Answer of Assignment 5
分类器集成 + 数据聚类

2020E8017782032_ 蒲尧

第一部分：简述题

1. 请简述 adaboost 算法的设计思想和主要计算步骤。

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。

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。

第二部分：计算机编程（第 1 题必做，第 2 题选做）

1. 现有 1000 个二维空间的数据点，可以采用如下 MATLAB 代码来生成：

```
Sigma = [1, 0; 0, 1];
mu1 = [1, -1];
x1 = mvnrnd(mu1, Sigma, 200);
mu2 = [5.5, -4.5];
x2 = mvnrnd(mu2, Sigma, 200);
mu3 = [1, 4];
x3 = mvnrnd(mu3, Sigma, 200);
mu4 = [6, 4.5];
x4 = mvnrnd(mu4, Sigma, 200);
mu5 = [9, 0.0];
x5 = mvnrnd(mu5, Sigma, 200);
% Obtain the 1000 data points to be clustered
X = [x1; x2; x3; x4; x5];
% Show the data point
plot(x1(:,1), x1(:,2), 'r.'); hold on;
plot(x2(:,1), x2(:,2), 'b.');
plot(x3(:,1), x3(:,2), 'k.');
plot(x4(:,1), x4(:,2), 'g.');
plot(x5(:,1), x5(:,2), 'm.');
```

在运行完上述代码之后，可以获得 1000 个数据点，它们存储于矩阵 X 之中。X 是一个行数为 1000 列数为 2 的矩阵。即是说，矩阵 X 的每一行为一个数据点。另外，从上述 MATLAB 中可见，各真实分

布的均值向量分别为 mu1, mu2, mu3, mu4, mu5。
提示：在实验中，生成一个数据矩阵 X 之后，就将其固定。后续实验均用此数据集，以便于分析算法。
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。

2. 关于谱聚类。有如下 200 个数据点，它们是通过两个半月形分布生成的。如图所示：

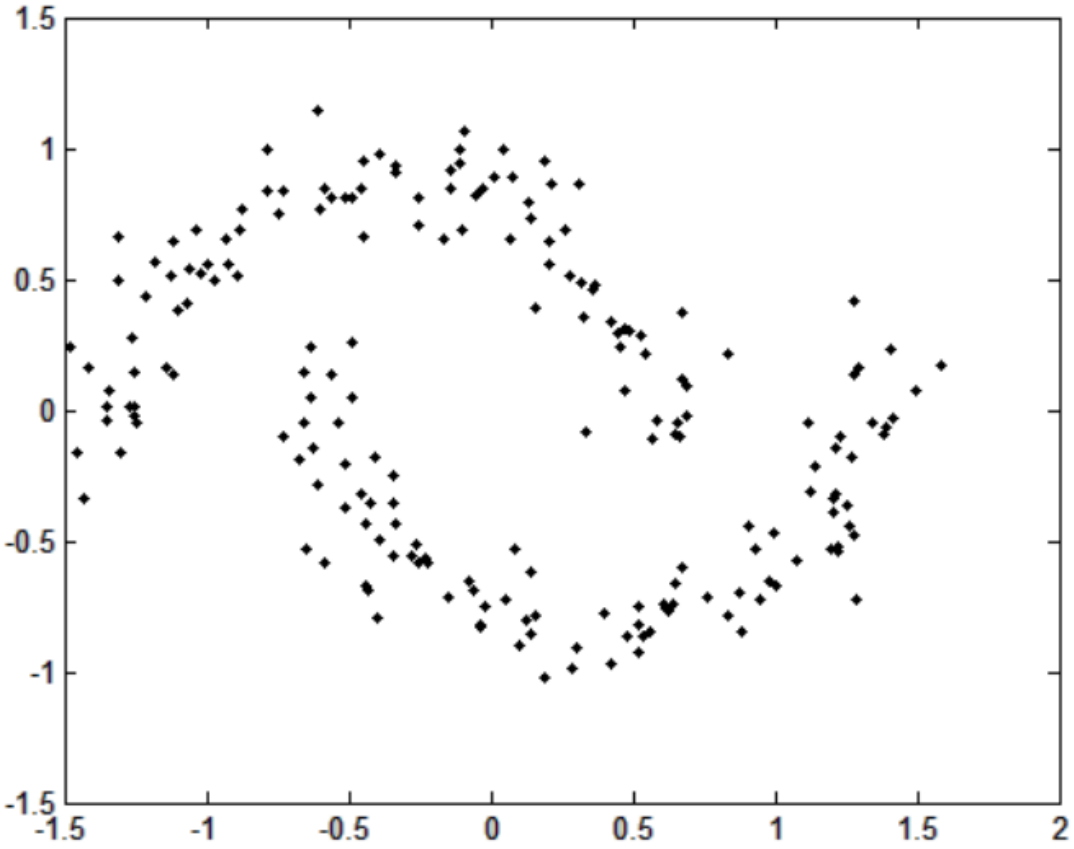

图 1: 本题数据点分布（具体附后）

(1). 请编写一个谱聚类算法，实现“Normalized Spectral Clustering—Algorithm 3 (Ng 算法)”。

(2). 设点对亲和性（即边权值）采用如下计算公式：

$$w_{ij} = exp\left(-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2}{2\sigma^2}\right)$$

同时，数据图采用 k-近邻方法来生成（即是说，对每个数据点 x_i ，首先在所有样本中找出不包含 x_i 的 k 个最邻近的样本点，然后 x_i 与每个邻近样本点均有一条边相连，从而完成图构造）。注意，为了保证亲和度矩阵 $\boldsymbol{W}$ 是对称矩阵，可以令 $\boldsymbol{W} = (\boldsymbol{W}^T + \boldsymbol{W})/2$, 其中，$\boldsymbol{W}^T$ 示表示 $\boldsymbol{W}$

的转置矩阵。假设已知前 100 个点为一个聚类，后 100 个点为一个聚类，请分析分别取不同的 σ 值和 k 值对聚类结果的影响。（本题可以给出关于聚类精度随着 σ 值和 k 值的变化曲线。在实验中，可以固定一个，变化另一个）。

附注 1： 聚类精度 Accu 计算如下：

$$Accu = \frac{n_1 + n_2}{n}$$

其中，n_1 表示正确的属于第一个聚类的样本点的个数；n_2 表示正确的属于第二个聚类的样本点的个数；n 表示样本点的总数。

附注 2： 200 个样本如下 (其中，$\boldsymbol{X}$ 的每一行代表一个数据点)：data 文件

第一部分回答

1. (1) Adaboost 算法的**设计思想**就是训练一系列弱分类器，然后组合成一个强分类器。具体有两方面的内容要做：

1) 如何改变训练数据的权重？提高前一轮弱分类器错分样本的权重，降低正确分类样本的权重；

2) 如何调整弱分类器的权重？采用加权多数表决的方法。即加大分类错误率较低的分类器的权重，减小错误率高的分类器权重。

(2) **主要计算步骤：**

1) 初始化训练数据权值分布：$D_1 = \{w_{11}, w_{12}, ..., w_{1n}\}, w_1 i = 1/n, i = 1, 2, ..., n$

2) 对于 m=1,2,...,M

2a) 学习具有权值分布 D_m 的训练数据，学习基本分类器：$G_m(\boldsymbol{X}) : \boldsymbol{X} \to \{-1, +1\}$

2b) 计算 $G_m(\boldsymbol{X})$ 在训练数据集上的分类加权错误率：$e_m = \sum_{i=1}^{n} w_{mi} \underbrace{I(G_m(\boldsymbol{X}_i \neq y_i))}_{truth \quad function}$

2c) 计算 $G_m(\boldsymbol{X})$ 的贡献系数，即第 m 个分类器的权值：$\alpha_m = 1/2 \ln \frac{1-e_m}{e_m}$，$e_m$ 越大，权重 α_m 越小

2d) 更新训练数据集的权重分布：$D_{m+1} = \{w_{m+1,1}, w_{m+1,2}, ..., w_{m+1,n}\}$

$$w_{m+1,i} = \frac{w_{mi}}{Z_m} \times exp(-\alpha y_i G_m(X_i))$$

$$Where \quad Z_m = \sum_{i=1}^{n} w_{mi} exp(-\alpha y_i G_m(X_i))$$

3) 构造基本分类器的线性组合：$f(\boldsymbol{X}) = \sum_{m=1}^{M} \alpha_m G_m(\boldsymbol{X})$。

对于二分类器：$G(\boldsymbol{X}) = sign(f(\boldsymbol{X})) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(\boldsymbol{X})\right)$。

2. (1) K-Means 聚类算法的**原理**

假设每个样本以概率为 1 属于某个类，则

$$P\left(\omega_i|x_k,\hat{\mu}\right)=\begin{cases}1, & x_k\in\omega_i;\\ 0, & x_k\notin\omega_i\end{cases}$$

$$\hat{P}\left(\omega_i\right)=n_i/n$$

$$\hat{\mu}_i=\frac{1}{n_i}\sum_{k=1}^{n_i}x_k^i$$

$$\hat{\Sigma}_i=\frac{1}{n_i}\sum_{k=1}^{n_i}\left(x_k^i-\hat{mu}_i\right)\left(x_k^i-\hat{mu}_i\right)^T$$

K-Means 聚类算法是一种基于距离的聚类算法，即先随机给出类中心，通过不断计算距离归类和重新计算类中心的迭代最终确定趋于正确的类中心。

(2) K-Means 聚类算法的计算步骤

先引入两个假设：

1) 各类出现的先验概率均相等

2) 每个样本以概率 1 属于某个类（即后验概率 0-1 近似）

步骤：

1) 根据给定的 K 值划分，可以随机选择 K 个点作为类中心；

2) 计算每个点到各个类中心的距离，将该点归到距离最近的类中心的 Cluster 类下；

3) 重新计算每个 Cluster 的类中心；

4) 重复 2 和 3，直到类中心在某个精度范围不变化或者到达最大迭代次数。

(3) **影响因素**

-必须事先给定簇的个数，且对初始值敏感；

-不适合于发现非凸曲面的簇以及大小相差很大的簇；

-对噪声、孤立数据点、野点很敏感。

3. (1) 谱聚类算法的**原理**

-从图切割的角度，聚类就是要找到一种合理的分割图的方法，分割后能形成若干个子图。连接不同子图的边的权重尽可能小，子图内部边权重尽可能大。

-谱聚类算法建立在图论中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。

-谱聚类算法建立在点对亲和性基础之上，理论上能对任意分布形状的样本空间进行聚类。

-算法的核心是将原始的数据点 x_i 转换为在特征空间的数据点 y_i ，在新的空间对原始数据进行描述

(2) 谱聚类算法的**计算步骤**

-利用点对之间的相似性，构建亲和度矩阵 $\mathbb{W}$；

-构建拉普拉斯矩阵：

- Classical: $\mathbb{L}=\mathbb{D}-\mathbb{W}$
- Shi's Algorithm: $\mathbb{L}_{rw}=\mathbb{D}^{-1}\mathbb{L}$
- Ng's Algorithm: $\mathbb{L}_{sym}=\mathbb{D}^{-1/2}\mathbb{L}\mathbb{D}^{-1/2}$

-求解拉普拉斯矩阵最小的 k 个特征值$\lambda_1, \lambda_2, ..., \lambda_k$ 对应的特征向量$\vec{u}_1, \vec{u}_2, ..., \vec{u}_k$（通常舍弃零特征所对应的分量全相等的特征向量）；

- Classical: $\mathbb{L}\vec{u} = \lambda\vec{u}$
- Shi's Algorithm: $\mathbb{L}_{rw}\vec{u} = \lambda\mathbb{D}\vec{u}$
- Ng's Algorithm: $\mathbb{L}_{sym}\vec{u} = \lambda\vec{u}$

-由这些特征向量构成样本点的新特征 (Ng 的特征向量需要归一化)，大小 n×k，采用 K-means 等聚类方法完成最后的聚类。

(3) **影响因素**

相似度的计算方法，拉普拉斯矩阵、特征向量的归一化方法，局部链接数目，聚类数目和方法，等等。

第二部分回答

1. 代码见如下文件：MATLAB 文件 1 主程序，MATLAB 文件 2KMeans 函数，Kmeans 的 Python 实现。效果图如下，十字是最初算得的类中心，颜色与类别有差异；圆圈是重排序之后的类中心，颜色与类别无差异。由实验结果可知，初始化的均值会影响最终的类中心收敛，下面的第一幅图显然收敛到了两类的中间，而第二幅图由于初始的均值设置合理聚类准确度较高。说明一定的先验知识对聚类很有帮助：

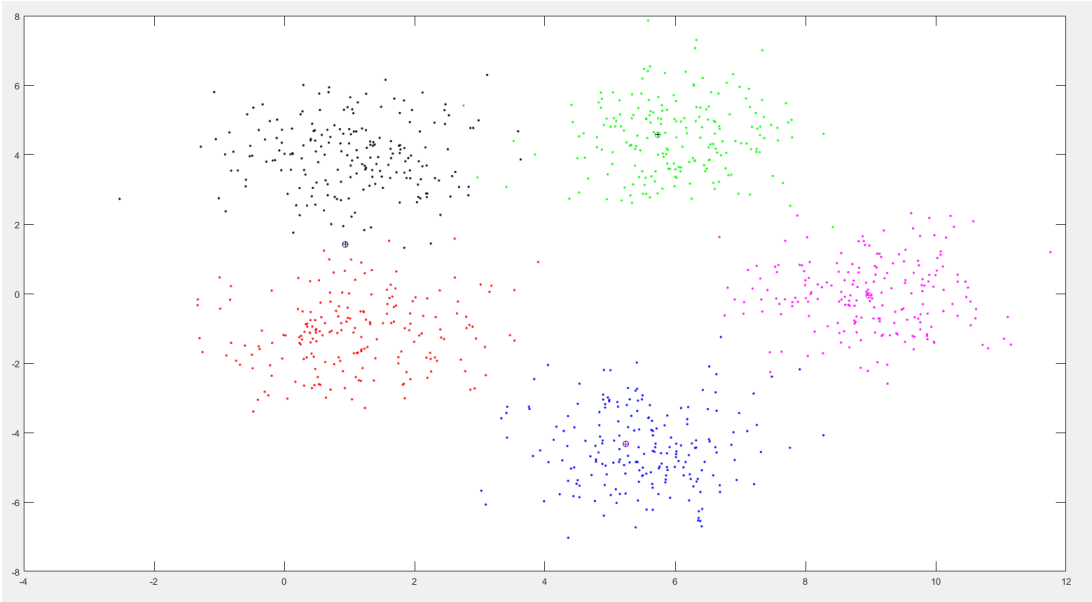

图 2: 聚类准确率 77.8%，均方误差 1.1389 图像

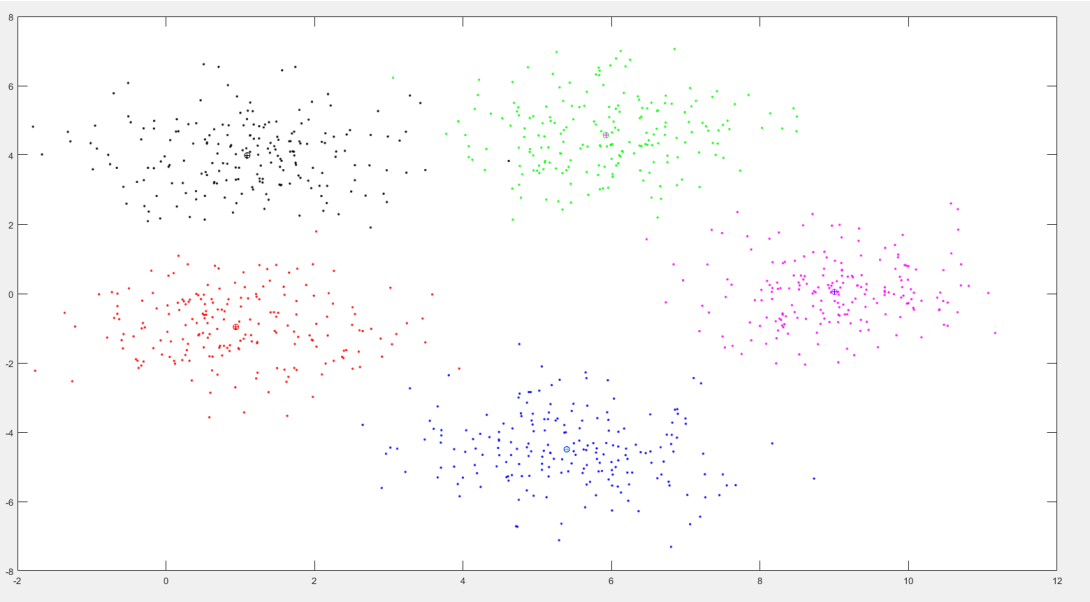

图 3: 聚类准确率 99.5%，均方误差 0.0121 图像

2. 代码见如下文件: Ng 算法谱聚类的 Python 实现。效果图，准确度随着 σ 和 k 变化的曲线如下：

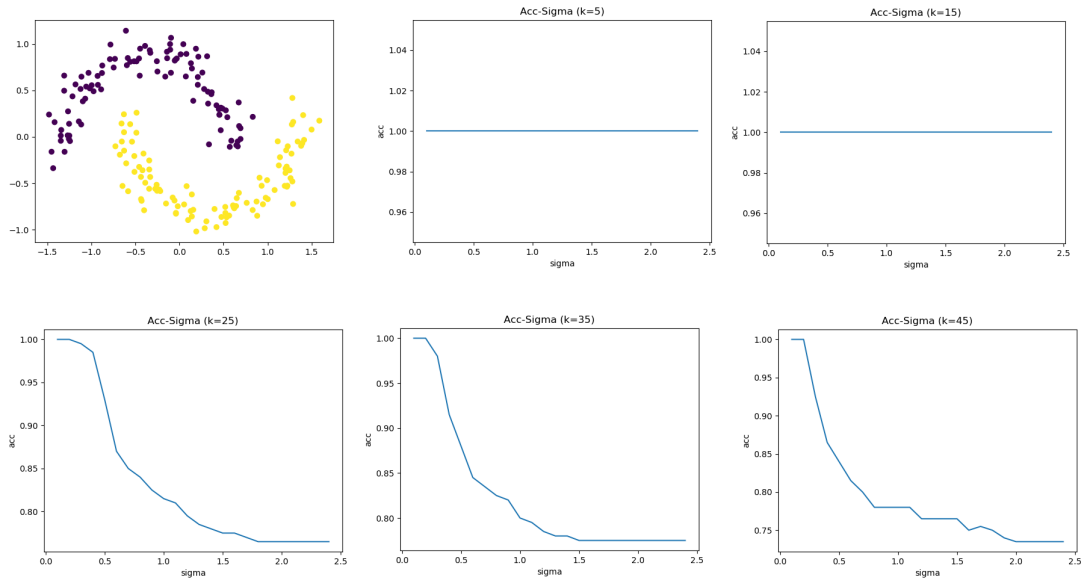

图 4: Ng 算法谱分类效果图