

姓名：方盛俊

学号：201300035

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 $\text{Ent}(D)$ 定义如下

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

请证明信息熵的上下界为

$$0 \leq \text{Ent}(D) \leq \log_2 |\mathcal{Y}| \quad (2)$$

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 $\{a^1, a^2, \dots, a^V\}$ ，请考教材 4.2.1 节相关符号的定义证明：

$$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v) \geq 0 \quad (3)$$

即信息增益非负。

解：

1. 对于属性值均为有限取值的离散值的训练集来说，存在与训练集一致的深度有限决策树。因为对于有限取值的离散值，每一层都会减少一种待选的属性，所以深度必然有限。

对于属性值为有无限中取值时，例如有一种属性是连续值时，不一定存在与训练集一致的深度有限决策树。

例如我们构造一个只有单个属性和单个标记的训练集 $D = \{(x_i, y_i)\}$ ，其中

$$y = D(x) = \begin{cases} 1, & x \in \mathbb{Q} \\ 0, & x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

而 $x \in [0, 1]$, 即定义域为 $[0, 1]$ 的 Dirichlet 函数.

我们这样取出我们的无穷多个训练集样本: 从 $x_1 = 0$ 开始取, 此时 $i = 1$, 不断取出比 x_{2i-1} 大且相邻的有理数 x_{2i+1} , 其中 $i = 1, 2, \dots$, 并且在两个相邻的有理数 x_{2i-1} 和 x_{2i+1} 之间任取一个无理数 x_{2i} , 并使得 $x_{2i-1} < x_{2i} < x_{2i+1}$. 而它们对应的 $y_i = D(x_i)$.

这样, 我们就构造出了一个标记为 1 和 0 交替出现的无穷个样本的训练集.

对于一个这样的训练集, 我们使用处理连续值属性的决策树算法, 我们在训练集中不断地对属性 x 进行划分, 无论划分区间多小, 也不可能得到一个标记 y 完全为 1 或 0 的子集, 因此决策树算法会不断继续下去, 生成深度无限的决策树.

2. 因为 $0 \leq p_k \leq 1$, 则有 $p_k \log_2 p_k \leq 0$, 因此

$$-\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k \geq 0$$

令 $-\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k = 0$, 则有 $p_k \log_2 p_k = 0$

即有当每一个 $p_k = 0$ 或 $p_k = 1$ 时等号成立.

对于原式

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

显然在 $0 \leq p_k \leq 1$ 时是上凸函数, 因此是一个凸优化问题.

对应拉格朗日函数为

$$L(\mathbf{p}, \lambda) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k + \lambda \left(\sum_{k=1}^{|\mathcal{Y}|} p_k - 1 \right)$$

将其转化为矩阵形式则有

$$L(\mathbf{p}, \lambda) = -\mathbf{p}^T \log_2 \mathbf{p} + \lambda(\mathbf{1}^T \mathbf{p} - 1)$$

对其求微分得

$$\begin{aligned} dL(\mathbf{p}, \lambda) &= \text{tr}(-\mathbf{p}^T (d \log_2 \mathbf{p}) - (d\mathbf{p})^T \log_2 \mathbf{p} + \lambda \mathbf{1}^T d\mathbf{p}) \\ &= -\text{tr}(\mathbf{p}^T (\frac{1}{\ln 2} \ln' \mathbf{p} \odot d\mathbf{p})) - \text{tr}((d\mathbf{p})^T \log_2 \mathbf{p}) + \lambda \text{tr}(\mathbf{1}^T d\mathbf{p}) \\ &= -\frac{1}{\ln 2} \text{tr}(\mathbf{p}^T (\ln' \mathbf{p} \odot d\mathbf{p})) - \frac{1}{\ln 2} \text{tr}((d\mathbf{p})^T \ln \mathbf{p}) + \frac{1}{\ln 2} \text{tr}(\lambda \ln 2 \mathbf{1}^T d\mathbf{p}) \\ &= -\frac{1}{\ln 2} \text{tr}((\mathbf{p} \odot \ln' \mathbf{p})^T d\mathbf{p}) - \frac{1}{\ln 2} \text{tr}(\ln \mathbf{p}^T d\mathbf{p}) - \frac{1}{\ln 2} \text{tr}(-\lambda \ln 2 \mathbf{1}^T d\mathbf{p}) \\ &= -\frac{1}{\ln 2} \text{tr}(\mathbf{1}^T d\mathbf{p}) - \frac{1}{\ln 2} \text{tr}(\ln \mathbf{p}^T d\mathbf{p}) - \frac{1}{\ln 2} \text{tr}(-\lambda \ln 2 \mathbf{1}^T d\mathbf{p}) \\ &= \text{tr}(-\frac{1}{\ln 2} ((1 - \lambda \ln 2) \mathbf{1} + \ln \mathbf{p})^T d\mathbf{p}) \end{aligned}$$

因此有

$$\frac{\partial L(\mathbf{p}, \lambda)}{\partial \mathbf{p}} = -\frac{1}{\ln 2} ((1 - \lambda \ln 2) \mathbf{1} + \ln \mathbf{p})$$

令 $\frac{\partial L(\mathbf{p}, \lambda)}{\partial \mathbf{p}} = 0$ 即可知 \mathbf{p} 各分量相同, 即 $p_i = p_j, i \neq j$

再由我们知道 $\sum_{k=1}^{|\mathcal{Y}|} p_k = 1$ 则有

$$p_k = \frac{1}{|\mathcal{Y}|}$$

即当 $p_k = \frac{1}{|\mathcal{Y}|}$ 时我们取得最大值

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

3.

$$\begin{aligned}
 & \text{Gain}(D, a) \\
 &= \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v) \\
 &= - \sum_{v=1}^V \frac{|D^v|}{|D|} \sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k + \sum_{v=1}^V \frac{|D^v|}{|D|} \sum_{k=1}^{|\mathcal{Y}|} p_k^v \log_2 p_k^v \\
 &= - \sum_{k=1}^{|\mathcal{Y}|} \sum_{v=1}^V \frac{|D^v|}{|D|} \frac{|D_k|}{|D|} \log_2 \frac{|D_k|}{|D|} + \sum_{k=1}^{|\mathcal{Y}|} \sum_{v=1}^V \frac{|D^v|}{|D|} \frac{|D_k^v|}{|D^v|} \log_2 \frac{|D_k^v|}{|D^v|} \\
 &= - \sum_{k=1}^{|\mathcal{Y}|} \sum_{v=1}^V \frac{|D_k^v|}{|D|} \frac{|D|}{|D_k^v|} \frac{|D^v|}{|D|} \frac{|D_k|}{|D|} \log_2 \frac{|D_k|}{|D|} + \sum_{k=1}^{|\mathcal{Y}|} \sum_{v=1}^V \frac{|D_k^v|}{|D|} \log_2 \frac{|D_k^v|}{|D^v|} \\
 &= - \sum_{k=1}^{|\mathcal{Y}|} \sum_{v=1}^V \frac{|D_k^v|}{|D|} \log_2 \frac{|D_k|}{|D|} \cdot 2^{\frac{|D^v|}{|D_k^v|} \frac{|D_k|}{|D|}} + \sum_{k=1}^{|\mathcal{Y}|} \sum_{v=1}^V \frac{|D_k^v|}{|D|} \log_2 \frac{|D_k^v|}{|D^v|} \\
 &= - \sum_{k=1}^{|\mathcal{Y}|} \sum_{v=1}^V \frac{|D_k^v|}{|D|} \log_2 \frac{|D^v|}{|D_k^v|} \frac{|D_k|}{|D|} \cdot 2^{\frac{|D^v|}{|D_k^v|} \frac{|D_k|}{|D|}} \\
 &\geq - \log_2 \sum_{k=1}^{|\mathcal{Y}|} \sum_{v=1}^V \frac{|D_k^v|}{|D|} \frac{|D^v|}{|D_k^v|} \frac{|D_k|}{|D|} \cdot 2^{\frac{|D^v|}{|D_k^v|} \frac{|D_k|}{|D|}} \\
 &= - \frac{|D^v|}{|D_k^v|} \frac{|D_k|}{|D|} \log_2 \sum_{v=1}^V \frac{|D^v|}{|D|} \sum_{k=1}^{|\mathcal{Y}|} \frac{|D_k|}{|D|} \\
 &= - \frac{|D^v|}{|D_k^v|} \frac{|D_k|}{|D|} \log_2 1 \cdot 1 \\
 &= 0
 \end{aligned}$$

其中不等号使用了 Jensen 不等式, 即 $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$, 其中 f 是凸函数.

因此信息增益 $\text{Gain}(D, a)$ 非负.

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

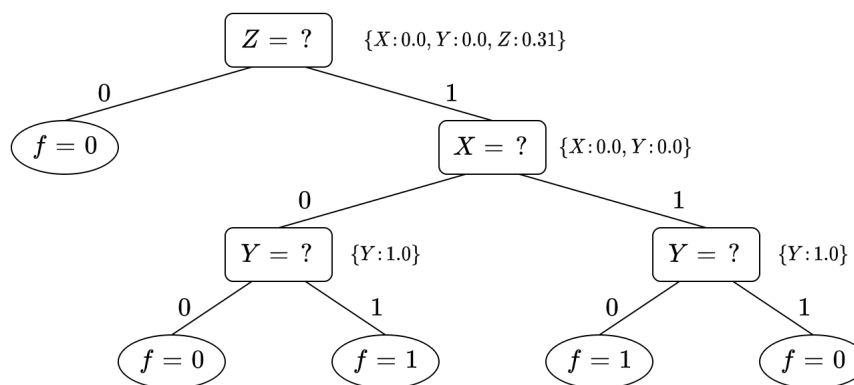
1. 请使用信息增益作为划分准则画出决策树的生成过程. 当两个属性信息增益相同时, 依据字母顺序选择属性.
2. 请使用基尼指数作为划分准则画出决策树的生成过程, 当两个属性基尼指数相同时, 依据字母顺序选择属性.

解:

1. 通过写了一个 Python 程序计算得出:

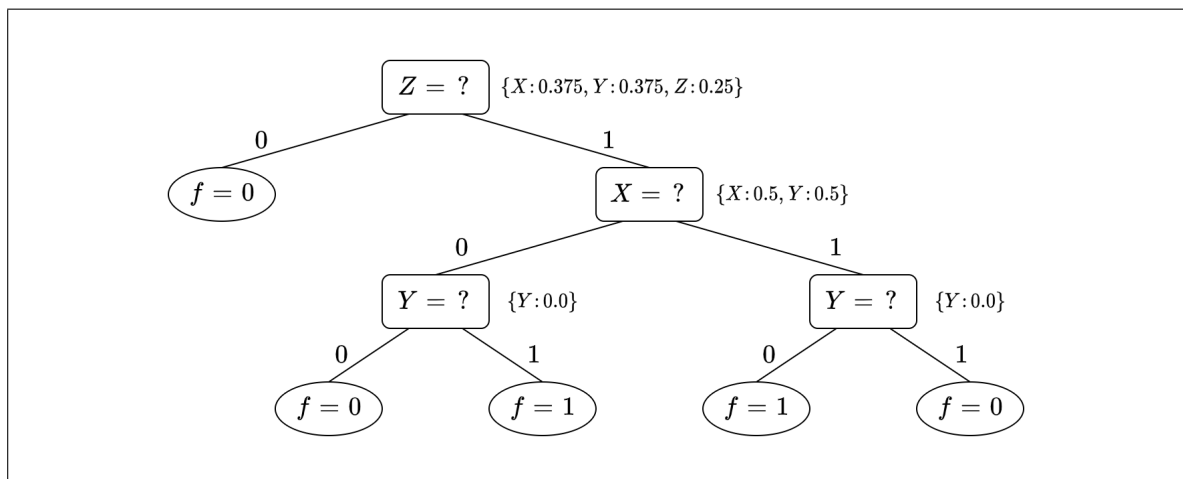
对于一开始的八个样本, X 信息增益为 0, Y 信息增益为 0, Z 信息增益为 0.31, 因此首先选择 Z .

其他同理, 最后得出如下图:



2. 对于一开始的八个样本, X 基尼指数为 0.375, Y 信息增益为 0.375, Z 信息增益为 0.25, 因此首先选择最小的 Z .

其他同理, 最后得出如下图:



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:

(a) 训练集				(b) 验证集			
编号	爱运动	爱学习	成绩高	编号	爱运动	爱学习	成绩高
1	是	是	是	6	是	是	是
2	否	是	是	7	否	是	否
3	是	否	否	8	是	否	否
4	是	否	否	9	否	否	否
5	否	否	是				

Table 2: 人造数据集

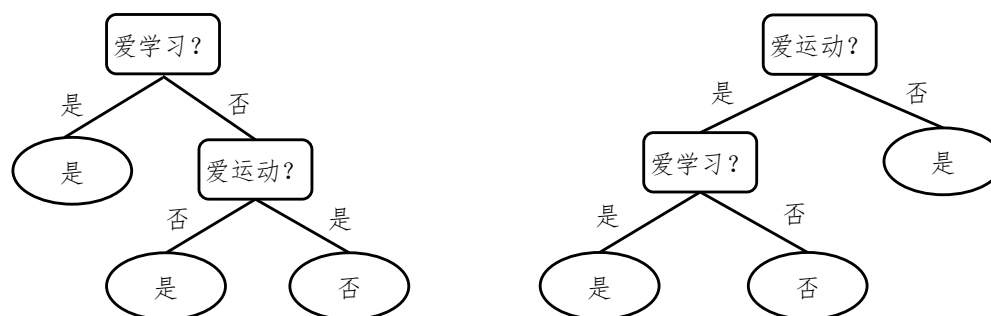


Figure 1: 人造数据决策树结果

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.
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. 对于一开始的五个样本, ”爱运动” 的信息增益为 0.42, ”爱学习” 的信息增益也为 0.42, 所以两棵决策树的第一层无论是选择”爱运动” 还是”爱学习”, 均是正确的.

对于左边的决策树, 第二层的”爱运动” 的信息增益为 0.92, 因此左边的决策树验证完毕.

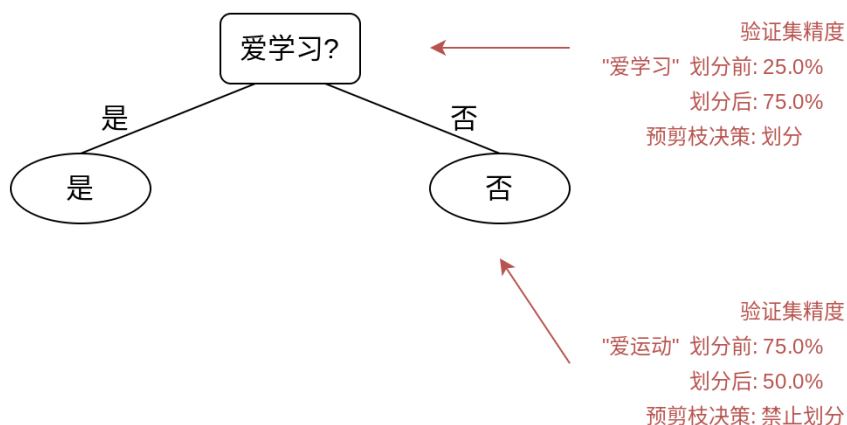
对于右边的决策树, 第二层的”爱学习” 的信息增益为 0.92, 因此右边的决策树验证完毕.

2. 对于左边的决策树:

首先是预剪枝, 对于”预剪枝” 若不进行剪枝, 那么每个样本都会被标记为”是”, 编号 {6} 的样例被分类正确, 精度为 25.0%. 若进行划分, 那么第 6 个或第 7 个样例会被分类错误, 第 8 和第 9 个样例分类正确, 精度为 75.0%, 因此我们进行划分.

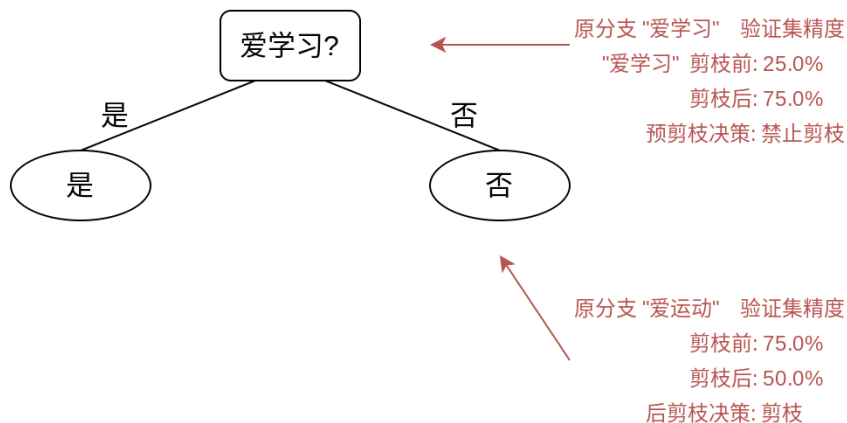
对于”爱学习”, 如果进行剪枝, 那么第 9 个样例会从正确转为错误, 使得验证集精度变为 50%, 因此我们禁止划分.

预剪枝



同理有后剪枝的结果:

后剪枝



对于右边的决策树:

首先是预剪枝, 对于"预剪枝"若不进行剪枝, 那么每个样本都会被标记为"是", 编号 {6} 的样例被分类正确, 精度为 25.0%. 若进行划分, 只有第 8 个样例会被分类正确, 精度为 25.0%, 因此我

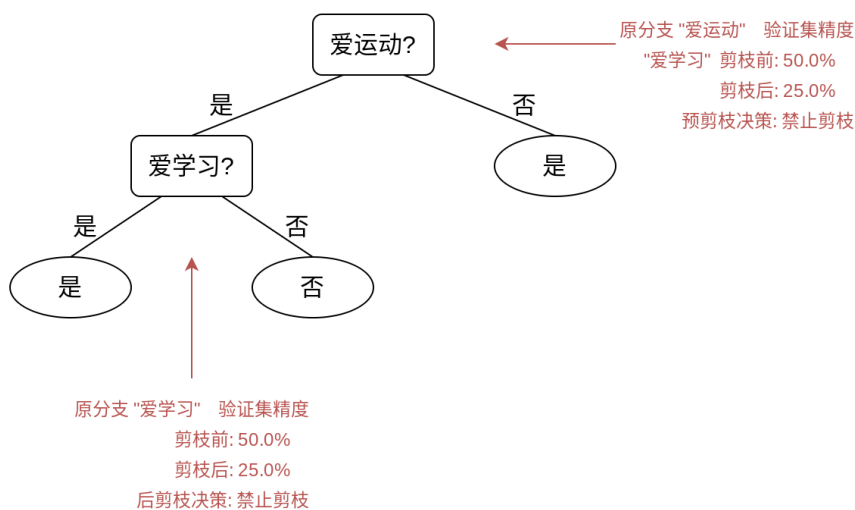
们禁止划分.

预剪枝



同理有后剪枝的结果:

后剪枝



3. 对于左边的决策树来说, 预剪枝和后剪枝在训练集上的准确率均为 80%, 在测试集上的准确率均为 75%.

对于右边的决策树来说, 预剪枝在训练集上的准确率均为 60%, 在测试集上的准确率均为 25%. 后剪枝在训练集上的准确率均为 100%, 在测试集上的准确率均为 50%.

因此我们可知, 后剪枝的拟合能力较强.

四. (20 points) 连续与缺失值

- 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

- 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

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

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。

- 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. 对于第一层,

计算可得 $T_a = \{3.5, 5.0, 7.5\}$

并且有 $\text{Ent}(D) = - \sum_{k=1}^2 \frac{1}{2} \log_2 \frac{1}{2} = 1.0$

同理可以算出

$$\text{Gain}(D, a, 3.5) = 0.31$$

$$\text{Gain}(D, a, 5.0) = 0.00$$

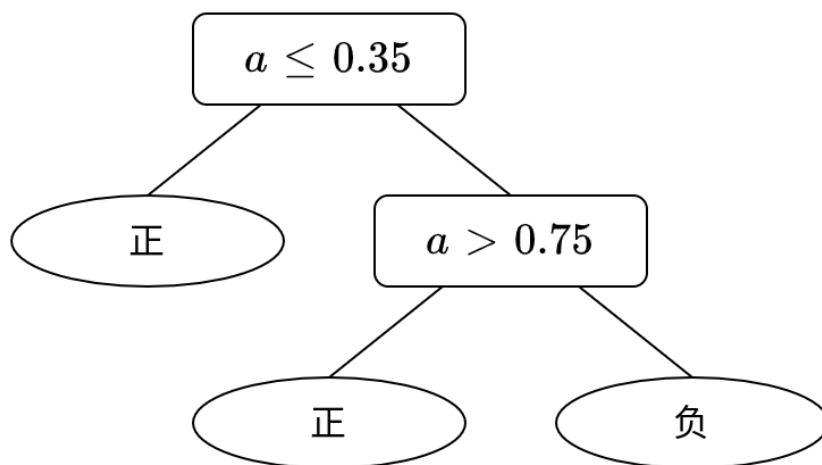
$$\text{Gain}(D, a, 7.5) = 0.31$$

因此第一层是以 3.5 为划分点进行划分.

对于第二层,

同理可以算出 $\text{Gain}(D', a, 5.0) = 0.25, \text{Gain}(D', a, 7.5) = 0.92$

因此第一层是以 7.5 为划分点进行划分.



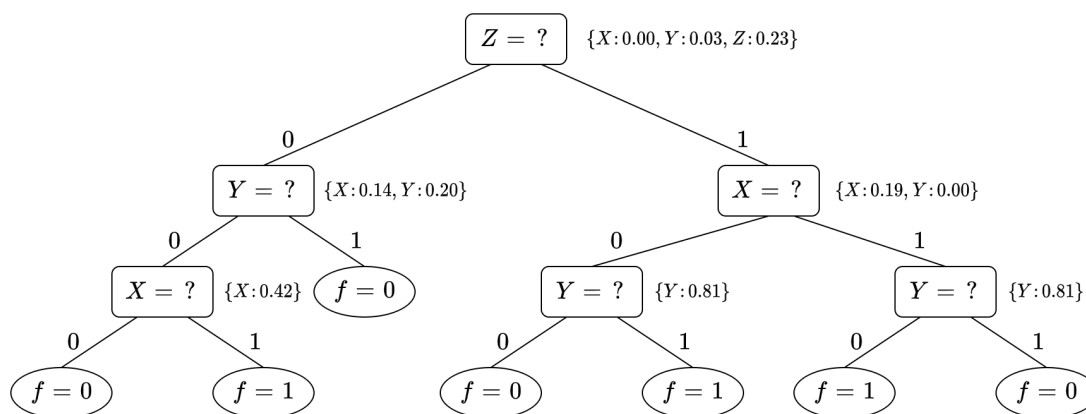
2. 使用公式

$$\text{Gain}(D, a) = \rho \times (\text{Ent}(\tilde{D}) - \sum_{v=1}^V \tilde{r}_v \text{Ent}(\tilde{D}^v))$$

$\text{Gain}(D, X) = 0.0$, $\text{Gain}(D, Y) = 0.03$, $\text{Gain}(D, Z) = 0.24$

因此我们第一层选择 Z 节点进行划分, 并且通过 $w'_x = \tilde{r}_v \cdot w_x$ 不断调整权重值.

其他层同理, 最终可得



与问题二第 1 小问构造的决策树进行比较我们可以看出, 使用带缺失值的样本进行决策, 由于带缺失值样本会被不断复制到不同的分支中, 并且要维持每个样本的权重, 所以需要更大的计算量.

局限性就是, 由于带缺失值样本进入到了不同的分支中, 可能会导致一些分支出现了本不该有的节点, 导致过拟合, 例如这里的 $Z = 0$ 分支.

3. 用类似于构造决策树的方式, 碰到缺失值, 就划分给每个分支对应的权重.

下面我们统一使用问题二左边的决策树进行讨论.

对于编号 6 的样例, 有 $0.4 + 0.6 \times 0.0 = 40\%$ 的概率, 标签为”是”, 有 60% 的概率, 标签为”否”. 则标签为”否”.

对于编号 7 的样例, 有 100% 的概率, 标签为”是”.

对于编号 8 的样例, 有 $0.4 + 0.6 \times 1.0 = 100\%$ 的概率, 标签为”是”.

对于编号 9 的样例, 有 33% 的概率, 标签为”是”, 有 67% 的概率, 标签为”否”. 选择标签为”否”.

该方法的特点是, 并不能确定性地输出一个标签, 只能输出选择各个标签的各个概率值, 并且计算量也较大.

该方法的局限性是, 由于不能确定性地输出标签, 只能输出概率, 在不同标签对应概率相近时, 很有可能分类错误.

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 $\mathbf{x}_i = (x_{i1}; x_{i2})$.

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: