姓名：麻超
学号：201300066

一. (20 points) **利用信息熵进行决策树划分**

1. 对于不含冲突样本（即属性值相同但标记不同的样本）的训练集, 必存在与训练集一致（训练误差为 0）的决策树. 如果训练集可以包含无穷多个样本, 是否一定存在与训练集一致的深度有限的决策树? 并说明理由 (仅考虑每次划分仅包含一次属性判断的决策树).
2. 信息熵 $\mathrm{Ent}(D)$ 定义如下

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

请证明信息熵的上下界为

$$0 \leq \mathrm{Ent}(D) \leq \log_2 |\mathcal{Y}| \tag{2}$$

并给出等号成立的条件.

3. 在 ID3 决策树的生成过程中, 需要计算信息增益（information gain）以生成新的结点. 设离散属性 a 有 V 个可能取值 $\{a^1, a^2, \cdots, a^V\}$, 请考教材 4.2.1 节相关符号的定义证明:

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}\left(D^v\right) \geq 0 \tag{3}$$

即信息增益非负.

解：

1. 存在.

   已知对于不含冲突样本的训练集, 存在和训练集一致的决策树, 那么当决策树达到最大深度之后, 每一个叶节点必然会对应唯一的一个样本或者所有属性都和标记相同的多个样本, 即叶节点标记与样本自身标记相同, 此时训练误差为 0, 也就是训练集和决策树一致. 因为样本特征数量有限, 所以属性数量同时决定了决策树深度最大值, 故决策树深度有限.

2. $\mathrm{Ent}(D) = f(p_1, p_2, ..., p_n) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$
   当 $0 \le p_k \le 1$ 时, 此问题为凸优化问题, 其 Largrange 函数如下:

$$L(p_1, p_2, ...p_{|\mathcal{Y}|}, \lambda) = \sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k + \lambda(\sum_{k=1}^{|\mathcal{Y}|} p_k - 1)$$

$$\frac{\partial L}{\partial p_n} = \frac{\partial}{\partial p_1}[\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k + \lambda(\sum_{k=1}^{|\mathcal{Y}|} -1)] = 0$$

$$\lambda = -\log_2 p_n - \frac{1}{\ln 2}$$

$$\because \sum_{k=1}^{|\mathcal{Y}|} p_k = 1$$

$$\therefore p_1 = p_2 = ... = p_n = \frac{1}{|\mathcal{Y}|}$$

$$\therefore f(p_1, p_2, ..., p_n) = -\sum_{k=1}^{|\mathcal{Y}|} \frac{1}{|\mathcal{Y}|} \log_2 \frac{1}{|\mathcal{Y}|} = \log_2 |\mathcal{Y}|$$

$$\therefore \quad \frac{1}{|\mathcal{Y}|}, \qquad p_1 = p_2 = ... = p_n = \frac{1}{|\mathcal{Y}|}$$

   由于 $\sum_{k=1}^{|\mathcal{Y}|} p_k = 1$ 且 $0 \le x_k \le 1$ 对 $\mathrm{Ent}(D)$ 的每一项 $-p_k \log_2 p_k \ge 0$.
   $\therefore \mathrm{Ent}(D) \ge 0$, 当 $p_1 = 1, p_2 = ... = p_{|\mathcal{Y}|} = 0$ 时取等.

3. 设 n 为样本类数, 将 $\mathrm{Ent}(D) = -\sum_{k=1}^{n} p_k \log_2 p_k$ 代入可得:

$$\begin{aligned} Gain(D, a) &= -\sum_{k=1}^{n} p_k \log_2 p_k - \sum_{v=1}^{V} \frac{|D^v|}{|D|}(-\sum_{k=1}^{n} p_{vk} \log_2 p_{vk}) \\ &= -\sum_{k=1}^{n} p_k \log_2 p_k + \sum_{k=1}^{n}\sum_{v=1}^{V} \frac{|D^v|}{|D|} p_{vk} \log_2 p_{vk} \\ &= \sum_{k=1}^{n}(\sum_{v=1}^{V} \frac{|D^v|}{|D|} p_{vk} \log_2 p_{vk} - p_k \log_2 p_k) \end{aligned}$$

其中 $\sum_{k=1}^{n}=1, \sum_{v=1}^{V}\frac{|D^v|}{|D|}=1, \sum_{k=1}^{n}\sum_{v=1}^{V}\frac{|D^v|}{|D|}p_{vk}=1, p_k=\sum_{v=1}^{V}\frac{|D^v|}{|D|}p_{vk}$
设 $f(p)=p\log_2 p$, 那么 $f'(p)=\log_2 p+\frac{1}{\ln 2}$, 则 $f(p)$ 为凸函数. 由 $Jensen$ 不等式,$f(p_k)=f(\sum_{v=1}^{V}\frac{|D^v|}{|D|}p_{vk})\leq\sum_{v=1}^{V}\frac{|D^v|}{|D|}f(p_{vk})$,
$p_k\log_2 p_k=\sum_{v=1}^{V}\frac{|D^v|}{|D|}p_{vk}\log_2(\sum_{v=1}^{V}\frac{|D^v|}{|D|}p_{vk})\leq\sum_{v=1}^{V}\frac{|D^v|}{|D|}p_{vk}\log_2 p_{vk}$
移项可得,$\sum_{v=1}^{V}\frac{|D^v|}{|D|}p_{vk}\log_2 p_{vk}-p_k\log_2 p_k\geq 0$
代入可得 $Gain(D,a)=\sum_{k=1}^{n}(\sum_{v=1}^{V}\frac{|D^v|}{|D|}p_{vk}\log_2 p_{vk}-p_k\log_2 p_k)\geq 0$. 得证.

二. (15 points) **决策树划分计算**

本题主要展现决策树在不同划分标准下划分的具体计算过程. 假设一个包含三个布尔属性 X,Y,Z 的属性空间, 目标函数 $f=f(X,Y,Z)$ 作为标记空间, 它们形成的数据集如1所示.

编号	X	Y	Z	f	编号	X	Y	Z	f
1	1	0	1	1	5	0	1	0	0
2	1	1	0	0	6	0	0	1	0
3	0	0	0	0	7	1	0	0	0
4	0	1	1	1	8	1	1	1	0

Table 1: 布尔运算样例表

1. 请使用信息增益作为划分准则画出决策树的生成过程. 当两个属性信息增益相同时, 依据字母顺序选择属性.
2. 请使用基尼指数作为划分准则画出决策树的生成过程, 当两个属性基尼指数相同时, 依据字母顺序选择属性.

解:

1. 布尔属性集合 X,Y,Z
根节点的信息熵 $\mathrm{Ent}(D)=-(\frac{1}{4}\log_2\frac{1}{4}+\frac{3}{4}\log_2\frac{3}{4})=0.811$
分别以 X,Y,Z 对 D 进行划分, 得到的子集分别为 D^0, D^1.

$$
\begin{aligned}
X:&\mathrm{Ent}(D^0) = -(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}) = 0.811\\
&\mathrm{Ent}(D^1) = \mathrm{Ent}(D^0) = 0.811\\
&Gain(D,X) = 0.811 - (\frac{1}{2}\times 0.811 + \frac{1}{2}\times 0.811) = 0\\
Y:&\mathrm{Ent}(D^0) = -(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}) = 0.811\\
&\mathrm{Ent}(D^1) = \mathrm{Ent}(D^0) = 0.811\\
&Gain(D,Y) = 0.811 - (\frac{1}{2}\times 0.811 + \frac{1}{2}\times 0.811) = 0\\
Z:&\mathrm{Ent}(D^0) = -(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}) = 1.000\\
&\mathrm{Ent}(D^1) = -(0\log_2 0 + 1\log_2 1) = 0.000\\
&Gain(D,Z) = 0.811 - (\frac{1}{2}\times 1.000 + \frac{1}{2}\times 0.000) = 0.311
\end{aligned}
$$

故以 Z 划分, 得到两个数据集,$Z = 1 : \{1,4,6,8\}, Z = 0 : \{2,3,5,7\}$ 分别以 X,Y 对 1,4,6,8 进行划分, 得到的子集为 D_1^0, D_1^1.
$\mathrm{Ent}(D_1) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1.000$

$$
\begin{aligned}
X_1:&\mathrm{Ent}(D_1^0) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1.000\\
&\mathrm{Ent}(D_1^1) = \mathrm{Ent}(D_1^0) = 1.000\\
&Gain(D_1,X) = 1.000 - (\frac{1}{2}\times 1.000 + \frac{1}{2}\times 1.000) = 0\\
Y_1:&Gain(D_1,Y) = Gain(D_1,X) = 0.
\end{aligned}
$$

决策树如下:

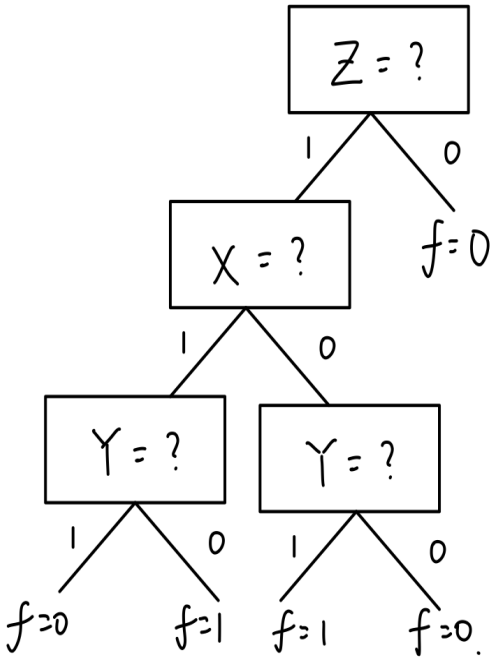

Figure 1: 第 2 题图 1

2. 分别以 X,Y,Z 对 D 进行划分, 得到的子集分别为 D^0, D^1.

$$Gini(D, X) = 2 \times (1 - (\frac{1}{4})^2 - (\frac{3}{4})^2) = 0.375$$
$$Gini(D, Y) = 2 \times (1 - (\frac{1}{4})^2 - (\frac{3}{4})^2) = 0.375$$
$$Gini(D, Z) = 2 \times (1 - (\frac{0}{4})^2 - (\frac{4}{4})^2) = 0.25$$

故以 Z 划分. 然后分别以 X,Y 对 1,4,6,8 进行划分, 得到的子集为 D_1^0, D_1^1, 此时 X 与 Y 的基尼指数相同, 故以 X 划分, 决策树如下:

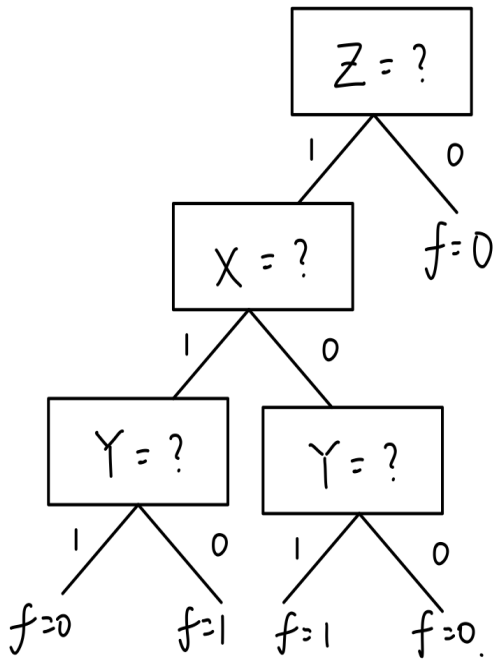

Figure 2: 第 2 题图 2

三. (25 points) **决策树剪枝处理**

教材 4.3 节介绍了决策树剪枝相关内容, 给定包含 5 个样例的人造数据集如表3a所示, 其中“爱运动”、“爱学习”是属性, “成绩高”是标记. 验证集如表3b所示. 使用信息增益为划分准则产生如图3所示的两棵决策树. 请回答以下问题:

(a) 训练集

编号	爱运动	爱学习	成绩高
1	是	是	是
2	否	是	是
3	是	否	否
4	是	否	否
5	否	否	是

(b) 验证集

编号	爱运动	爱学习	成绩高
6	是	是	是
7	否	是	否
8	是	否	否
9	否	否	否

Table 2: 人造数据集

1. 请验证这两棵决策树的产生过程.
2. 对图3的结果基于该验证集进行预剪枝、后剪枝, 给出剪枝后的决策树.

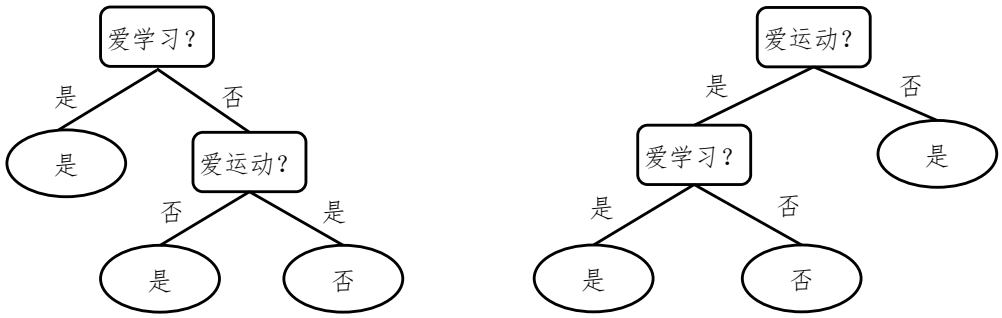

Figure 3: 人造数据决策树结果

3. 比较预剪枝、后剪枝的结果, 每种剪枝方法在训练集、验证集上的准确率分别为多少? 哪种方法拟合能力较强?

解:

1. 对训练集, 设爱运动为 X, 爱学习为 Y, 是为 1, 否为 0. 则有:
   $\mathrm{Ent}(D) = -(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}) = 0.971$
   以爱学习划分时, 有 $Gain(D,Y) = -(\frac{2}{5}\log_2\frac{2}{5}+\frac{3}{5}\log_2\frac{3}{5})+\frac{3}{5}(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3})$
   以爱运动划分时, 有 $Gain(D,X) = -(\frac{2}{5}\log_2\frac{2}{5}+\frac{3}{5}\log_2\frac{3}{5})+\frac{3}{5}(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3})$
   因此其对根节点属性划分有两种选择, 可以发现以这两种划分时得到的数据集是一样的, 所以可以证明这两个决策树产生过程都是合理的.
2. 以下四幅图分别表示左侧决策树预剪枝、后剪枝, 右侧决策树预剪枝、后剪枝:

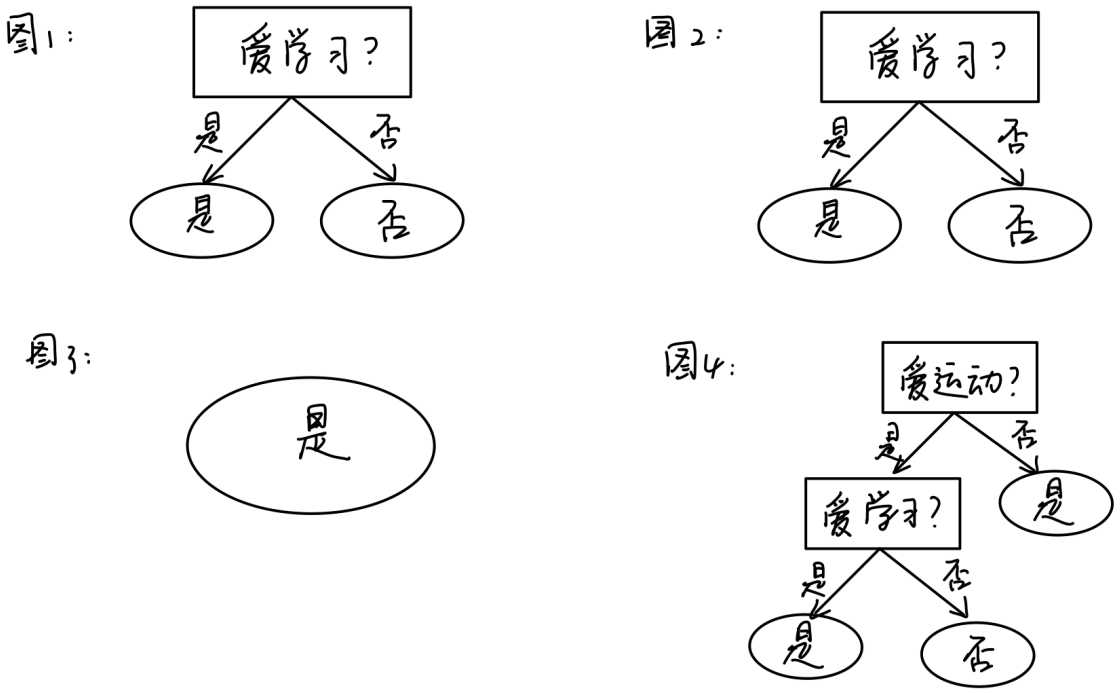

Figure 4: 第 3 题图

3. 决策树 1:

	训练集	测试集
预剪枝	0.8	0.75
后剪枝	0.8	0.75

决策树 2:

	训练集	测试集
预剪枝	0.6	0.25
后剪枝	1.0	0.5

因此可以发现后剪枝拟合能力强于预剪枝.

四. (20 points) **连续与缺失值**

1. 考虑如表 4所示数据集，仅包含一个连续属性，请给出将该属性“数字”作为划分标准时的决策树划分结果。

属性	类别
3	正
4	负
6	负
9	正

Table 4: 连续属性数据集

2. 请阐述决策树如何处理训练时存在缺失值的情况，具体如下：考虑表 1的数据集，如果发生部分缺失，变成如表 5所示数据集（假设 X, Y, Z 只有 0 和 1 两种取值）. 在这种情况下，请考虑如何处理数据中的缺失值，并结合问题 二第 1 小问的答案进行对比，论述方法的特点以及是否有局限性。

X	Y	Z	f
1	0	-	1
-	1	0	0
0	-	0	0
0	1	1	1
-	1	0	0
0	0	-	0
1	-	0	0
1	1	1	0

Table 5: 缺失数据集

3. 请阐述决策树如何处理测试时存在缺失值的情况，具体如下：对于问题 三训练出的决策树，考虑表 6所示的含有缺失值的测试集，输出其标签，并论述方法的特点以及是否有局限性。

编号	爱运动	爱学习	成绩高
6	是	-	
7	-	是	
8	否	-	
9	-	否	

Table 6: 缺失数据集

解：

1. 进行连续值处理，可以得到划分点集合 $T_a = \{3.5, 5, 7.5\}$

属性信息增益计算结果如下:

$$\begin{aligned} Gain(D,a,3.5) &= -(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2})+ \\ &\quad \frac{3}{4}(\frac{1}{3}log_2\frac{1}{3}+\frac{2}{3}log_2\frac{2}{3})] = 0.311 \\ Gain(D,a,5) &= -(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2})+ \\ &\quad 2\times\frac{1}{2}(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2})] = 0 \\ Gain(D,a,7.5) &= -(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2})+ \\ &\quad \frac{3}{4}(\frac{1}{3}log_2\frac{1}{3}+\frac{2}{3}log_2\frac{2}{3})] = 0.311 \end{aligned}$$

故可以选择 3.5 和 7.5 两个划分点, 当选取 3.5 为划分点时得到两个分支:$D^0=\{3\}, D^1=\{4,6,9\}$

选取 7.5 为划分点时得到两个分支:$D^0=\{3,4,6\}, D^1=\{9\}$

2. 计算信息增益可以得到:

$$\begin{aligned} X:\mathrm{Ent}(\tilde{D}) &= -(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}) = 0.918 \\ Gain(D,X) &= \frac{3}{4}\times(0.918-[\frac{3}{6}(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3})+ \\ &\quad \frac{3}{6}(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3})]) = 0 \\ Y:\mathrm{Ent}(\tilde{D}) &= -(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}) = 0.918 \\ Gain(D,Y) &= \frac{3}{4}\times(0.918-[\frac{2}{6}(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}) \\ &\quad +\frac{4}{6}(-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4})]) = 0.0329 \\ Z:\mathrm{Ent}(\tilde{D}) &= -(\frac{1}{6}\log_2\frac{1}{6}+\frac{5}{6}\log_2\frac{5}{6}) = 0.650 \\ Gain(D,Z) &= \frac{3}{4}\times(0.650-[\frac{4}{6}(-0-0)+\frac{2}{6}(\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2})]) = 0.238 \end{aligned}$$

所以第一个划分属性为 Z, 得到两个分支, 其中编号为 2,3,5,7 的进入"Z=0" 分支, 编号为 4,8 的进入"Z=1" 分支, 编号为 1,6 的同

事进入两个分支, 在两个分支的权重分别为 $\frac{2}{3}, \frac{1}{3}$
对于分支"Z=0" 计算信息增益有:

$$Gain(D^0, X) = \frac{2}{3}(\mathrm{Ent}(\tilde{D}^0) + \frac{1}{2}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}))$$
$$Gain(D^0, Y) = \frac{2}{3}(\mathrm{Ent}(\tilde{D}^0) + \frac{2}{5}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}))$$

所以由于 $Gain(D^0, Y)$ 大, 即此时第二个划分属性为 Y, 同理对另一个分支, 此时第二个划分属性为 X. 决策树如下:

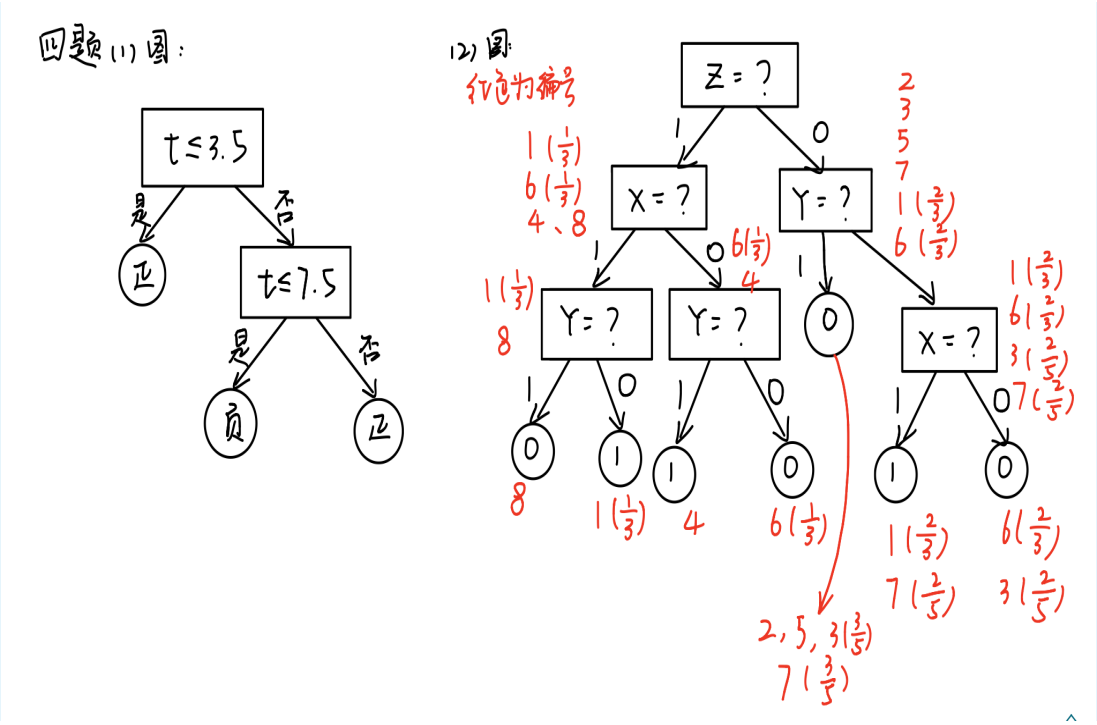

Figure 5: 第 4 题图

由此可见, 大部分的样本依旧分类正确, 仅有样本 1,3,6,7 号以不同的概率进入不同节点, 其中 1,3,6 号样本所有概率都取值为 0, 仅有 7 号样本有 $\frac{2}{5}$ 的概率分类错误. 所以该方法具有较高泛化能力, 表现较好.

3. 当测试时出现缺失值时, 就同时探查所有的可能分支, 计算每个类别的概率, 取概率最大的类别赋值给该样本. 对于第三题的验证集若部分属性缺失, 则输出如下:

	爱运动	爱学习	成绩高
6	是	-	一半是一半否
7	-	否	是
8	否	-	是
9	-	否	一半是一半否

	爱运动	爱学习	成绩高
6	是	-	一半是一半否
7	-	否	是
8	否	-	是
9	-	否	一半是一半否

五. (20 points) **多变量决策树**

考虑如下包含 10 个样本的数据集, 每一列表示一个样本, 每个样本具有二个属性, 即 $\boldsymbol{x}_i = (x_{i1};\ x_{i2})$.

编号	1	2	3	4	5	6	7	8	9	10
A_1	24	53	23	25	32	52	22	43	52	48
A_2	40	52	25	77	48	110	38	44	27	65
标记	1	0	0	1	1	1	1	0	0	1

1. 计算根结点的熵;
2. 构建分类决策树, 描述分类规则和分类误差;
3. 根据 $\alpha x_1 + \beta x_2 - 1$, 构建多变量决策树, 描述树的深度以及 α 和 β 的值.

解:

1.
$$\mathrm{Ent}(\mathit{Root}) = -(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}) = 0.971.$$

2. 对 A_1, A_2 两个特征进行离散处理, 并根据基尼系数作为划分属性的评价指标. 根据其基尼系数值得到最小的划分标准 $(A_2)32.5$ 作为第一轮决策树划分标准, 分类之后, 第二轮中最小的划分标准

(A_1)52.5 作为第二轮决策树划分标准, 依次类推, 每次删除成功分类的点并递归进行该过程, 直到可以将所有的特征都分类完成, 决策树如下:

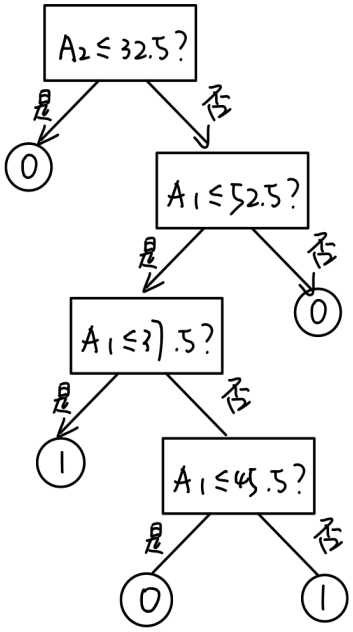

Figure 6: 第 5 题图

3.