其中最大的作为选择的属性。

信息增益计算：

$$\text{G}ain(D,\text{年龄}) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$
$$= \text{Ent}(D) - (\frac{5}{15} Ent(\text{青年}) + \frac{5}{15} Ent(\text{中年}) + \frac{5}{15} Ent(\text{老年}))$$

$$\text{Ent}(\text{青年}) = -\frac{2}{5}\log_2\frac{2}{5} + (-\frac{3}{5}\log_2\frac{3}{5}) = 0.9709505944546686$$
$$\text{Ent}(\text{中年}) = -\frac{2}{5}\log_2\frac{2}{5} + (-\frac{3}{5}\log_2\frac{3}{5}) = 0.9709505944546686$$
$$\text{Ent}(\text{老年}) = -\frac{4}{5}\log_2\frac{4}{5} + (-\frac{1}{5}\log_2\frac{1}{5}) = 0.7219280948873623$$

所以计算得：

$$\text{G}ain(D,\text{年龄}) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$
$$= \text{Ent}(D) - (\frac{5}{15} Ent(\text{青年}) + \frac{5}{15} Ent(\text{中年}) + \frac{5}{15} Ent(\text{老年}))$$
$$= 0.9709505944546686 - (\frac{5}{15} * 0.9709505944546686 + \frac{5}{15} * 0.9709505944546686 + \frac{5}{15} * 0.7219280948873623)$$
$$= 0.08300749985576883$$

$$\text{G}ain(D,\text{有工作}) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$
$$= \text{Ent}(D) - (\frac{5}{15} Ent(\text{有工作}) + \frac{10}{15} Ent(\text{无工作}))$$

$$\text{Ent}(\text{有工作}) = -\frac{0}{5}\log_2\frac{0}{5} + (-\frac{5}{5}\log_2\frac{5}{5}) = 0$$
$$\text{Ent}(\text{无工作}) = -\frac{4}{10}\log_2\frac{4}{10} + (-\frac{6}{10}\log_2\frac{6}{10}) = 0.9709505944546686$$

假如最后我们计算出来年龄 A1、有工作 A2、有自己房子 A3、信贷情况 A4，4 个特征的信息增益比分别是：0.083，0.324，0.420，0.363，**相比较来说其中特征 A3（有自己房子）的信息增益比最大，所以我们选择特征 A3 为最有特征**。

5、信息增益率(C4.5 算法)

信息增益原则对于每个分支节点，都会乘以其权重，也就是说，由于权重之和为 1，所以分支节点分的越多，即每个节点数据越小，纯度可能越高。这样会导致信息熵准则偏爱那些取值数目较多的属性。

为了解决该问题，这里引入了信息增益率，定义如下：

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

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

相当于引入了修正项 IV(a)，它是对于属性 a 的固有值。

需要注意的是，信息增益率原则可能对取值数目较少的属性更加偏爱,为了解决这个问题，可以先找出信息增益在平均值以上的属性，在从中选择信息增益率最高的。

6、基尼指数

在 CART 决策树中，使用基尼指数来选择属性，首先定义数据集 D 的基尼值：

$$Gini(D)=\sum_{k=1}^{|\gamma|}\sum_{k^1!=k}p_k p_{k^1}=1-\sum_{k=1}^{|\gamma|}{p_k}^2$$

基尼值的计算：

范例 1：

Label=0	5	P(0)=0.5
Label=1	5	P(1)=0.5

$$Gini(D)=1-((\frac{1}{2})^2+(\frac{1}{2})^2)=0.5$$

范例 2：

Label=0	2	P(0)=0.2
Label=1	8	P(1)=0.8

$$Gini(D)=1-((\frac{2}{10})^2+(\frac{8}{10})^2)=0.32$$

范例 3：

Label=0	0	P(0)=0
Label=1	10	P(1)=1

$$Gini(D)=1-((\frac{0}{10})^2+(\frac{10}{10})^2)=0$$

基尼指数(**基尼值的加权平均**)：

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

基尼指数越小，说明纯度越高，我们可以通过**选择基尼指数小的属性来划分子节点**。

综合案例：

Id	有房者	婚姻	年收入	是否拖欠贷款
1	是	单身	125k	否
2	否	已婚	100k	否
3	否	单身	70k	否
4	是	已婚	120k	否
5	否	离异	95k	是
6	否	已婚	60k	否
7	是	离异	220k	否
8	否	单身	85k	是
9	否	已婚	75k	否
10	否	单身	90k	是

基尼值计算：

有房者：

是(3) : Label(是否拖欠贷款)：是(0) 否(3)

$$Gini(D) = 1 - (0^2 + (\frac{3}{3})^2) = 0$$

否(7): Label(是否拖欠贷款)：是(3) 否(4)

$$Gini(D) = 1 - ((\frac{3}{7})^2 + (\frac{4}{7})^2) = \frac{25}{49}$$

Gini 加权平均：

$$Gini(有房者) = \frac{3}{10} \cdot 0 + \frac{7}{10} \cdot \frac{25}{49} = \frac{5}{14}$$

同理，计算婚姻：

单身/离异(6): Label(是否拖欠贷款)：是(3) 否(3)

$$Gini(D) = 1 - ((\frac{1}{2})^2 + (\frac{1}{2})^2) = 0.5$$

已婚(4): Label(是否拖欠贷款)：是(0) 否(4)

$$Gini(D) = 1 - ((\frac{0}{4})^2 + (\frac{4}{4})^2) = 0$$

Gini 加权平均：

$$\text{Gi}ni(\text{婚姻}) = \frac{6}{10} \cdot 0.5 + \frac{4}{10} \cdot 0 = \frac{3}{10}$$

同理，**计算年收入**(规定>=80k 不会拖欠贷款，<80k 的会拖欠贷款)：

>=80k(7): Label(是否拖欠贷款)：是(3) 否(4)

$$\text{Gi}ni(D) = 1 - ((\frac{3}{7})^2 + (\frac{4}{7})^2) = \frac{25}{49}$$

<80k(3): Label(是否拖欠贷款)：是(0) 否(3)

$$\text{Gi}ni(D) = 1 - ((\frac{0}{3})^2 + (\frac{3}{3})^2) = 0$$

Gini 加权平均：

$$\text{Gi}ni(\text{婚姻}) = \frac{7}{10} \cdot \frac{25}{49} + \frac{3}{10} \cdot 0 = \frac{5}{14}$$

比较三者的 Gini 的加权平均值：最小值是**婚姻，所以婚姻作为根节点。**