第二次 课后作业

提交截止时间：2022 年 11 月 13 日 20：00

问题一：考虑三个正态分布函数 $\mathcal{N}_1(\boldsymbol{\mu}_1, \Sigma_1), \mathcal{N}_2(\boldsymbol{\mu}_2, \Sigma_2)$ 和 $\mathcal{N}_3(\boldsymbol{\mu}_3, \Sigma_3)$，具体参数如下：

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 1.2 & 0.4 \\ 0.4 & 1.8 \end{bmatrix}$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 2.1 \\ 1.9 \end{bmatrix}, \boldsymbol{\mu}_3 = \begin{bmatrix} -1.5 \\ 2.0 \end{bmatrix}$$

按照随机样本生成规则为：前两个样本使用 $\mathcal{N}_2$ 生成，第三个样本使用 $\mathcal{N}_1$ 生成，第四个样本使用 $\mathcal{N}_3$ 生成。重复上述规则生成 500 个样本。

随机样本遵循的概率密度函数建模为混合模型如下：

$$p(\boldsymbol{x}) = \sum_{i=1}^{3} P_i \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$$

编程解决下述问题：

1) （10 分）绘制所生成 500 个随机样本的散布图；
2) （30 分）给出对所生成随机样本集合混合概率分布模型的参数估计 $\hat{\boldsymbol{\mu}}_i, \hat{\Sigma}_i, \hat{P}_i$，其中 $i = 1,2,3$；
3) （20 分）分析真实参数和估计参数存在差异的原因，并提出解决办法。

问题二：应用最近邻分类方法解决文件 HW#2.mat 中样本的分类问题。

HW#2.mat 文件中包含六个数组 c_1, c_2, c_3 和 t_1, t_2, t_3，每个数组的维数均为 500×2，其中数组 $c_i (i = 1,2,3)$ 为训练样本集合，$t_i (i = 1,2,3)$ 为测试样本集合，i 表示类标签。

编程解决下述问题：

1) （20 分）给出具有最佳测试性能的 k 值 k_{opt} 及对应的分类错误率；
2) （10 分）解释当 $k = 1, k_{opt}, 50$ 三种情形下测试性能的不同；
3) （10 分）比较 1) 中性能结果和最大似然分类器（0-1 损失函数且相等先验概率假设下的 Bayes 分类器）在该数据集上的分类结果（假设三个类别均满足正态分布）。