《模式识别》作业六

姓名：谷绍伟　　学号：202418020428007

1　简述题

1. 请简述 Adaboost 算法的设计思想和主要计算步骤。

答：设计思想：给定训练集，寻找比较粗糙的分类规则（弱分类器）要比寻找精确的分类规则要简单得多。因此从弱学习算法出发，反复学习，得到一系列弱分类器；然后组合这些弱分类器，构成一个强分类器。

策略：改变训练数据的概率（权重）分布，针对不同的训练数据的分布，调用弱学习算法来学习一系列分类器。

主要计算步骤：训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$，$x_i \in R^d, y_i \in \{-1, +1\}$，计算过程如下：

- 初始化训练数据的权值分布：$D_1 = \{w_{11}, w_{12}, \ldots, w_{1n}\}$，$w_{1i} = 1/n, i = 1, \ldots, n$；
- 对于 $m = 1, 2, \ldots, M$：
  - 使用具有权值分布 D_m 的训练数据，学习弱分类器 $G_m(\mathbf{x})$：　$\boldsymbol{X} \to \{-1, +1\}$
  - 计算 $Gm(x)$ 在训练数据集上的加权分类错误率：$e_m = P(G_m(\mathbf{x}_i) \neq y_i) = \sum_{i=1}^{n} w_{mi} I(G_m(\mathbf{x}_i) \neq y_i)$
  - 计算 $Gm(x)$ 的贡献系数：$\alpha_m = \frac{1}{2} \ln \frac{1-e_m}{e_m}$，其中 α_m 表示 $Gm(x)$ 在最终分类器中的重要性。当 $e_m \leq 0.5$ 时，$\alpha_m \geq 0$，同时 α_m 将随着 e_m 的减小而增大。即分类误差率越小的弱分类器在最终分类器中的作用越大；
  - 更新训练数据集的权重分布：$D_{m+1} = \{w_{m+1,1}, w_{m+1,2}, ..., w_{m+1,n}\}$，计算过程如下：

$$
\begin{aligned}
W_{m+1,i} &= \frac{w_{m,i}}{Z_m} \times \begin{matrix} \exp(-\alpha_m), \text{if } G_m(\mathbf{x}_i) = y_i \\ \exp(\alpha_m), \text{if } G_m(\mathbf{x}_i) \neq y_i \end{matrix} \\
&= \frac{W_{m,i}}{Z_m} \times \exp(-\alpha_m y_i G_m(\mathbf{x}_i))
\end{aligned} \tag{1}
$$

  其中，Z_m 是归一化因子，它使 D_{m+1} 成为一个概率分布：$Z_m = \sum_{i=1}^{n} w_{mi} \exp(-\alpha_m y_i G_m(\mathbf{x}_i))$；
- 构建弱分类器的线性组合：$f(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m G_m(\mathbf{x})$，对于两类分类问题，得到最终的分类器：$G(\mathbf{x}) = \text{sign}\left(f(\mathbf{x}) \right) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(\mathbf{x}) \right)$

2. 请从混合高斯密度函数估计的角度，简述 K-Means 聚类算法的原理 (请主要用文字描述，条理清晰)；请给出 K-Means 聚类算法的计算步骤；请说明哪些因素会影响 K-Means 算法的聚类性能。

答：K-Means 算法可以看作是一种简化的混合高斯模型。在混合高斯模型中，目标是估计每个高斯分布的参数（均值、协方差等）以最大化数据的似然函数。K - Means 聚类算法可以看作是一种简化的混合高斯模型估计方法。它假设每个聚类（相当于一个高斯分布）的数据点具有相同的协方差矩阵（通常假设为球形，即协方差矩阵是对角矩阵且对角元素相等），并且每个聚类中心对应一个高斯分布的均值。K - Means 算法通过迭代地寻找聚类中心（近似高斯分布的均值），使得数据点到其所属聚类中心的距离之和最小，这类似于在混合高斯模型中优化每个高斯成分的均值向量。

计算步骤：

- 初始化：确定聚类的数目 K，随机地从数据集中选择 K 个数据点作为初始聚类中心 $\mu_1, \mu_2, \cdots, \mu_K$；
- 将数据分配到聚类：
  - 对于数据集中的每个数据点 x_i，计算它到每个聚类中心 μ_j（$j = 1, 2, \cdots, K$）的距离 $d(x_i, \mu_j)$（通常使用欧几里得距离 $d(x_i, \mu_j) = \sqrt{\sum_{l=1}^{n}(x_{i,l} - \mu_{j,l})^2}$，其中 n 是数据点的维度）。
  - 将数据点 x_i 分配到距离它最近的聚类中心所属的聚类，即如果 $d(x_i, \mu_{j^*}) = \min_{j=1,\cdots,K} d(x_i, \mu_j)$，则 x_i 属于第 j^* 个聚类。
- 更新聚类中心：对于每个聚类 C_j（$j = 1, 2, \cdots, K$），重新计算其聚类中心 μ_j，$\mu_j = \frac{1}{|C_j|}\sum_{x_i \in C_j} x_i$，其中 $|C_j|$ 是聚类 C_j 中的数据点个数。
- 重复 “分配数据点到聚类” 和 “更新聚类中心” 这两个步骤，直到聚类中心不再发生显著变化（例如，两次迭代之间聚类中心的移动距离小于某个设定的阈值）或者达到最大迭代次数。

影响 K - Means 算法聚类性能的因素：

聚类数目 K 的选择：如果 K 选择过小，会导致聚类结果过于粗糙，不同的数据分布可能被合并到一个聚类中，无法很好地揭示数据的内在结构；如果 K 选择过大，可能会使聚类过于细碎，每个聚类中的数据点过少，甚至可能出现一些聚类只包含一个或几个离群点的情况。

数据的分布形状：当数据分布不是球形或者具有复杂的形状（如环形、螺旋形等）时，K - Means 算法基于距离的划分方式可能无法很好地拟合数据。因为它假设聚类是球形的，对于非球形的数据分布，可能会产生不合理的聚类结果。

离群点和噪声：噪声数据会干扰聚类中心的计算和数据点的分配。离群点可能会使聚类中心发生偏移，或者导致聚类算法将其单独划分为一个聚类，从而影响其他正常数据点的聚类效果。

3. 请简述谱聚类算法的原理，给出一种谱聚类算法（经典算法、Shi 算法和 Ng 算法之一）的计算步骤；请指出哪些因素会影响聚类的性能。

答：谱聚类算法建立在图论的谱图理论基础之上，其本质是将聚类问题转化为一个图上的关于顶点划分的最优问题。其核心是利用数据点之间的相似性结构，找到数据的自然分组。

Shi 算法（Normalized Cuts 谱聚类算法）计算步骤：

- 构建相似性矩阵：给定数据集 $X = \{x_1, x_2, \cdots, x_n\}$，利用高斯核函数确定相似性度量 w_{ij}。
- 计算度矩阵和拉普拉斯矩阵：
  - 计算度矩阵 D。$d_{ii} = \sum_{j=1}^{n} w_{ij}$，即对角线上的元素 d_{ii} 是相似性矩阵 W 中第 i 行元素之和。
  - 构建未归一化的拉普拉斯矩阵 $L = D - W$。然后计算归一化的拉普拉斯矩阵，采用对称归一化的拉普拉斯矩阵 $L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$；
- 特征分解：对 L_{sym} 进行特征分解，得到特征值 $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ 和对应的特征向量 $v_1, v_2, \cdots, v_n$。
- 选择特征向量和构建新空间：选择最小的非零特征值对应的 k 个特征向量 $v_1, v_2, \cdots, v_k$，将这些特征向量组成矩阵 $V = [v_1, v_2, \cdots, v_k] \in R^{n \times k}$。矩阵 V 的每一行看作是数据点在新的 k 维表示空间中的坐标。
- 聚类划分。对新空间中的数据点应用 K - Means 等聚类算法进行聚类。例如，在新空间中计算每个数据点到聚类中心的距离，将数据点分配到距离最近的聚类中心所属的聚类，最终得到聚类结果

影响聚类性能的因素：

相似性度量和核函数参数：相似性度量函数的选择直接影响图的构建和数据点之间的关联程度。不同的相似性度量可能适用于不同类型的数据。

数据分布和噪声：对于数据分布复杂（如非凸形状、存在多个密度层次等）的数据集，谱聚类算法可能需要更合适的相似性度量和参数调整才能得到好的结果。噪声数据可能会干扰相似性计算，导致错误的边权重和聚类划分。

2 编程题

题目 1

现有 1000 个二维空间的数据点，可以采用 MATLAB 代码来生成 (略)。

请完成如下工作：

1. 编写一个程序，实现经典的 K-均值聚类算法；

2. 令聚类个数等于 5，采用不同的初始值，报告聚类精度、以及最后获得的聚类中心，并计算所获得的聚类中心与对应的真实分布的均值之间的误差。

计算程序见压缩包中的 k-means 文件夹，原始数据的分布如图 1所示。

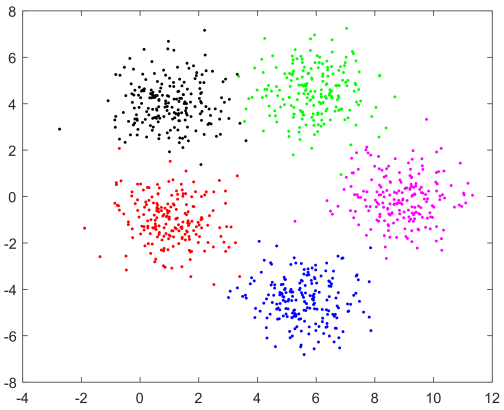

Figure 1: 原始数据的分布

设置聚类数目为 5, 随机初始化聚类中心, 运行代码, k-means 的结果如图 2所示, 图中红色点为预测的聚类中心。坐标分别为 $(1.039, -1.008), (8.885, -0.081), (5.532, -4.445)$, $(1.026, 4.000)$, $(5.924, 4.521)$。

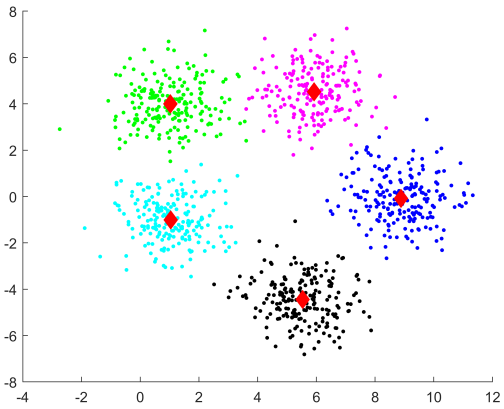

Figure 2: Caption

计算聚类中心与真实中心的差距，用平方根误差为指标，与原始聚类中心的误差分别为 $0.0396, 0.0636, 0.0258, 0.0790, 0.1405$，误差较小。

题目 2

关于谱聚类。有 200 个数据点，它们是通过两个半月形分布生成的。

1. 请编写一个谱聚类算法，实现 “Normalized Spectral Clustering—Algorithm 3 (Ng 算法)”。**2**. 请分析分别取不同的 σ 值和 k 值对聚类结果的影响。计算程序见压缩包中的 NG 文件夹，原始数据的分布如图 3所示。

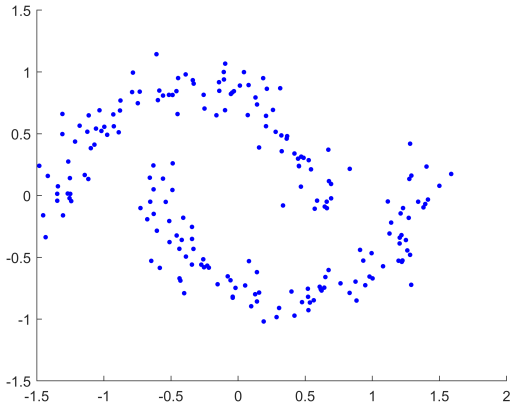

Figure 3: 原始数据分布

设置 $sigma = 0.03$，$k = 2$，编写代码实现的谱分类结果如图 4。

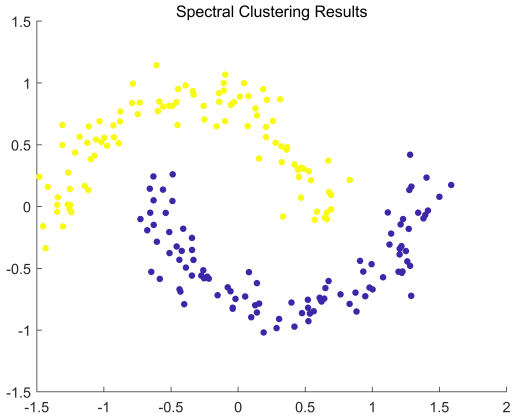

Figure 4: 谱聚类算法结果

进一步，固定 $k = 2$，改变 σ 的值，统计聚类精度，得到的结果如图 5所示。

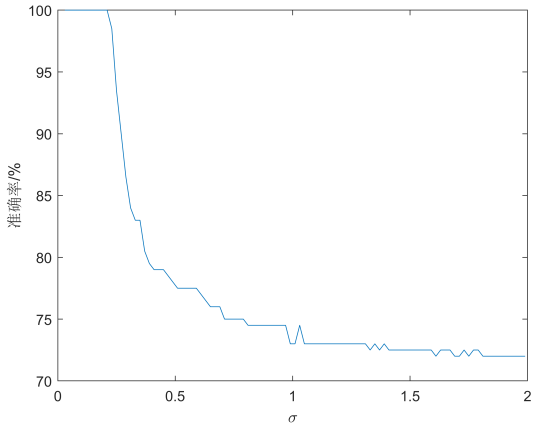

Figure 5: 改变 σ 得到的聚类精度