姓名：杜兴豪
学号：201300096

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 $\left\{a^1, a^2, \cdots, a^V\right\}$, 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.

一定存在。由于样本中属性个数为有限个，记属性个数为 d。若不包含冲突样本，则这些属性的任意组合最多只有 $\prod_d m_i$ 种，其中 m_i 为每种属性的取值数目。则无穷多个样本中剩下的均为重复样本，可以忽略，下证该决策树为有限的。

由题干信息可知道，对于不含冲突样本的训练集，一定存在一棵和训练集一致的决策树。考虑最坏情况，对每种样本都独立判断（有一个

独立的叶子节点)，一共仅 $\prod_d m_i$ 个叶子节点，因此树为有限的。
2.
由信息熵的定义有:

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

因为 $0 \le p_k \le 1$, 我们令:

$$f(x) = x^x = e^{x\ln x}, 0 \le x \le 1$$

讨论 $f(x)$ 取值情况如下:

$$f'(x) = \frac{\partial f(x)}{\partial x} = x^x(\ln x + 1)$$

由 $f(x) = e^{x\ln x} > 0$, 知

$$\begin{cases} f'(x) < 0 & 0 < x < e^{-1} \\ f'(x) > 0 & x > e^{-1} \end{cases}$$

因此

$$\max f(x) = \max\{f(0), f(1)\} = 1$$

我们有:

$$\mathrm{Ent}(D) \ge -\log_2 \prod_{k=1}^{|\mathcal{Y}|} 1 = 0$$

又 p_k 为第 k 类样本在总体 D 中所占的比例，因此有:

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

故取等号时，$p_k = |\mathcal{Y}| = 1$，即一共仅一类时取等号, 下证:

$$\mathrm{Ent}(D) \le \log_2 |\mathcal{Y}|$$

即解凸优化问题:

$$\max -\log_2 \prod_{k=1}^{|\mathcal{Y}|} {p_k}^{p_k}$$

$$\textbf{s.t.} \sum_{k=1}^{|\mathcal{Y}|} p_k = 1$$

构造 **Lagrange** 对偶问题：

$$\max \ L(p_1,\ldots,p_{|\mathcal{Y}|},\lambda) = -\log_2 \prod_{k=1}^{|\mathcal{Y}|} {p_k}^{p_k} + \lambda\left(\sum_{k=1}^{|\mathcal{Y}|} p_k - 1\right)$$

分别求偏导有：

$$\frac{\partial L(p_1,\ldots,p_{|\mathcal{Y}|},\lambda)}{\partial p_k} = -\log_2 p_k - 1 + \lambda$$

$$\frac{\partial L(p_1,\ldots,p_{|\mathcal{Y}|},\lambda)}{\partial \lambda} = \sum_{k=1}^{|\mathcal{Y}|} p_k - 1$$

令上式全为 0，有：

$$p_k = 2^{\lambda-1}$$

$$\sum_{k=1}^{|\mathcal{Y}|} p_k = |\mathcal{Y}|2^{\lambda-1} = 1$$

得到对偶问题的解（同时也是原优化问题的解）为：

$$p_1 = p_2 = \cdots = p_{|\mathcal{Y}|} = \frac{1}{|\mathcal{Y}|}$$

此时得到信息熵的最大值：

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

即

$$\mathrm{Ent}(D) \le \log_2|\mathcal{Y}|$$

由优化问题同样可以知道取等条件为：

$$p_1 = p_2 = \cdots = p_{|\mathcal{Y}|} = \frac{1}{|\mathcal{Y}|}$$

3.

假设全集 D 中，第 k 类样本的元素个数为 m_k, 则在 D^v 中第 k 类样本所占比例为：

$$p_k^v = \frac{m_k}{|D^v|}$$

则其信息熵可以表示为:

$$\operatorname{Ent}(D^v) = -\sum_{k=1}^{|\mathcal{Y}|} \frac{m_k}{|D^v|} \log_2 \frac{m_k}{|D^v|}$$

则有

$$\sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Ent}(D^v) = -\sum_{v=1}^{V} \sum_{k=1}^{|\mathcal{Y}|} \frac{m_k}{|D|} \log_2 \frac{m_k}{|D^v|}$$

而

$$\operatorname{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k = -\sum_{v=1}^{V} \sum_{k=1}^{|\mathcal{Y}|} \frac{m_k}{|D|} \log_2 \frac{m_k}{|D|}$$

即证明

$$-\log_2 \frac{m_k}{|D^v|} \leq -\log_2 \frac{m_k}{|D|}$$
$$\iff \frac{1}{|D^v|} \geq \frac{1}{|D|}$$

由于 $|D^v|$ 为包含 D 中所有在属性 a 上取值为 a^v 的样本，因此一定有

$$|D| \geq |D^v|$$

上式成立，证明完毕。

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.

令全集为 D，$|D| = 8$，其中正例有 2 个，反例 6 个，则当前根节点的信息熵为:

$$\mathrm{Ent}(D) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.8113$$

以 X 属性划分，有两个取值，其中当 $X = 1$ 时有 4 个样例，正例 1 个，反例 3 个; 当 $X = 0$ 时共 4 个样例，正例 1 个，计算信息熵为:

$$\mathrm{Ent}(D^{X_1}) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.8113$$
$$\mathrm{Ent}(D^{X_0}) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.8113$$

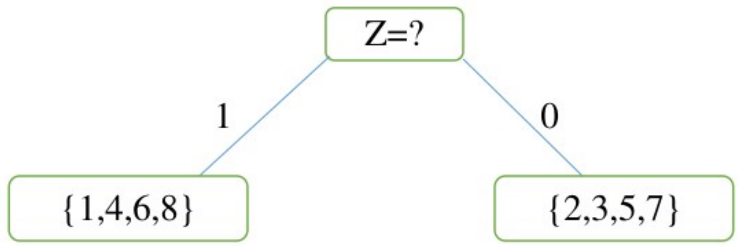
因此可计算属性 X 的信息增益为:

$$\begin{aligned}\mathrm{Gain}(D, X) &= \mathrm{Ent}(D) - \sum_{v=0}^{1}\frac{|D^{X_v}|}{|D|}\mathrm{Ent}(D^{X_v}) \\ &= 0.8113 - (\frac{1}{2}\times 0.8113 + \frac{1}{2}\times 0.8113) \\ &= 0\end{aligned}$$

类似可计算属性 Y，Z 的信息增益为:

$$\begin{aligned}\mathrm{Gain}(D, Y) &= 0 \\ \mathrm{Gain}(D, Z) &= 0.3113\end{aligned}$$

属性 Z 的信息增益最大，则选用 Z 对根节点进行划分，得到划分情况如图所示:



其中各分支的信息熵为:

$$\mathrm{Ent}(D^{Z_1}) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

$$\mathrm{Ent}(D^{Z_0}) = -(0 + \log_2 1) = 0$$

在 $Z = 1$ 分支下，计算 X,Y 的信息增益如下:

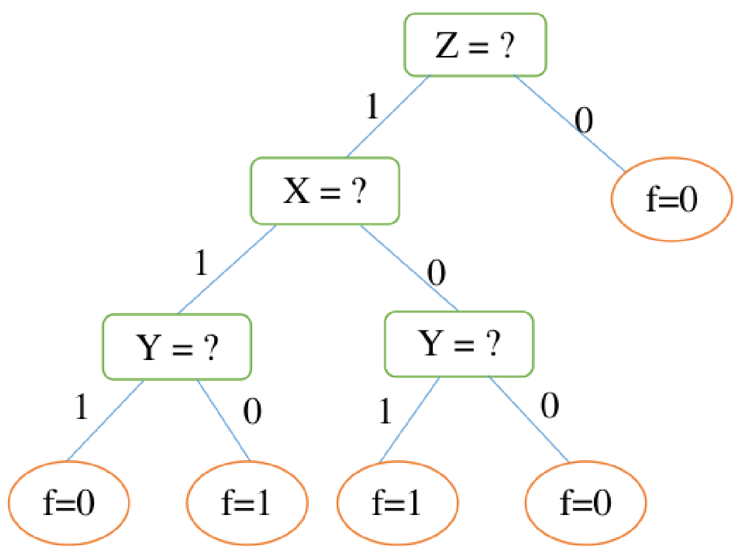
$$\mathrm{Gain}(D^{Z_1}, X) = 1$$

$$\mathrm{Gain}(D^{Z_1}, Y) = 1$$

信息增益相同，以字典序选取 X 作为当前划分属性对 $Z = 1$ 分支进行划分。

在 $Z = 0$ 分支下，观察到所有样例取值 f 均为 0，因此不再继续划分。

最后在 X 分支下以 Y 划分，得到最终结果如下:



2.

$$\mathrm{Gini}(D) = 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2 = 1 - \left((\frac{1}{4})^2 + (\frac{3}{4})^2 \right) = 0.375$$

以 X 属性划分，其取值为 $X = 1$ 和 $X = 0$ 的基尼值为:

$$\mathrm{Gini}(D^{X_1}) = 1 - \left((\frac{1}{4})^2 + (\frac{3}{4})^2 \right) = 0.375$$

$$\mathrm{Gini}(D^{X_0}) = 1 - \left((\frac{1}{4})^2 + (\frac{3}{4})^2 \right) = 0.375$$

则 X 属性的基尼指数为:

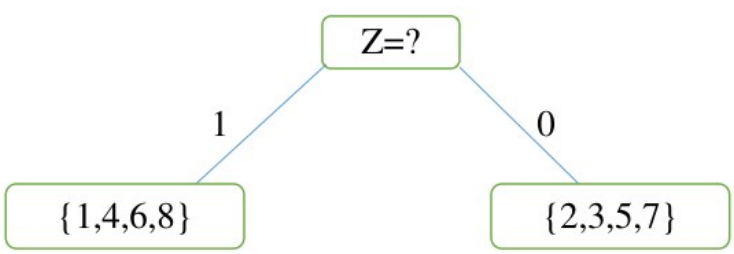
$$\begin{aligned}
\mathrm{Gini_index}(D, X) &= \sum_{v=1}^{V} \frac{|D^{X_v}|}{|D|} \mathrm{Gini}(D^{X_v}) \\
&= \frac{1}{2} \times 0.375 + \frac{1}{2} \times 0.375 \\
&= 0.375
\end{aligned}$$

类似可计算 Y，Z 属性的基尼指数为:

$$\mathrm{Gini\_index}(D, Y) = \frac{1}{2} \times 0.375 + \frac{1}{2} \times 0.375 = 0.375$$

$$\mathrm{Gini\_index}(D, Z) = \frac{1}{2} \times 0 + \frac{1}{2} \times 0.5 = 0.25$$

属性 Z 的基尼指数最小，则选用 Z 对根节点进行划分，得到划分情况如图:



其中各分支的基尼值为:

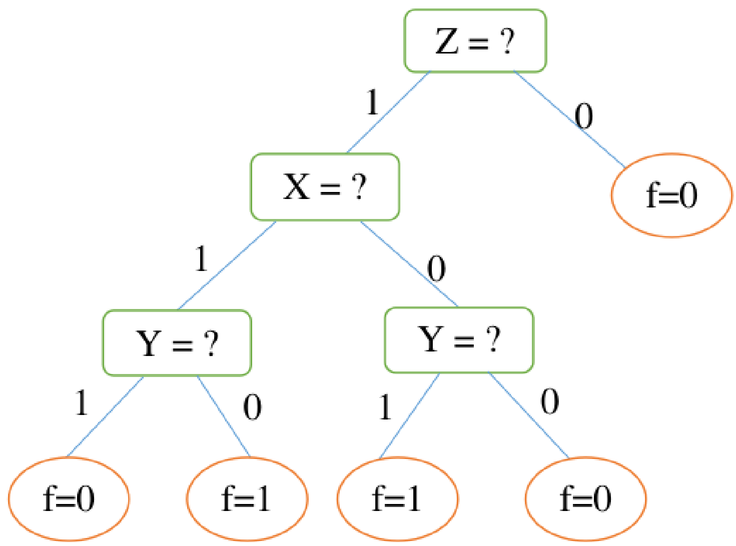
$$\mathrm{Gini}(D^{Z_1}) = 0.25$$

$$\mathrm{Gini}(D^{Z_0}) = 0$$

观察到 $Z=0$ 时的基尼值已经为最小，是最纯净的情况，因此不用再细分。
当 $Z=1$ 时，计算 X，Y 的基尼指数如下:

$$\text{Gini\_index}(D^{Z_1}, X) = \frac{1}{2} \times 0.5 + \frac{1}{2} \times 0.5 = 0.5$$
$$\text{Gini\_index}(D^{Z_1}, Y) = \frac{1}{2} \times 0.5 + \frac{1}{2} \times 0.5 = 0.5$$

基尼指数相同，则以字典序选取 X 来划分 $Z=1$ 的情况。
最后在 X 分支下以 Y 进行划分，得到最终结果如图:



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1所示的两棵决策树. 请回答以下问题:

1. 请验证这两棵决策树的产生过程.
2. 对图1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.

(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

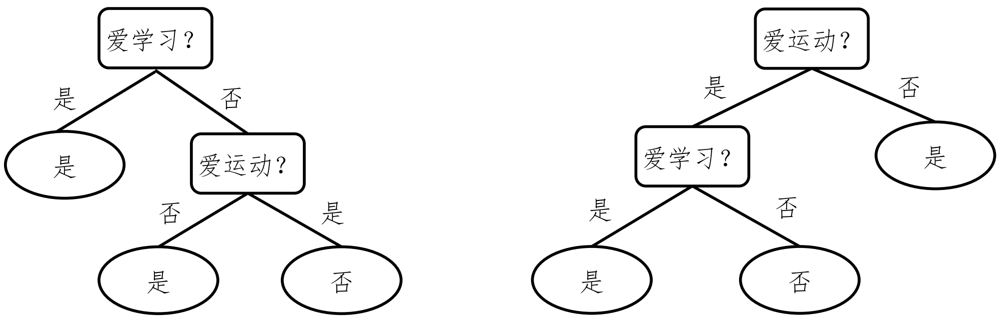


Figure 1: 人造数据决策树结果

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.

令全集为 D, 则其信息熵可以表示为:

$$\mathrm{Ent}(D) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.97095$$

若选择爱学习来划分原数据集，以属性值“是”“否”得到其信息熵分别为:

$$\begin{aligned}
\mathrm{Ent}(D^1) &= -\left(0 + \log_2 1\right) = 0 \\
\mathrm{Ent}(D^2) &= -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.91830
\end{aligned}$$

则其信息增益为:

$$\begin{aligned}
\mathrm{Gain}(D, 爱学习) &= \mathrm{Ent}(D) - \frac{2}{5}\mathrm{Ent}(D^1) - \frac{3}{5}\mathrm{Ent}(D^2) \\
&= 0.41997
\end{aligned}$$

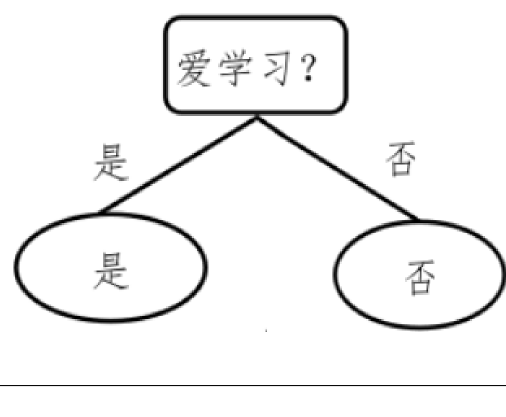
同理可以计算出选择爱运动划分数据集的信息增益为：

$$\mathrm{Gain}(D, \text{爱运动}) = 0.41997$$

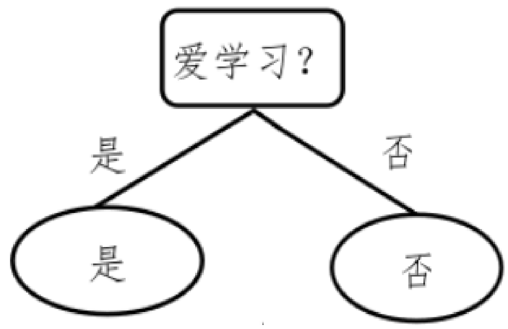
两者的信息增益相同，因此划分根节点采用哪个都合理。
若选用爱学习来划分根节点
由

$$\mathrm{Ent}(D^1) = 0$$

知道当爱学习为“是”的时候，样例绝对有序，因此不再继续展开。
当爱学习为“否”的时候，以爱运动划分数据集，得到图一所示结果为合理。
若采用爱运动来划分根节点
同样能看出当爱运动为“否”的时候，样例中所有的标签均为“是”，因此不再展开。
当爱运动为“是”的时候，以爱学习划分，得到图二所示结果也合理。
2.
预剪枝：
若不以爱学习进行划分，将其标记为训练集中出现最多的“是”，在验证集的精度为 25%，
而使用“爱学习”进行划分之后，“是”被判断为“是”（“是”爱学习被判断为“是”成绩好），“否”被判断为“否”，其验证集精度为 75%>25%，因此需要使用“爱学习”划分根节点。
若当爱学习为“否”的时候，不使用“爱运动”划分数据集，而将其标记为最多的“否”，则其验证集精度为 75%
而使用“爱运动”进行划分之后，验证集精度为 50%<75%，因此不需要继续以“爱运动”划分数据。
最终得到预剪枝后的图像如下：



后剪枝:
易知完整决策树的验证集精度为 50%
若将“爱运动”领衔的分支剪除，将其替换为训练集爱学习下“否”中标记最多的“否”，此时验证集精度提升为 75%>50%，因此需要剪枝。
若将“爱学习”领衔的分支剪除，将其替换为训练集中标记最多的“是”，此时验证集精度变为 25%<75%，不剪枝。
最终后剪枝图像如下:



3.
两种方法得到的决策树完全相同，因此具有相同的准确率和拟合能力。训练集上准确率为 80%, 测试集上准确率为 75%

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

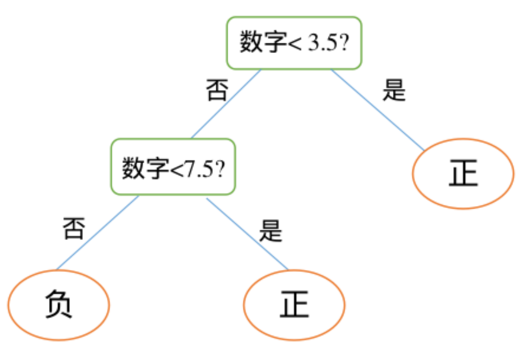
3. 请阐述决策树如何处理测试时存在缺失值的情况，具体如下：对于问题 三训练出的决策树，考虑表 6所示的含有缺失值的测试集，输出其标签，并论述方法的特点以及是否有局限性。

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.



2.

假设 $\hat{D}$ 为在 X 属性上没有缺失的样本子集，则以这个集合为训练集计算其信息熵如下：

$$\mathrm{Ent}(\hat{D}) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.91830$$

以 X 属性划分，设 $\hat{D}^1, \hat{D}^0$ 为 X 取值为 1，0 时的子集，则其信息熵:

$$\mathrm{Ent}(\hat{D}^1) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.91830$$

$$\mathrm{Ent}(\hat{D}^0) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.91830$$

因此得到信息增益为:

$$\begin{aligned}
\mathrm{Gain}(D,X) &= \frac{|\hat{D}|}{|D|}\mathrm{Gain}(\hat{D},X) \\
&= \frac{|\hat{D}|}{|D|}\left(\mathrm{Ent}(\hat{D}) - \sum_{v=1}^{V}\frac{|\hat{D}^v|}{|\hat{D}|}\mathrm{Ent}(\hat{D}^v)\right) \\
&= \frac{3}{4}\times\left(0.91830 - \left(\frac{1}{2}\times 0.91830 + \frac{1}{2}\times 0.91830\right)\right) \\
&= 0
\end{aligned}$$

同理可计算 Y，Z 的信息增益为:

$$\mathrm{Gain}(D,Y) = \frac{3}{4}\times\mathrm{Gain}(\hat{D},Y) = 0.03308$$

$$\mathrm{Gain}(D,Z) = \frac{3}{4}\times\mathrm{Gain}(\hat{D},Z) = 0.43873$$

因此选用信息增益最大的 Z 进行划分。在原数据集中 Z 缺失的 1、6 号元素同时进入 Z 的两个分支，其权重在 $Z=1$ 分支中为 1/3，在 $Z=0$ 分支中为 2/3

现在对 $Z=0$ 分支继续划分, 该分支中取值为 1 的权重为:

$$p_1 = \frac{\frac{2}{3}}{4 + 2\times\frac{2}{3}} = \frac{1}{8}$$

因此其信息熵为:

$$\mathrm{Ent}(D^{Z_0}) = -\left(\frac{1}{8}\log_2\frac{1}{8} + \frac{7}{8}\log_2\frac{7}{8}\right) = 0.54356$$

使用 X 属性来划分数据集 D^{Z_0}, 我们有: 无缺训练集 $\hat{D}^{Z_0}$ 为 {1,3,6,7}，分别计算 X 取值为 0，1 时的信息熵如下:

$$\mathrm{Ent}(\hat{D}^{Z_0X_0}) = -(0 + \log_2 1) = 0$$

$$\mathrm{Ent}(\hat{D}^{Z_0X_1}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.97095$$

因此可以计算出其信息增益为:

$$\mathrm{Gain}(D^{Z_0}, X) = \frac{10}{16}\times\left(0.54356 - \left(\frac{1}{2}\times 0 + \frac{1}{2}\times 0.97095\right)\right) = 0.03630$$

同理可以计算出 Y 的信息增益为:

$$\mathrm{Gain}(D^{Z_0}, Y) = \frac{10}{16}\times\left(0.54356 - \left(\frac{3}{5}\times 0 + \frac{2}{5}\times 1\right)\right) = 0.08972$$

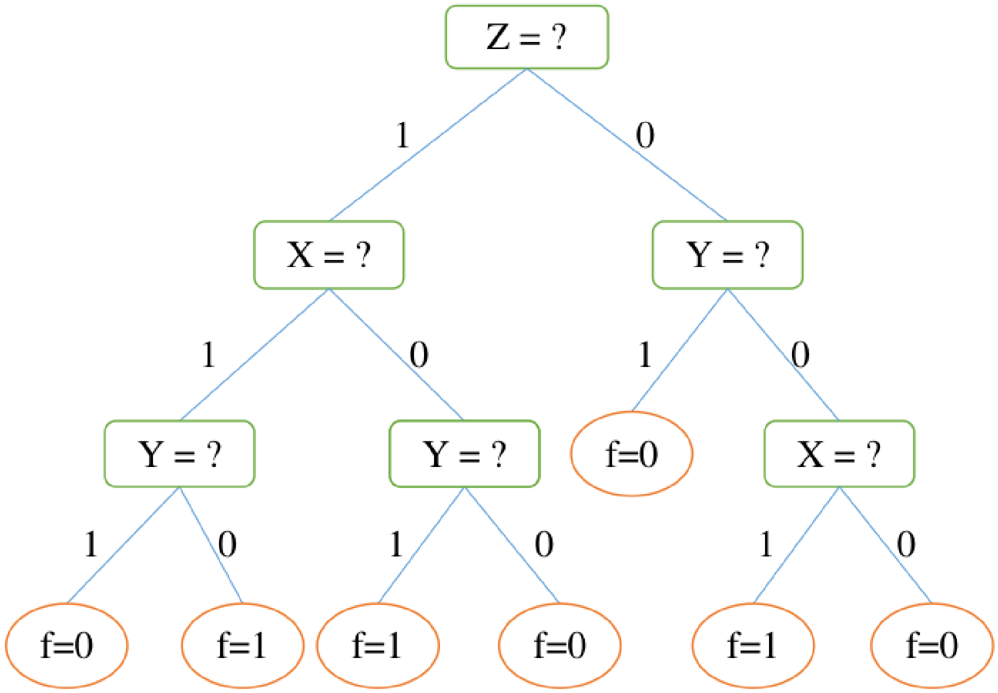
因此选择信息增益较大的 Y 来划分 $Z=0$ 分支。原先 Y 缺失的样本同时进入两个分支，在 $Y=1$ 分支中权重为原来的 2/3，在 $Y=0$ 分支中为原来的 1/3。
观察到 $Y=1$ 分支中的所有结果均为 0，因此不再展开。
$Y=0$ 分支再使用 X 做进一步划分，不多赘述。
然后对 $Z=1$ 分支讨论划分方法。

$$\mathrm{Ent}(D^{Z_1}) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

X，Y 在此样本中均无缺失，计算其信息增益分别为:

$$\mathrm{Gain}(D^{Z_1}, X) = 0.18872$$
$$\mathrm{Gain}(D^{Z_1}, Y) = 0$$

因此选用 X 划分 $Z=1$ 分支。再使用 Y 划分 X 的分支，不多赘述。
得到最终结果如下图:



处理方法使用：

在无缺数据集下处理数据，以频率猜测缺失样本的属性值，将有缺样本以猜测的频率分发权重，放入所有的分支中，从而可以利用好有缺样本的无缺部分。

对比结果：

当数据缺失时，在 $Z=0$ 处并没有直接判断为 $f=0$，而是继续用属性 Y 继续划分数据集，其余部分保持一致。

特点：

不用将缺失数据集丢弃，极大地提高了数据的利用率，并且学习出的决策树在大体上与无缺数据集保持一致，性能良好。

对方法局限性的思考：

当同一属性的样本缺失太多条时，这样的预测将变得难以保持稳定，猜测补全的结果很有可能和真实数据集发生较大的偏差。

3.

以频率估计概率，将原测试集缺失样本以其出现概率分发权重，创造更大的测试集，每条样本具有不同的权重。计算成功率时以加权和计算。得到的新测试集以及对应标签如下：

编号	爱运动	爱学习	成绩高	权重
6	是	是	是	0.5
	是	否	否	0.5
7	是	是	是	0.5
	否	是	是	0.5
8	否	是	是	0.5
	否	否	否	0.5
9	是	否	否	0.5
	否	否	否	0.5

特点为可以利用到缺失样本，利用率高。局限性是当测试集分布和真实情况很不同时，预测结果有可能发生较大偏差。

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:

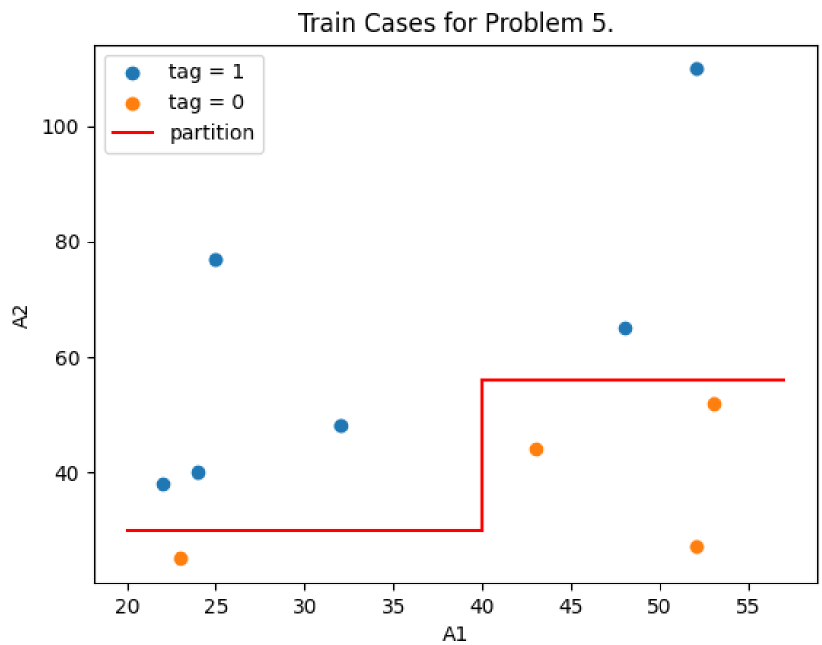
1.

令训练集为 D，则根节点的熵为：

$$
\begin{aligned}
\operatorname{Ent}(D) &= -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k \\
&= -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) \\
&= 0.97095
\end{aligned}
$$

2.

根据训练样本的图像，以及考虑的划分候选点集合可以得到以下分类边界：



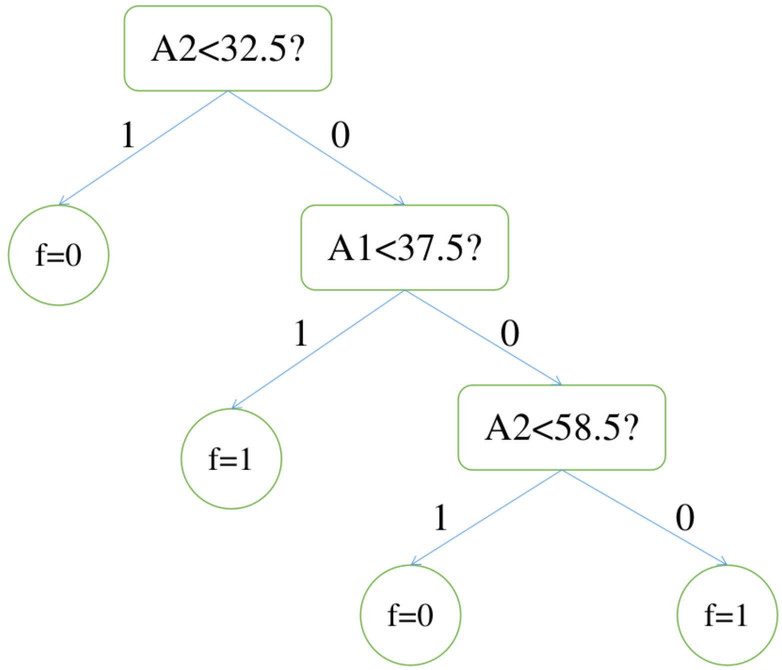
得到分类规则为：

先以 $A_2 < 32.5$ 为界，满足的为标记为 0 的点，否则进入分支；

再以 $A_1 < 37.5$ 为界限，满足的为标记为 1 的点，否则再进入分支；

最后以 $A_2 < 58.5$ 为界限，满足的标记为 0，否则标记为 1。

可以得到分类决策树为：



由图可以知道分类误差为 0%

3.

由图可知该分类可以通过一条直线进行划分。尝试线性分类方法。令

$$X = \left(A_1^\top, A_2^\top, 1\right) \in \mathbb{R}^{10\times 3}$$

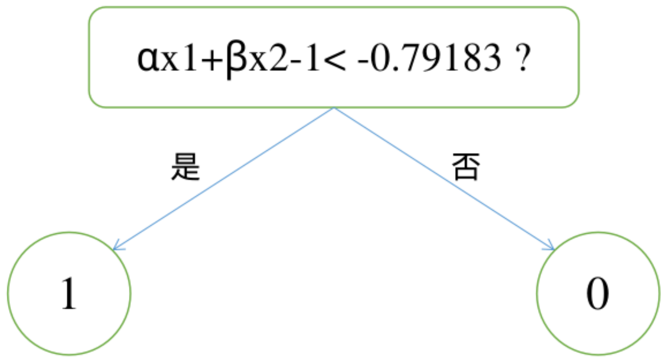
参数 α, β 合并为

$$w = \begin{pmatrix} \alpha \\ \beta \\ b \end{pmatrix} \in \mathbb{R}^{3\times 1}$$

学得参数

$$w^* = (X^\top X)^{-1} X^\top y = \begin{pmatrix} -0.02264422 \\ 0.01454232 \\ 0.68196814 \end{pmatrix}$$

由于题目要求为 $\alpha x_1 + \beta x_2 - 1$, 因此需要在原分类器输出结果的判别条件上除以 -0.68196814，得到如下的决策树:



经验证，正确率为 100%，无需继续划分。
因此树深度为 2，$\alpha = 0.03320422, \beta = -0.02132404$