第七章**作业**：（共 12 分）（多选（1 分）×2，大题（10 分））

1.【多选】下列属于无监督学习方法的是__**AC**___ 。
A. PCA　　B. FDA　　C. k-means　　D. kNN。

2.【多选】下面关于聚类算法的描述，__**AD**___ 是错误的。
A. k-medoids 聚类的结果易受噪音或离群点的影响。
B. k-means 算法比 k-medoids 算法更高效。
C. k-means 只能发现球状的簇，而 DBSCAN 能够发现任意形状的簇。
D.算法 DBSCAN 对噪声/离群点比较敏感。

3. 简述有监督学习和无监督学习的区别。(2 分)

- **有监督学习 (supervised learning)：**
  - 依赖于已经标注好类别标签的样本构成的训练集
  - 旨在从训练集中学习到具体的决策规则
  - 常用于：分类、回归
- **无监督学习 (unsupervised learning)：**
  - 训练集中样本的类别标记未知
  - 旨在发现训练集中内在的结构或规律
  - 常用于：聚类、概率密度估计

4. 给定一个样本集{(0,0), (0,1), (1,0), (2, 1), (2, 3), (3, 4)}，试用基于欧式距离的 k-均值聚类算法将该数据集聚成 2 类，要求初始的聚类中心选择 c1=(0,0) , c2 = (0,1)。请写出详细的计算过程，给出算法每次迭代结束后，将哪些样本点聚在一个簇中，并注明每次迭代的聚类中心。（8 分）

解答：记每个样本依次为 x1, x2, x3, x4, x5, x6,

第一次迭代中：初始聚类中心为 c1=(0,0) , c2 = (0,1)

||x1-c1|| < ||x1-c2||, x1 属于簇 1；

||x2-c1|| > ||x2-c2||, x1 属于簇 2；

||x3-c1|| < ||x1-c2||, x1 属于簇 1；

||x4-c1|| > ||x4-c2||, x1 属于簇 2；

||x5-c1|| > ||x5-c2||, x1 属于簇 2；

||x6-c1|| > ||x6-c2||, x1 属于簇 2；

第一次迭代结束后，簇 1 中包含 x1，x3；簇 2 中包含 x2，x4，x5，x6.　　**(2.5 分)**

计算下次迭代的聚类中心：c1=(0.5, 0)，c2 = (7/4, 9/4)，第二次迭代：

||x1-c1|| < ||x1-c2||, x1 属于簇 1；

||x2-c1|| < ||x2-c2||, x1 属于簇 1；

||x3-c1|| < ||x1-c2||, x1 属于簇 1；

||x4-c1|| > ||x4-c2||, x1 属于簇 2；

||x5-c1|| > ||x5-c2||, x1 属于簇 2；

||x6-c1|| > ||x6-c2||, x1 属于簇 2；

第二次迭代结束后，簇 1 中包含 x1，x2，x3；簇 2 中包含 x4，x5，x6。 **(2.5 分)**

重新计算聚类中心为 c1 = (1/3, 1/3), c2 = (7/3, 8/3)，第三次迭代：

||x1-c1|| < ||x1-c2||, x1 属于簇 1；

||x2-c1|| < ||x2-c2||, x1 属于簇 1；

||x3-c1|| < ||x1-c2||, x1 属于簇 1；

||x4-c1|| > ||x4-c2||, x1 属于簇 2；

||x5-c1|| > ||x5-c2||, x1 属于簇 2；

||x6-c1|| > ||x6-c2||, x1 属于簇 2；

重新计算聚类中心，与上次迭代的聚类中心相同，所以算法终止，最终得到的聚类结果为：
簇 1 中包含 x1，x2，x3；簇 2 中包含 x4，x5，x6。 **(3 分)**

第八章作业：（共 10 分）

1. 对某地区的人种情况进行调查，得到如下的一张表：

人员	眼睛颜色	头发颜色	所属人种
1	黑色	黑色	黄种人
2	蓝色	金色	白种人
3	灰色	金色	白种人
4	蓝色	红色	白种人
5	灰色	红色	白种人
6	黑色	金色	混血
7	灰色	黑色	混血
8	蓝色	黑色	混血

其中表中每一行信息表示一个人员的信息，包括眼睛颜色和头发颜色，最后一列表明该人员属于哪种人种。根据该表构建一棵决策树，能根据一个人的眼睛和头发的颜色来判断其所属人种。要求在选择划分属性时，使用信息增益的准则，计算熵时以 2 为底。请写出详细的构建过程，并画出最终得到的决策树。
（7 分）

解答：记训练集为 D，则

Ent(D) = - 1/8 * log2 (1/8)-4/8 * log2 (4/8)-3/8 * log2 (3/8) = 1.4056

Ent(D|“眼睛颜色”)

= - 2/8 *(1/2 * log2(1/2) *2) - 3/8 * (2/3 * log2 (2/3) + 1/3 * log2(1/3))*2

= 0.9387

gain(D, “眼睛颜色”) = Ent(D) - Ent(D|“眼睛颜色”) = 0.4669 **(2 分)**

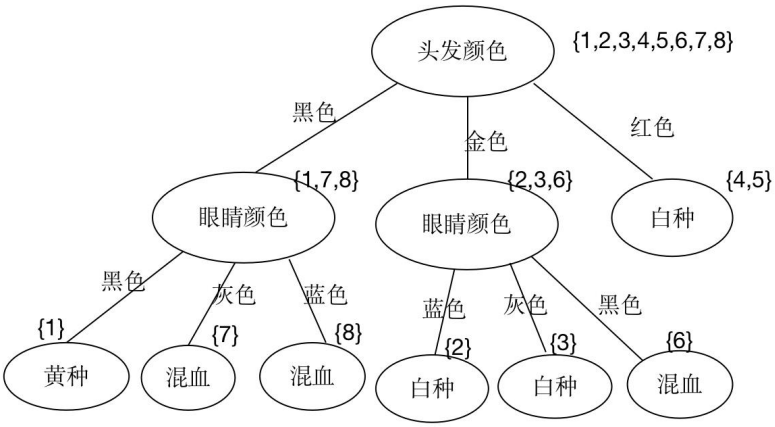
Ent(D|“头发颜色”)

= - 3/8 * (2/3 * log2 (2/3) + 1/3 * log2(1/3))*2 + 0

= 0.6887

gain(D, “头发颜色”) = Ent(D) - Ent(D|“头发颜色”) = 0.7169 **(2 分)**

所以，在根结点处，选择“头发颜色”作为划分属性。因为“头发颜色”取值个数为 3 个，所以生成 3 个子结点，然后依次对每个子结点处的样本进行划分，此时只能选择“眼睛颜色”作为划分属性，其取值个数为 3 个，所以也是生成 3 个子结点，最终得到的决策树如下图所示：



(3 分)

2. 设 X 是一个离散型随机变量，其概率分布为：

$$P(X = x_i) = p_i, \quad i = 1,2,\cdots m$$

求 X 的熵的最大值，其中熵以纳特为单位，求解过程按步骤计分。(3 分)