作业 1：数据认知与预处理

注意：要求给出必要的公式和中间计算过程，所有结果精确到小数点 2 位。

1. 现对数据集中的“年龄”收集到如下观测值：$32,20,30,29,18,21,24,26$。

a. 对以上属性分别进行均值为 0，方差为 1 的标准化（方差计算用有偏）。

答：令均值为μ，标准差为σ，则通过以下公式进行标准化：

$$x'_j = \frac{x_j - \mu}{\sigma}$$

根据题目，得到μ=25，σ=4.77(有偏)

代入以上公式，得到标准化后的观测值为：

$\boldsymbol{x}'$ =[1.47， -1.05，1.05，0.84， -1.47，-0.84，-0.21，0.21]

b. 对以上属性进行[0,1]标准化，即最小值为 0， 最大值为 1。

答：令 max 和 min 表示原观测值的最大和最小值，则通过以下公式得到标准化后的x'_j

$$x'_j = \frac{x_j - min}{max - min}$$

从题目可知 max=32, min=18 （如果题目给定最大最小值则用题目给定值），代入以上公式得到标准化后的观测值为：

$\boldsymbol{x}'$ =[1， 0.14，0.86，0.79， 0，0.21，0.43，0.57]

2. 现采集到以下蘑菇数据集，其中 NA 表示缺失值，请用每个类别的中位数/众数填补缺失值。假设颜色的可能取值为：{红色、褐色、白色、棕色}。

编号	尺寸	颜色	类别
1	2.3	红色	有毒
2	2.4	褐色	无毒
3	1.8	红色	有毒
4	NA	褐色	无毒
5	1.6	白色	无毒
6	2.4	NA	有毒
7	1.5	棕色	无毒

答：

表格中一共有两个缺失值，分别是标记为“无毒”的样本 4 的尺寸缺失和标记为“有毒”的样本 6 的颜色缺失。

属于“无毒”的样本 {2，5，7}在尺寸上有取值，按从小到大排序为：1.5， 1.6， 2.4，所以“无毒”类别对应尺寸的中位数是 1.6。根据题意，样本 4 的尺寸缺失值用 1.6 来填充。

属于“有毒”的样本{1，3}在颜色上有取值，均为红色，所以该类别颜色的众数是红色。根据题意，样本 6 的颜色用红色填充。

3. 给定$\boldsymbol{x} = [1,0,-1]^T, \boldsymbol{y} = [-1,1,0]^T, \boldsymbol{z} = [-2,1,-1]^T$，计算这三个向量两两之间的欧式距离、曼哈顿距离和余弦相似度。

答：

欧式距离：

$$d_{Euc}(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(1-(-1))^2 + (0-1)^2 + (-1-0)^2} = \sqrt{6} = 2.45$$

$$d_{Euc}(\boldsymbol{x},\boldsymbol{z}) = \sqrt{(1-(-2))^2 + (0-1)^2 + (-1-(-1))^2} = \sqrt{10} = 3.16$$

$$d_{Euc}(\boldsymbol{y},\boldsymbol{z}) = \sqrt{(-1-(-2))^2 + (1-1)^2 + (0-(-1))^2} = \sqrt{2} = 1.41$$

曼哈顿距离：

$$d_{Manh}(\boldsymbol{x},\boldsymbol{y}) = |1-(-1)| + |0-1| + |-1-0| = 4$$
$$d_{Manh}(\boldsymbol{x},\boldsymbol{z}) = |1-(-2)| + |0-1| + |-1-(-1)| = 4$$
$$d_{Manh}(\boldsymbol{y},\boldsymbol{z}) = |-1-(-2)| + |1-1| + |0-(-1)| = 2$$

余弦相似度：

$$s_{cos}(\boldsymbol{x},\boldsymbol{y}) = \frac{x^T y}{\|x\|\|y\|} = \frac{-1}{\sqrt{2}\sqrt{2}} = -\frac{1}{2} = -0.5$$

$$s_{cos}(\boldsymbol{x},\boldsymbol{z}) = \frac{x^T z}{\|x\|\|y\|} = \frac{-2+1}{\sqrt{2}\sqrt{6}} = -\frac{\sqrt{3}}{6} = -0.29$$

$$s_{cos}(\boldsymbol{y},\boldsymbol{z}) = \frac{y^T z}{\|y\|\|z\|} = \frac{2+1}{\sqrt{2}\sqrt{6}} = \frac{\sqrt{3}}{2} = 0.87$$

4. 简答：（1）哪些原因导致实际应用中收集到的数据往往存在噪声和缺失值？（2）归一化的主要作用是什么？

答（1）导致数据中存在噪声和缺失值的原因主要包括：数据采集设备故障、数据传输、文件转换时发生的丢失、人为或计算机输入错误等。

（2）不同特征的取值范围可能存在较大不同，归一化的主要作用是使所有特征取值服从同一分布或同一个取值范围，避免某些特征信息被忽略，提高算法建模时的有效性。

作业 3：朴素贝叶斯分类、k 近邻分类

1. 根据顾客的年龄、收入、是否是学生来预测是否购买电脑。基于以下训练集请**先构造**朴素贝叶斯分类器，再给出测试集的预测结果。注意：年龄为连续属性，正态分布概率函数为$p(\mu,\sigma) = \frac{\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}$，方差计算采用 $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2$，其中$\mu$为均值，其他两个类别属性取值分别为“收入”={高、中、低}，“学生”={是、否}。

训练集

编号	类别：是否买电脑	年龄	收入	学生
1	否	22	高	否
2	否	24	高	否
3	是	57	中	否
4	是	54	低	是
5	否	50	低	是
6	是	37	低	是
7	是	34	高	是
8	否	48	中	否

测试集

编号	类别：是否买电脑	年龄	收入	学生
9	是	30	低	是
10	是	53	中	是
11	是	24	中	是

a.分类器构造，即计算所有可能用到的概率；

注意：该题的预测目标是**“是否买电脑”，即这一列才是类别标签**，最后一列“学生”只是其中一个属性。

计算每个类的先验概率：

$P(C_i) = |C_{i,D}|/|D|$，$P(C_{是}) = 4/8 = 0.5$，$P(C_{否}) = 4/8 = 0.5$

计算离散属性“收入”、“是否是学生”的各个取值在每个类的概率：

$P(收入 = 低 / C_{是}) = 2/4 = 0.5$，$P(收入 = 低 / C_{否}) = 1/4 = 0.25$

$P(收入 = 中 / C_{是}) = 1/4 = 0.25$，$P(收入 = 中 / C_{否}) = 1/4 = 0.25$

$P(收入 = 高 / C_{是}) = 1/4 = 0.25$，$P(收入 = 高 / C_{否}) = 2/4 = 0.5$

$P(学生 = 否 / C_{是}) = 1/4 = 0.25$，$P(学生 = 否 / C_{否}) = 3/4 = 0.75$

$P(学生 = 是 / C_{是}) = 3/4 = 0.75$，$P(学生 = 是 / C_{否}) = 1/4 = 0.25$

计算连续属性“年龄”对每个类的均值和标准差

$\mu_{年龄,C_{是}} = \frac{1}{4}(57+54+37+34) = 45.5$，

$\sigma^2_{年龄,C_{是}} = \frac{1}{4-1}((57-45.5)^2 + (54-45.5)^2 + (37-45.5)^2 + (34-45.5)^2) = 136.33$，

$\sigma_{年龄,C_{是}} = 11.68$ 。

$\mu_{年龄,C_{否}} = \frac{1}{4}(22+24+50+48) = 36$，

$\sigma^2_{年龄,C_{否}} = \frac{1}{4-1}((22-36)^2 + (24-36)^2 + (50-36)^2 + (48-36)^2) = 226.67$，

$\sigma_{年龄,C_{否}} = 15.06$。

构造完毕，得到以下表格：

		是否买电脑？=是（P=0.5）	是否买电脑？=否 (P=0.5)
年龄（均值、标准差）		（45.5， 11.68）	（36， 15.06）
收入	高	0.25	0.5
	中	0.25	0.25
	低	0.5	0.25
学生	是	0.75	0.25
	否	0.25	0.75

b. 基于以上各概率，预测顾客是否会买电脑（通过查询以上概率计算测试样本$\boldsymbol{P(c)P(x|c)}$并给出预测结果）。

测试：计算 $\boldsymbol{P(c)P(x|c) = P(c)\prod_i P(x_i|c)}$， 选择对应值最大的类。

样本 9：

根据$\mu_{年龄,\ C_{是}} = 45.5$，$\sigma_{年龄,\ C_{是}} = 11.68$ 计算得到 $P\left(年龄 / C_{是}\right) = \frac{\exp\left(\frac{-(30-45.5)^2}{2\cdot 136.33}\right)}{\sqrt{2\pi}\cdot 11.68} = 0.0142$

根据$\mu_{年龄,\ C_{否}} = 36$，$\sigma_{年龄,\ C_{否}} = 15.06$ 计算得到 $P\left(年龄 / C_{否}\right) = \frac{\exp\left(\frac{-(30-36)^2}{2\cdot 226.67}\right)}{\sqrt{2\pi}\cdot 15.06} = 0.0245$

根据 $P(收入 = 低 / C_{是}) = 0.5$，$P(学生 = 是 / C_{是}) = 0.75$ ，得到

$$P\left(X / C_{是}\right) = P\left(年龄 / C_{是}\right) P\left(收入 = 低 / C_{是}\right) P\left(学生 = 是 / C_{是}\right) = 0.0142 \cdot 0.5 \cdot 0.75 = 0.0053$$

根据 $P(收入 = 低 / C_{否}) = 0.25$，$P(学生 = 是 / C_{否}) = 0.25$ ，得到

$$P\left(X / C_{否}\right) = P\left(年龄 / C_{否}\right) P\left(收入 = 低 / C_{否}\right) P\left(学生 = 是 / C_{否}\right) = 0.0245 \cdot 0.25 \cdot 0.25 = 0.0015$$

$$P\left(C_{是}\right) P\left(X / C_{是}\right) = 0.5 \cdot 0.0053 = 0.00265$$

$$P\left(C_{否}\right) P\left(X / C_{否}\right) = 0.5 \cdot 0.0015 = 0.00075$$

$$P\left(C_{是}\right) P\left(X / C_{是}\right) > P\left(C_{否}\right) P\left(X / C_{否}\right)$$

预测“是”，实际“是”，正确。

样本 10：

$$P\left(年龄\ /\ C_{是}\right)=\frac{\exp\left(-\frac{(53-45.5)^2}{2\cdot136.33}\right)}{\sqrt{2\pi}\cdot11.68}=0.0278$$

$$P\left(年龄\ /\ C_{否}\right)=\frac{\exp\left(-\frac{(53-36)^2}{2\cdot226.67}\right)}{\sqrt{2\pi}\cdot15.06}=0.0140$$

$P\left(收入=中\ /\ C_{是}\right)=0.25$，$P\left(学生=是\ /\ C_{是}\right)=0.75$ ，得到

$P\left(X\ /\ C_{是}\right)=0.0278\cdot0.25\cdot0.75=0.0052$，

$P\left(收入=中\ /\ C_{否}\right)=0.25$，$P\left(学生=是\ /\ C_{否}\right)=0.25$，得到

$P\left(X\ /\ C_{否}\right)=0.0140\cdot0.25\cdot0.25=0.0009$

$P\left(C_{是}\right)P\left(X\ /\ C_{是}\right)=0.5\cdot\ 0.0052=0.0026$，$P\left(C_{否}\right)P\left(X\ /\ C_{否}\right)=0.5\cdot0.0009=0.00045$，

$$P\left(C_{是}\right)P\left(X\ /\ C_{是}\right)>P\left(C_{否}\right)P\left(X\ /\ C_{否}\right)$$

预测“是”，实际“是”，正确。

样本 11：

$$P\left(年龄\ /\ C_{是}\right)=\frac{\exp\left(-\frac{(24-45.5)^2}{2\cdot136.33}\right)}{\sqrt{2\pi}\cdot11.68}=0.0063,$$

$$P\left(年龄\ /\ C_{否}\right)=\frac{\exp\left(-\frac{(24-36)^2}{2\cdot226.67}\right)}{\sqrt{2\pi}\cdot15.06}=0.0193$$

$P(收入=中\ /\ C_{是})=0.25$， $P(学生=是\ /\ C_{是})=0.75$，得到

$$P\left(X\ /\ C_{是}\right)=0.0063\cdot0.25\cdot0.75=0.0012$$

$P(收入=中\ /\ C_{否})=0.25$, $P(学生=是\ /\ C_{否})=0.25$ ，得到

$P\left(X\ /\ C_{否}\right)=0.0193\cdot0.25\cdot0.25=0.00121$

$P\left(C_{是}\right)P\left(X\ /\ C_{是}\right)=0.5\cdot\ 0.0012=0.0006$，$P\left(C_{否}\right)P\left(X\ /\ C_{否}\right)=0.5\cdot0.00121=0.00061$

$$P\left(C_{否}\right)P\left(X\ /\ C_{否}\right)>P\left(C_{是}\right)P\left(X\ /\ C_{是}\right)$$

预测“否”，实际“是”，错误。

2. 如何用 k 近邻算法对以上数据集进行分类？给出实现步骤（包括选用哪种距离度量，k 的设定等），但不用计算。

答：K 近邻算法需要计算样本之间的距离/相似度，这里总共有三个特征，其中“年龄”为连续型特征，“收入”为等级型特征，“学生”是类别型特征。
第一步是把“收入”和“学生”转换成连续特征。“收入”的高、中、低可以分别用 1，0.5，0.1 来替换，“学生”的“是”和“否”可以用长度为 2 的独热编码[1, 0] 和 [0, 1] 来代替，比如把测试例 9 变换成[30, 0.1, 1, 0]。
由于距离的计算与特征取值范围相关，考虑把所有特征线性变化到[0, 1]区间。通过前一步操作，只有年龄的取值范围不在[0,1]区间，所以只要把这个特征进行 0-1 规范化。min-max 规范化需要确定原数据集的最小值和最大值。由于测试样本（新样本）的“年龄”可能大于/小于训练集中看到的“年龄”范围，所以统一把“最大年龄”设置成 60，大于 60 的用 60 替代，“最小年龄”设成 10，小于 10 的用 10 代替，然后进行 0-1 规范化。利用以上方法，测试例 9 中的年龄 30 被映射到 0.4。
给定一个测试样本，进行以上两步变换后，计算其到所有训练样本的距离，这里没有特殊要求，考虑通用的欧式距离。选择 K 个距离最小的样本作为近邻，把该样本标记为近邻中出现最多的类别。由于该数据集很小，K 的取值可以从 2，3，4 中尝试，最后选择验证集上准确率最高的 K。

3. 朴素贝叶斯的核心假设是什么？有什么优缺点？
答：朴素贝叶斯假设给定类别情况下所有属性之间相互独立。
这个假设的优点是简化问题，降低“维度灾难”导致的过拟合风险。缺点是不能描述属性之间的相关性，当实际问题中存在相互关联的属性时，模拟准确率不够理想。

作业 2：决策树构建、模型评估

1. 为什么不能用训练集中的样本来测试训练好的模型？请给出留出法和 K 折交叉验证法的具体步骤。
答：模型测试是为了评估模型的泛化能力。用训练集中的样本进行测试得到的是训练误差，训练误差小可能出现过拟合，因此训练误差不能有效评估模型的泛化能力。

留出法：
一般采用分层采样。从每个类别中按照给定的比例（无放回）随机采样一部分样本作为测试集，每个类别中未被采样到的样本作为训练集。一般用多次留出法的结果取平均。

K 折交叉验证法：
把训练集随机分成 K 个大小一样不相交的子集，每次分别用其中 1 个子集作为测试集，剩下的 K-1 个子集中的所有样本作为训练集，总共训练 K 次，对 K 个结果取平均。

2. 假设有一个测试集，其真实标签和模型预测出来的标签分别如下，请给出混淆矩阵，并计算查准率(precision)、查全率(recall)、和 F1 度量(F1-measure)。

编号	真实标签	预测标签
1	是	否
2	否	否
3	否	否
4	是	是
5	是	否
6	否	否
7	否	是
8	是	是

混淆矩阵：

	预测正例	预测反例
真实正例	TP=|{4,8}|=2	FN=|{1，5}|=2
真实反例	FP=|{7}|=1	TN=|{2，3，6}|=3

$$Precision=\frac{TP}{TP+FP}=\frac{2}{2+1}=0.67$$

$$Recall=\frac{TP}{TP+FN}=\frac{2}{2+2}=0.5$$

$$F1-measure=\frac{2\times Presicion\times Recall}{resicion+Recall}=\frac{2\times0.67\times0.5}{0.67+0.5}=0.57$$

3. 基于顾客的“年龄”、“收入”、“学生”这三个属性构造决策树用于预测一个学生是否会买电脑。其中三个属性的取值范围分别为：“年龄”={青年、中年、老年}，“收入”={高、中、低}，“学生”={是、否}。具体要求如下：
a.用以下训练集基于信息增益构造决策树，给出中间步骤，并画出决策树。
b.基于上面构造的决策树，对测试集中样本进行预测，即顾客是否会买电脑， 并计算 accuracy。
*方便大家计算，现给出以下对数的具体数值：$\log_2(3)=1.5850$，$\log_2(5)=2.3219$

训练集

编号	**类别:** 是否买电脑	年龄	收入	学生
1	否	青年	高	否
2	否	青年	高	否
3	是	中年	高	否
4	是	老年	中	否
5	是	老年	中	是
6	否	老年	低	是
7	否	青年	中	否
8	是	老年	中	是

测试集

编号	**类别:** 是否买电脑	年龄	收入	学生
1	是	中年	低	是
2	是	青年	中	是
3	否	老年	低	否

参考答案

a.以下给出具体的构造过程。注意：该题目中**第一列为要预测的类别**。

划分前数据集：D={1,2,3,4,5,6,7,8}，其信息熵$Ent(D)=-\sum_{i=1}^{n}p_i\ log_2(p_i)=-\frac{4}{8}\log_2\left(\frac{4}{8}\right)-\frac{4}{8}\log_2\left(\frac{4}{8}\right)=1$。

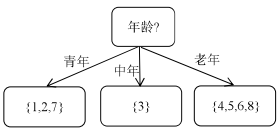
根据$Ent(D,a)=\sum_{v=1}^{V}\frac{|D_v|}{|D|}Ent(D_v)$，以及当前可用属性集$A=\{年龄、收入、学生\}$，计算 A 中每个特征划分后的信息熵如下：

$$Ent(D,年龄)=\frac{3}{8}\left(-\frac{3}{3}log_2\left(\frac{3}{3}\right)\right)+\frac{1}{8}\left(-\frac{1}{1}log_2\left(\frac{1}{1}\right)\right)+\frac{4}{8}\left(-\frac{1}{4}log_2\left(\frac{1}{4}\right)-\frac{3}{4}log_2\left(\frac{3}{4}\right)\right)=0.4056$$

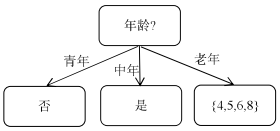
$$Ent(D,收入)=\frac{1}{8}\left(-\frac{1}{1}log_2\left(\frac{1}{1}\right)\right)+\frac{4}{8}\left(-\frac{1}{4}log_2\left(\frac{1}{4}\right)-\frac{3}{4}log_2\left(\frac{3}{4}\right)\right)+\frac{3}{8}\left(-\frac{2}{3}log_2\left(\frac{2}{3}\right)-\frac{1}{3}log_2\left(\frac{1}{3}\right)\right)=0.75$$

$$Ent(D,学生)=\frac{5}{8}\left(-\frac{3}{5}log_2\left(\frac{3}{5}\right)-\frac{2}{5}log_2\left(\frac{2}{5}\right)\right)+\frac{3}{8}\left(-\frac{1}{3}log_2\left(\frac{1}{3}\right)-\frac{2}{3}log_2\left(\frac{2}{3}\right)\right)=0.9512$$

根据信息增益：$\text{Gain}(D,a)=\text{Ent}(D)-\text{Ent}(D,a)$，得到每个特征的信息增益为：$\text{Gain}(D,年龄)=1-0.4056=0.5944$，$\text{Gain}(D,收入)=1-0.75=0.25$，$\text{Gain}(D,学生)=1-0.9512=0.0488$
“年龄”属性的信息增益最高，所以作为当前分裂属性，得到



其中，“青年”和“中年”两个节点样本类别已经一致，所以作为叶子节点，得到：



对“老年”节点再划分，此时 D={4,5,6,8}，$\text{Ent}(D) = -\frac{1}{4}log_2\left(\frac{1}{4}\right) - \frac{3}{4}log_2\left(\frac{3}{4}\right) = 0.8113$。

当前可用属性集更新为$A = \{$收入，学生$\}$，计算 A 中每个特征划分后的信息熵如下：

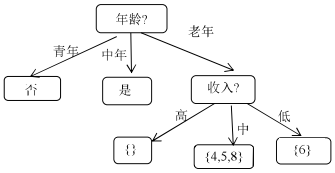
$$Ent(D, 收入) = \frac{1}{4}\left(-\frac{1}{1}log_2\left(\frac{1}{1}\right)\right) + \frac{3}{4}\left(-\frac{3}{3}log_2\left(\frac{3}{3}\right)\right) = 0$$

$$Ent(D, 学生) = \frac{1}{4}\left(-\frac{1}{1}log_2\left(\frac{1}{1}\right)\right) + \frac{3}{4}\left(-\frac{1}{3}log_2\left(\frac{1}{3}\right) - \frac{2}{3}log_2\left(\frac{2}{3}\right)\right) = 0.6887$$

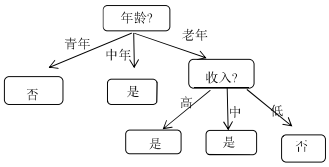
得到每个特征的信息增益为：

$\text{Gain}(D, 收入) = 0.8113 - 0 = 0.8113$，$\text{Gain}(D, 学生) = 0.8113 - 0.6887 = 0.1226$

“收入”属性的信息增益最高，所以作为当前分裂属性，决策树更新为：



其中，“老年->收入=低”和“老年->收入=中”两个节点样本的类别一致，所以作为叶子节点，“老年->收入高”对应的训练样本数目为 0，此时用父亲节点“老年”的样本{4,5,6,8}中数目最多的类别“是”来标记。最后得到的决策树为：



由于当前所有节点均为叶子节点，决策树构造完成。

b.测试：

根据以上所构造决策树，从上至下对各个特征遍历，得到

序号 9：年龄“中年”，预测“是”，实际“是”，正确；

序号 10：年龄“青年”，预测“否”，实际“是”，错误；

序号 11：年龄“老年”、收入“低”，预测“否”，实际“否”，正确；

Accuracy=$\frac{2}{3}$=0.67。

作业 4：聚类

1. 请从应用场景、模型学习过程等方面阐述聚类与分类的区别。

答：聚类是一种无监督学习，基于数据集中样本之间的相似性对样本进行分组，使得同一个组/簇内样本之间相似性大于不同组/簇样本之间的相似性，一般在不知道类别信息的情况下（不知道数据集属于哪些类别），用作为对数据集的初步探索和基本处理。而分类是监督学习，需要足够多的标签数据作为监督信息来训练模型，使其能够较准确的预测未知样本的类别标签。分类用于已知数据集所相关的类别以及有足够多的标记样本的情况下。

2. 对以下数据集进行标准的 k 均值聚类（$d_{ik} = \|x_i - \mu_k\|^2$）。假设 k=2，两个均值$\mu_1$和$\mu_2$分别初始化为第 2、4 个样本，即$\mu_1 = [2,\ 2,\ 1]$，$\mu_2 = [1,\ 0,\ 2]$，给出迭代过程以及收敛时的均值和簇。

序号	属性 1	属性 2	属性 3
1	1	2	1
2	2	2	1
3	0	1	2
4	1	0	2

答：基于初始均值，计算每个样本到两个均值的距离分别为：

$$d_{11} = (1-2)^2 + (2-2)^2 + (1-1)^2 = 1$$
$$d_{12} = (1-1)^2 + (2-0)^2 + (1-2)^2 = 5$$

$argmin_f\ d_{1f} = 1$，*所以第 1 个样本被分到第 1 个类；*

$$d_{21} = (2-2)^2 + (2-2)^2 + (1-1)^2 = 0$$
$$d_{22} = (2-1)^2 + (2-0)^2 + (1-2)^2 = 6$$

$argmin_f\ d_{2f} = 1$，*所以第 2 个样本被分到第 1 个类；*

$$d_{31} = (0-2)^2 + (1-2)^2 + (2-1)^2 = 6$$
$$d_{32} = (0-1)^2 + (1-0)^2 + (2-2)^2 = 2$$

$argmin_f\ d_{3f} = 2$，*所以第 3 个样本被分到第 2 个类；*

$$d_{41} = (1-2)^2 + (0-2)^2 + (2-1)^2 = 6$$
$$d_{42} = (1-1)^2 + (0-0)^2 + (2-2)^2 = 0$$

$argmin_f\ d_{4f} = 1$，*所以第 4 个样本被分到第 2 个类；*

第一次得到划分：C1={1，2}，C2={3，4}

第一次更新后的均值为：

$$\mu_1 = \frac{1}{2}[(1+2), (2+2), (1+1)] = [1.5, 2, 1]$$

$$\mu_2 = \frac{1}{2}[(0+1), (1+0), (2+2)] = [0.5, 0.5, 2]$$

计算每个样本到当前均值的距离分别为：

$$d_{11} = (1-1.5)^2 + (2-2)^2 + (1-1)^2 = 0.25$$
$$d_{12} = (1-0.5)^2 + (2-0.5)^2 + (1-2)^2 = 3.5$$

$argmin_f\ d_{1f} = 1$，*所以第 1 个样本被分到第 1 个类；*

$$d_{21} = (1-1.5)^2 + (2-2)^2 + (1-1)^2 = 0.25$$
$$d_{22} = (2-0.5)^2 + (2-0.5)^2 + (1-2)^2 = 5.5$$

$argmin_f\ d_{2f} = 1$，*所以第 2 个样本被分到第 1 个类；*

$$d_{31} = (0-1.5)^2 + (1-2)^2 + (2-1)^2 = 4.25$$
$$d_{32} = (0-0.5)^2 + (1-0.5)^2 + (2-2)^2 = 0.5$$

$argmin_f\ d_{3f} = 2$，*所以第 3 个样本被分到第 2 个类；*

$$d_{41} = (1-1.5)^2 + (0-2)^2 + (2-1)^2 = 5.25$$
$$d_{42} = (1-0.5)^2 + (0-0.5)^2 + (2-2)^2 = 0.5$$

$argmin_f\ d_{4f} = 1$，*所以第 4 个样本被分到第 2 个类；*

第二次划分：C1={1，2}，C2={3，4}，与之前一样，说明已经收敛。

最后的均值为：

$\mu_1 = [1.5, 2, 1]$，　$\mu_2 = [0.5, 0.5, 2]$

3. 对以上数据集进行基于平均连接（average-linkage）的凝聚层次聚类。

提示：先计算两两样本之间的距离矩阵（欧式距离），给出每一步更新后的簇间距离矩阵。

答：先计算两两样本之间的（欧式）距离矩阵：

$$R = \begin{bmatrix} & 1.00 & 1.73 & 2.24 \\ 1.00 & & 2.45 & 2.45 \\ 1.73 & 2.45 & & 1.41 \\ 2.24 & 2.45 & 1.41 & \end{bmatrix}$$

初始化每个样本为一个簇：$C_1 = \{x_1\}$，$C_2 = \{x_2\}$，$C_3 = \{x_3\}$，$C_4 = \{x_4\}$

初始化簇间距离D等于R。

从D中找出平均距离最小的两个簇：C_1和C_2进行合并，得到新的划分：$C_1 = \{x_1, x_2\}$，

$C_2 = \{x_3\}$，$C_3 = \{x_4\}$，用平均距离更新D得

$$D = \begin{bmatrix} & 2.09 & 2.35 \\ 2.09 & & 1.41 \\ 2.35 & 1.41 & \end{bmatrix}$$

其中$d_{12} = \frac{1}{2}(r_{13} + r_{23}) = 2.09$，$d_{13} = \frac{1}{2}(r_{14} + r_{24}) = 2.35$，$d_{23} = r_{34}$

从D中找出平均距离最小的两个簇：C_2和C_3进行合并，得到新的划分$C_1 = \{x_1, x_2\}$，$C_2 = \{x_3, x_4\}$，用平均距离更新D得

$$D = \begin{bmatrix} & 2.22 \\ 2.22 & \end{bmatrix}$$

其中$d_{12} = \frac{1}{4}(r_{13} + r_{14} + r_{23} + r_{24}) = 2.22$

最后把C_1和C_2合并得到$C_1 = \{x_1, x_2, x_3, x_4\}$

基于以上过程得到以下树状结构：