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

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

解：已知信息熵的定义为：

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k \tag{1.1}$$

信息增益的定义为：

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{1.2}$$

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

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k = -\left(\frac{9}{15}\log_2\frac{9}{15} + \frac{6}{15}\log_2\frac{6}{15}\right) = 0.971$$

② 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

- D^1(年龄＝青年)，包含编号{1, 2, 3, 4, 5}，其中类别是否分别的概率为 $p_{是} = \frac{2}{5} = 0.4$，$p_{否} = \frac{3}{5} = 0.6$
- D^2(年龄＝中年)，包含编号{6, 7, 8, 9, 10}，其中类别是否分别的概率为 $p_{是} = \frac{1}{5} = 0.2$，$p_{否} = \frac{4}{5} = 0.8$
- D^3(年龄＝老年)，包含编号{11, 12, 13, 14, 15}，其中类别是否分别概率为 $p_{是} = \frac{2}{5} = 0.4$，$p_{否} = \frac{3}{5} = 0.6$

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

$$Ent(D^1) = -\sum_{k=1}^{|y|} p_k \log_2 p_k = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.971$$

$$Ent(D^2) = -\sum_{k=1}^{|y|} p_k \log_2 p_k = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$$

$$Ent(D^3) = -\sum_{k=1}^{|y|} p_k \log_2 p_k = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.971$$

从而计算出信息增益为：

$$Gain(D,\text{年龄}) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) = 0.971 - \frac{1}{3}(0.971*2 + 0.722) = 0.083$$

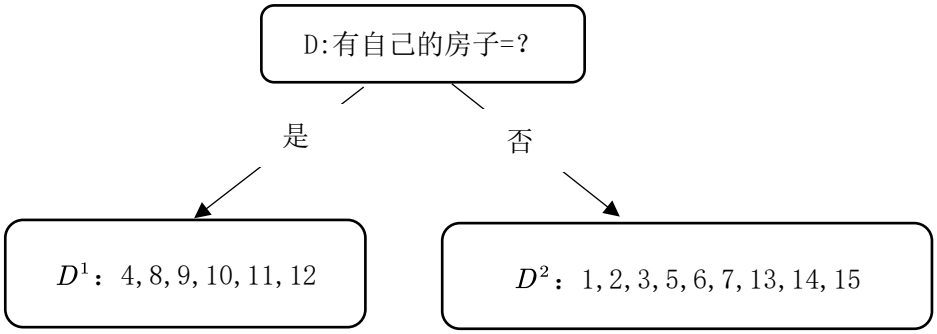
同理可以算出工作，房子，信贷情况的信息增益为：

$Gain(D,\text{有工作}) = 0.324$

$Gain(D,\text{有自己的房子}) = 0.420$

$Gain(D,\text{信贷情况}) = 0.363$

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



③ 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}

D^1数据集其类别标签都是是，故不再继续划分；

下对D^2数据集进行划分：

其$p_{是} = \frac{3}{9} = 0.34$，$p_{否} = \frac{6}{9} = 0.67$，$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k = 0.918$

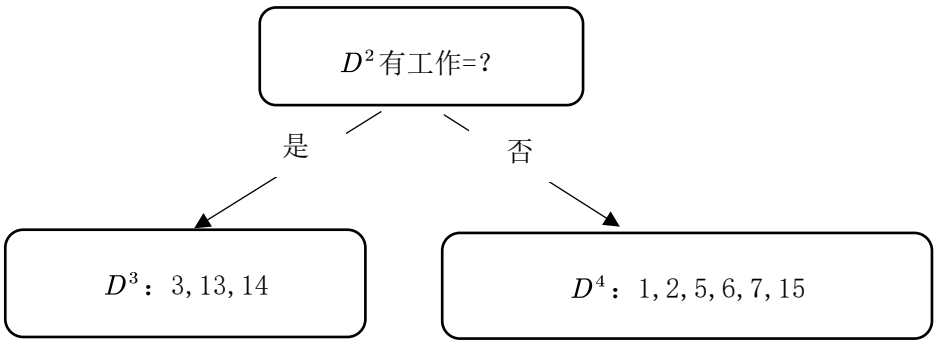
计算各属性信息增益如下：

$$Gain(D^2,\text{年龄}) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) = 0.252$$

$$Gain(D^2,\text{有工作}) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) = 0.918$$

$$Gain(D^2,\text{信贷情况}) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) = 0.474$$

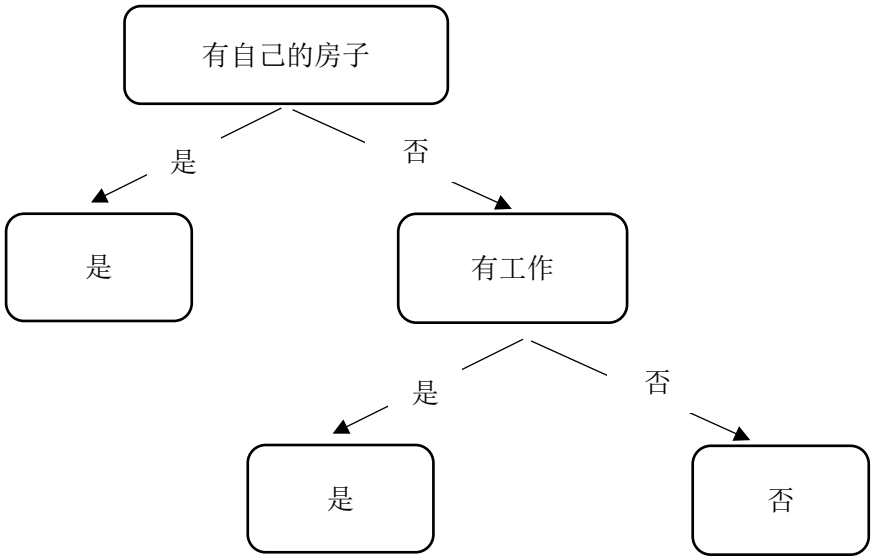
显然对于D^2数据集，属性“有工作”的信息增益最大，选其为划分属性，划分结果如下：



对于D^3数据集，其标签类别都是“是”，故不需要再次划分

对于D^4数据集，其标签类别都是“否”，故也不需要再次划分

④ 综上，决策树构建完成，绘制出决策树如下：



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

顾客 ID	性别	车型	衬衣尺码	类
1	男	家用	小	C0
2	男	运动	中	C0
3	男	运动	中	C0
4	男	运动	大	C0
5	男	运动	加大	C0
6	男	运动	加大	C0
7	女	运动	小	C0
8	女	运动	小	C0
9	女	运动	中	C0
10	女	豪华	大	C0
11	男	家用	大	C1
12	男	家用	加大	C1
13	男	家用	中	C1
14	男	豪华	加大	C1
15	女	豪华	小	C1
16	女	豪华	小	C1
17	女	豪华	中	C1
18	女	豪华	中	C1
19	女	豪华	中	C1
20	女	豪华	大	C1

解：已知$GINI(t) = 1 - \sum [p(j|t)]^2$

(a) 由上表可得整个训练样本可表示为

C0	C1
10	10

故$p(C0) = 10/20 = 0.5$ $p(C1) = 10/20 = 0.5$

训练样本集$GINI = 1 - p(C0)^2 - p(C1)^2 = 1 - 0.25 - 0.25 = 0.5$

(b) 由表格内容可知，共有 20 名顾客，且每个顾客的ID都不一样

因此$p(j|t) = 1$ 故$GINI = 1 - \sum [p(j|t)]^2 = 0$，

也即属性顾客ID的$GINI$指标值都是 0

(c) 由表格内容可得

男	女
10	10

故$p(男) = 10/20 = 0.5$ $p(女) = 10/20 = 0.5$

因此属性性别$GINI = 1 - p(男)^2 - p(女)^2 = 1 - 0.25 - 0.25 = 0.5$

(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$GINI$指标计算如下：

$p(C0) = 1/4 = 0.25$ $p(C1) = 3/4 = 0.75$

$GINI_{家用} = 1 - p(C0)^2 - p(C1)^2 = 1 - 0.25^2 - 0.75^2 = 0.375$

② 运动车型的$GINI$指标计算如下：

运动车型的 C1 类数量为 0，显然可得$GINI_{运动} = 0$

③ 豪华车型的$GINI$指标计算如下：

$p(C0) = 1/8 = 0.125$ $p(C1) = 7/8 = 0.875$

$GINI_{豪华} = 1 - p(C0)^2 - p(C1)^2 = 1 - 0.125^2 - 0.875^2 = 0.21875$

因此车型属性的总$GINI$计算如下;

$$GINI = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) = 4/20 * 0.375 + 8/20 * 0.21875 = 0.1625$$

(e) 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*GINI*指标计算如下：

$p(C0)=1/8=0.125 \qquad p(C1)=7/8=0.875$

$GINI_{小}=1-p(C0)^2-p(C1)^2=1-0.6^2-0.4^2=0.48$

② 中尺码的*GINI*指标计算如下：

$p(C0)=3/7=0.428571 \qquad p(C1)=4/7=0.571429$

$GINI_{中}=1-p(C0)^2-p(C1)^2=1-0.429^2-0.571^2=0.4898$

③ 大尺码的*GINI*指标计算如下：

$p(C0)=2/4=0.5 \qquad p(C1)=2/4=0.5$

$GINI_{大}=1-p(C0)^2-p(C1)^2=1-0.5^2-0.5^2=0.5$

④ 加大尺码的*GINI*指标计算如下：

$p(C0)=2/4=0.5 \qquad p(C1)=2/4=0.5$

$GINI_{加大}=1-p(C0)^2-p(C1)^2=1-0.5^2-0.5^2=0.5$

因此属性衬衣尺码的总*GINI*指标计算如下：

$$GINI=\sum_{i=1}^{k}\frac{n_i}{n}GINI(i)=5/20*0.48+7/20*0.4898+4/20*0.5*2=0.49143$$

(f) 有上述几个小问的解答可得

属性	性别	车型	衬衣尺码
GINI	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*GINI*指标值最低，其**子节点不纯度更高**，产生了**更纯的派生节点**

(g) 属性顾客*ID*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*ID*的*GINI*指标值最低，但不能作为属性测试条件