第四次作业——决策树

**2.1决策树如何进行“剪枝”处理？**

①剪枝（pruning）的目的是为了**避免决策树模型的过拟合**。

②决策树算法在学习的过程中为了尽可能的正确的分类训练样本，不停地对

结点进行划分，因此这会导致整棵树的分支过多，也就导致了过拟合。

③决策树的剪枝策略最基本的有两种：**预剪枝**（pre-pruning）和**后剪枝**

（post-pruning）：

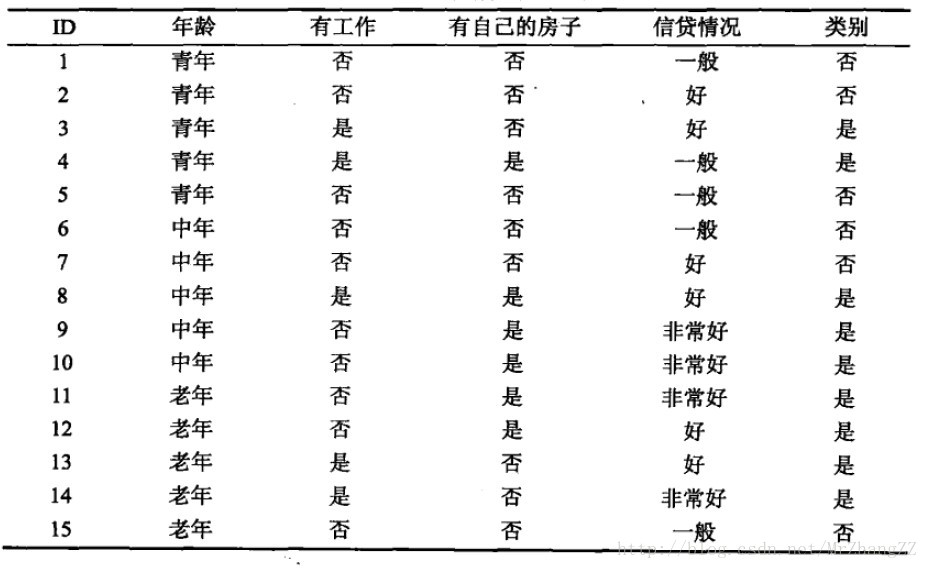
* **预剪枝**（pre-pruning）：预剪枝就是在构造决策树的过程中，先对每个结点在划分前进行估计，若果当前结点的划分**不能带来决策树模型泛华性能的提升**，则不对当前结点进行划分并且将当前结点标记为叶结点。
* **后剪枝**（post-pruning）：后剪枝就是先把整颗决策树构造完毕，然后**自底向上**的对非叶结点进行考察，若将该结点对应的子树换为叶结点**能够带来泛华性能的提升**，则把该子树替换为叶结点。

**2.2****试析使用“最小训练误差”作为决策树划分选择的缺陷。**

若以最小训练误差作为决策树划分的依据，由于训练集和真是情况总是会存在一定偏差，这使得这样得到的**决策树会存在过拟合**的情况，对于未知的数据的**泛化能力较差**。因此最小训练误差不适合用来作为决策树划分的依据

**2.3对表2-1的训练数据集，根据信息增益准则选择最优特征，列出计算步骤； 建立决策树，画出该决策树。**

表格 2-1：训练数据集



解：已知信息熵的定义为：



信息增益的定义为：



具体计算过程如下：

1. **计算整个训练集合的根节点信息熵：**

共有15条数据，类别中是有9条，否有6条

0.971

1. **计算属性集合{年龄，有工作，有自己的房子，信贷情况}，下简称为{年龄，工作，房子，信贷}中每个属性的信息增益**

以年龄为例，训练集D可划分为3个子集，分别为

* ，包含编号{1,2,3,4,5}，其中类别是否分别的概率为，
* ，包含编号{6,7,8,9,10}，其中类别是否分别的概率为，
* ，包含编号{11,12,13,14,15}，其中类别是否分别概率为，

根据式(1.1)可以算出根据年龄划分后的三个子集的信息熵为







从而计算出信息增益为：



同理可以算出工作，房子，信贷情况的信息增益为：







显然，属性“有自己的房子”信息增益最大，故选他为划分属性，划分结果

D:有自己的房子=？

是

否

：1,2,3,5,6,7,13,14,15

：4,8,9,10,11,12

1. **然后对每一个分支节点进行划分，可用属性为{年龄，有工作，信贷情况}**

数据集其类别标签都是是，故不再继续划分；

下对数据集进行划分：

其，，

计算各属性信息增益如下：







显然对于数据集，属性“有工作”的信息增益最大，选其为划分属性,划分结果如下：

有工作=？

是

否

：1,2,5,6,7,15

：3,13,14

对于数据集，其标签类别都是“是”，故不需要再次划分

对于数据集，其标签类别都是“否”，故也不需要再次划分

1. **综上，决策树构建完成，绘制出决策树如下：**

有自己的房子

否

是

否

是

是

是

有工作

否

**2.4 考虑表2-2中二元分类问题的训练样本。**

**(a)计算整个训练样本集的Gini指标值。**

**(b)计算属性顾客ID的Gini指标值。**

**(c)计算属性性别的Gini指标值。**

**(d)计算使用多路划分属性车型的Gini指标值。**

**(e)计算使用多路划分属性衬衣尺码的Gini指标值。**

**(f)下面哪个属性更好，性别、车型还是衬衣尺码？**

**(g)解释为什么属性顾客ID的Gini值最低，但是不能作为属性测试条件。**

表格2-2



解：已知

1. 由上表可得整个训练样本可表示为

|  |  |
| --- | --- |
| C0 | C1 |
| 10 | 10 |

故 

训练样本集

1. 由表格内容可知，共有20名顾客，且每个顾客的都不一样

因此，

也即属性顾客的指标值都是0

1. 由表格内容可得

|  |  |
| --- | --- |
| 男 | 女 |
| 10 | 10 |

故 

因此属性性别

(d)由表格内容可得

|  |  |  |  |
| --- | --- | --- | --- |
| 类 | 车型 | | |
| 家用 | 运动 | 豪华 |
| C0 | 1 | 8 | 1 |
| C1 | 3 | 0 | 7 |

1. 家用车型的指标计算如下：





1. 运动车型的指标计算如下：

运动车型的C1类数量为0，显然可得

1. 豪华车型的指标计算如下：





因此车型属性的总计算如下;



1. 由表格内容可得：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类 | 衬衣尺码 | | | |
| 小 | 中 | 大 | 加大 |
| C0 | 3 | 3 | 2 | 2 |
| C1 | 2 | 4 | 2 | 2 |

1. 小尺码的指标计算如下：





1. 中尺码的指标计算如下：





1. 大尺码的指标计算如下：





1. 加大尺码的指标计算如下：





因此属性衬衣尺码的总指标计算如下：



1. 有上述几个小问的解答可得

|  |  |  |  |
| --- | --- | --- | --- |
| 属性 | 性别 | 车型 | 衬衣尺码 |
|  | 0.5 | 0.1625 | 0.4914 |

由上表内容，显而易见可以得出**车型属性更好**，车型的指标值最低，其**子节点不纯度更高**，产生了**更纯的派生节点**

1. 属性顾客只是一个自己设定的标记，用于区分不同的顾客，没有预测性，因为**与每个划分相关联的记录很少，故不足以做出可靠的预测**。因此虽然属性顾客的指标值最低，但不能作为属性测试条件