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 $\operatorname{Ent}(D)$ 定义如下

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

请证明信息熵的上下界为

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

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:

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

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

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

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 $i = 1, 2, \cdots$, 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

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

显然在 $0 \leq p_k \leq 0$ 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

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

将其转化为矩阵形式则有

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

对其求微分得

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

因此有

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

令 $\frac{\partial L(\boldsymbol{p},\lambda)}{\partial\boldsymbol{p}}=0$ 即可知 $\boldsymbol{p}$ 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

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

3.

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

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 $\operatorname{Gain}(D,a)$ 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:

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

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?

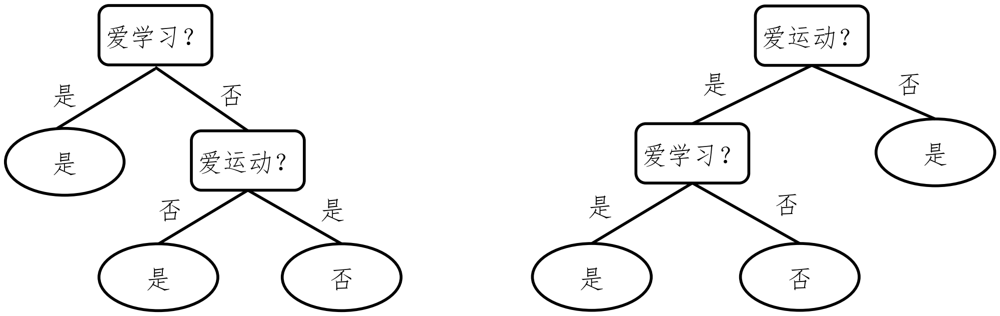

Figure 1: 人造数据决策树结果

解：

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

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。

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:

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: