1. 假定用于分析的数据包含属性 age，数据元组中 age 的值如下:
[22, 36, 25, 70, 19, 20, 33, 45, 16, 25, 20, 15, 35, 21, 30, 33, 25, 40, 22, 16, 13, 35, 35, 52, 33, 46, 35, 25]。分别按照等宽和等频的方法将上述数据划分到 4 个不同区间内。

2. 假设你有以下数据集，表示某个班级学生的数学成绩（单位：分）： [60, 70, 80, 90, 100]，请对该数据集进行标准化处理。

3. 对于下面的向量 x 和 y，计算指定的相似性或距离度量。
(a) x=(1，1，1，1)，y=(2，2，2，2)，计算余弦、皮尔森相关系数、欧几里得。
(b) x=(0，1，0，1)，y=(1，0，1，0)，计算余弦、皮尔森相关系数、欧几里得。
(c) x=0101010001，y=0100011000，计算两个二元向量之间的简单匹配系数和 Jaccard 相似度。

4. 考虑如下二分类问题的训练样本集
(a) 整个训练样本集关于类属性的熵是多少?
(b) 关于这些训练样本，a_1和a_2的信息增益是多少?
(c) 对于连续属性a_3，计算所有可能的划分的信息增益。
(d) 根据信息增益，哪个是最佳划分(在a_1、a_2和a_3中)?
(e) 根据分类误差，哪个是最佳划分(在a_1和a_2中)?
(f) 根据基尼指数，哪个是最佳划分(在a_1和a_2中)?

实例	a_1	a_2	a_3	类标号
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

5. 考虑下表中的一维数据集。
(a)根据 1-最近邻、3-最近邻、5-最近邻及 9-最近邻，对数据点x=5.0 分类(使用多数表决)。

(b)根据距离权衡每个最近邻x_i的影响$w_i = \frac{1}{|x_i - x|}$,使用距离加权表决方法重复前面的分析。

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	−	−	+	+	+	−	−	+	−	−