第八章：数据聚类 (第五次作业)

第一部分：问题、计算与证明

1. 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。
2. 请给出谱聚类算法（经典算法、Shi 算法和 Ng 算法）的计算步骤；请指出哪些因素会影响聚类性能。
3. 设有四个样本：$\mathbf{x}_1 = (4,5)^T$，$\mathbf{x}_2 = (1,4)^T$，$\mathbf{x}_3 = (0,1)^T$，$\mathbf{x}_4 = (5,0)^T$。现有如下三种划分：
   (1) $\omega_1 = \{\mathbf{x}_1, \mathbf{x}_2\}$，$\omega_2 = \{\mathbf{x}_3, \mathbf{x}_4\}$；
   (2) $\omega_1 = \{\mathbf{x}_1, \mathbf{x}_4\}$，$\omega_2 = \{\mathbf{x}_2, \mathbf{x}_3\}$；
   (3) $\omega_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$，$\omega_2 = \{\mathbf{x}_4\}$。
   证明：对于平方误差和准则，第三种划分最好；若采用类内散度矩阵的行列式最小准则（即 $\min |\mathbf{S}_w|$），则前两种划分较好。

第二部分：计算机编程

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。

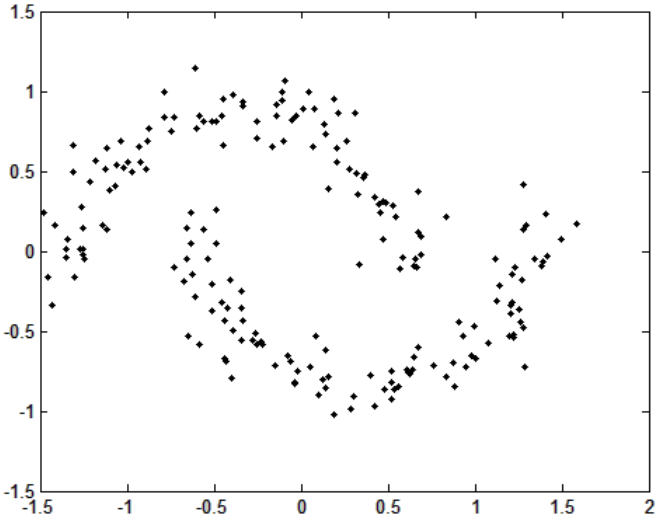
提示：在实验中，生成一个数据矩阵 X 之后，就将其固定。后续实验均用此数据集，以便于分析算法。

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，采用不同的初始值观察最后的聚类中心，给出你所估计的聚类中心，指出每个中心有多少个样本；指出你所得到聚类中心与对应的真实分布的均值之间的误差（对 5 个聚类，给出均方误差即可）。

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：



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：

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

同时，数据图采用 k-近邻方法来生成（即是说，对每个数据点 $\mathbf{x}_i$，首先在所有样本中找出不包含 $\mathbf{x}_i$ 的 k 个最邻近的样本点，然后 $\mathbf{x}_i$ 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 W 是对称矩阵，可以令 $W=(W^T+W)/2$，其中，W^T 表示 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，**X** 的每一行代表一个数据点)：

X = [-1.3046 -0.1606
-1.4341 -0.3372
-1.3475 -0.0421
-1.3426 0.0746
-1.2433 -0.0451
-1.3477 0.0140

-1.2695	0.0159
-1.2560	-0.0217
-1.4525	-0.1600
-1.4813	0.2405
-1.2533	0.1410
-1.1160	0.1336
-1.2506	0.0134
-1.4179	0.1585
-1.1427	0.1662
-1.2654	0.2749
-1.3093	0.4972
-0.8894	0.5117
-1.2147	0.4365
-1.0589	0.5416
-1.0974	0.3831
-1.1807	0.5650
-1.0185	0.5226
-1.1234	0.5162
-0.9260	0.5602
-1.1131	0.6486
-1.0704	0.4117
-1.3090	0.6600
-0.9298	0.6564
-0.7473	0.7469
-0.8765	0.7687
-1.0327	0.6897
-0.8800	0.6879
-0.7897	0.8371
-0.9968	0.5561
-0.7329	0.8403
-0.9754	0.4914
-0.5870	0.8492
-0.7836	0.9945
-0.5976	0.7718
-0.6082	1.1435
-0.5593	0.8088
-0.4512	0.6593
-0.4470	0.9503
-0.5153	0.8144
-0.4882	0.8137
-0.3907	0.9800
-0.3310	0.9054
-0.4595	0.8447
-0.2568	0.8148

-0.1615	0.6493
-0.1068	0.9994
-0.3376	0.9333
-0.0439	0.8333
-0.0312	0.8441
-0.1402	0.8472
-0.0961	1.0671
0.0436	0.9985
-0.2503	0.7038
-0.1422	0.9170
-0.1054	0.9390
-0.0536	0.8220
0.3137	0.8682
0.0720	0.6508
0.0782	0.8938
-0.0973	0.6904
0.1411	0.7367
0.1296	0.7934
0.2079	0.5609
0.2143	0.8637
0.0133	0.8896
0.2598	0.6927
0.1868	0.9497
0.1563	0.3887
0.2798	0.5154
0.2081	0.6446
0.3619	0.4604
0.4530	0.2400
0.3655	0.4805
0.4878	0.3051
0.3224	0.4864
0.4199	0.3399
0.4536	0.2371
0.4703	0.3144
0.3250	0.3584
0.4466	0.2982
0.6724	0.3708
0.5432	0.2119
0.6769	0.1165
0.5262	0.2860
0.8319	0.2159
0.6931	-0.0227
0.3347	-0.0805
0.6928	0.0936

0.4681	0.0717
0.6455	-0.0896
0.6603	-0.0498
0.6617	-0.1023
0.5885	-0.0414
0.5709	-0.1069
-0.5597	0.1367
-0.6311	0.2434
-0.4894	0.0448
-0.6578	0.1441
-0.4873	0.2604
-0.5392	-0.0516
-0.6547	-0.0502
-0.6288	-0.1486
-0.7275	-0.1015
-0.6298	0.0501
-0.3436	-0.2539
-0.6768	-0.1921
-0.6052	-0.2849
-0.4088	-0.1798
-0.5129	-0.2064
-0.3433	-0.3513
-0.5111	-0.3768
-0.4578	-0.3235
-0.3441	-0.5589
-0.4199	-0.3592
-0.3887	-0.4941
-0.6470	-0.5287
-0.4376	-0.4311
-0.2540	-0.5804
-0.5833	-0.5855
-0.2180	-0.5837
-0.3327	-0.4320
-0.4364	-0.6702
-0.4299	-0.6884
-0.4017	-0.7903
-0.2743	-0.5594
-0.2300	-0.5676
-0.0773	-0.6538
-0.0194	-0.7473
-0.1527	-0.7179
-0.2595	-0.5159
-0.0402	-0.8188
0.1271	-0.7987

0.0809	-0.5309
0.1404	-0.8573
0.1402	-0.6199
-0.0565	-0.6858
0.0977	-0.8960
0.1566	-0.7853
0.2867	-0.9838
0.1906	-1.0191
0.0526	-0.7277
-0.0401	-0.8271
0.3977	-0.7768
0.5240	-0.9277
0.3055	-0.9099
0.5350	-0.8659
0.5637	-0.8476
0.4200	-0.9719
0.5199	-0.7543
0.6421	-0.7453
0.6160	-0.7585
0.4763	-0.8627
0.5213	-0.8194
0.6220	-0.7689
0.6079	-0.7382
0.8823	-0.8498
0.6742	-0.6028
1.0800	-0.5724
0.7631	-0.7106
0.8309	-0.7879
0.9476	-0.7258
0.6503	-0.6596
0.9792	-0.6558
0.9327	-0.5266
0.9087	-0.4398
1.2249	-0.5257
1.0020	-0.6705
0.8749	-0.6966
0.9947	-0.4664
1.2899	-0.7225
1.2173	-0.3204
1.1954	-0.5276
1.2815	-0.4800
1.2609	-0.4442
1.2198	-0.5365
1.2043	-0.3410

1.2503 -0.3602

1.2035 -0.3879

1.3811 -0.0959

1.2714 -0.1807

1.4145 -0.0326

1.1280 -0.3075

1.3946 -0.0686

1.1405 -0.2190

1.2109 -0.1448

1.4998 0.0787

1.1156 -0.0481

1.5878 0.1745

1.2817 0.4193

1.3417 -0.0501

1.2290 -0.1015

1.4039 0.2343

1.2912 0.1612

1.2759 0.1342];