

计算机与信息学院实验报告

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实验课程： | 模式识别 | | | | |
| 实验编号： | 2 | | | | |
| 实验名称： | 实验2：聚类算法实验 | | | | |
| 实验人员： | 年级 | | 2018级 | | |
| 专业 | | 计算机科学与技术 | | |
| 学号 | | 18111207248 | | |
| 姓名 | | 吴钰 | | |
| 实验日期： | 2021.4.10 | | | | |
| 上交日期： | 2021.4.24 | | | | |
| 实 验 室： | 2060302 | | | | |
| 实验评价： |  | | | | |
|  | 实验成绩： |  | | 评定日期： |  |
|  | 指导教师： | 郑明 | | | |

一、实验目的

使用MATLAB实现聚类算法相关实验

二、实验环境

MATLAB2012

三、实验内容

1. 5个点组成的样本集X={[1,0],[0,1],[1,1],[2,1],[6,6]}，请将这5个样本分为2个类别*ω1*，*ω2*，即c=2。为了简化实验，在先验知识预判下，在此只考虑2个类别中的样本数分别为1个和4个的分类情况。编写MATLAB程序，遍历这5种情况，找到使得误差平方和最小的聚类情况，并对实验结果进行分析。
2. 编写近邻聚类算法，距离采用欧氏距离。采用二维特征空间中的10个样本对程序进行验证，X={[0,0],[3,8],[2,2],[1,1],[5,3],[4,8],[6,3],[5,4],[6,4],[7,5]}，通过不停测试阈值*T*的范围（*T*<1,1<*T*<10,*T*>10）,通过plot函数对聚类结果进行画图分析距离阈值*T*的变化和聚类效果之间的关系。

3、给出6个五位模式样本*X*1=[0,3,1,2,0]，*X*2=[1,3,0,1,0]，*X*3=[3,3,0,0,1]，*X*4=[1,1,0,2,0]，*X*5[3,2,1,2,1]，*X*6=[4,1,1,1,0]，并对这6个样本按照最短距离准则进行层次聚类。

4、使用MATLAB编程根据聚类准则函数实现C均值聚类算法，并分别对以下两个模式样本集X1和X2分别进行聚类分析，并通过调整不同的聚类数目c分析其与聚类效果之间的关系。X1={[2.4,3.8,4.9,4.7], [3.2,10.2,100.4,12.4], [4.6,7.8,8.9,9.0], [10.2,11.2,23.4,24.6]

X1={[1.2,3.4,5.6,7.8], [2.3,4.5,6.7,7.8], [11.2,12.3,12.4,14.5], [2.3,4.5,5.6,7.8]

四、实验设计

根据题意，本文的主要任务是：聚类算法实验。其中第1题的解决思路是利用误差平方和准则，算出2个聚类类别的全部模式样本与其相应类别模式均值之间的误差平方和；第2题的解决思路是利用近邻聚类算法：任取样本作为第一个聚类中心的初始值，计算样本到聚类中心的欧氏距离，不断比较更新聚类中心，直到N个样本都分类；第3题的解决思路是根据最短距离准则；第四题的解决思路是采用距离作为相似性指标，从而发现给定数据集中的K个类，且每个类的中心是根据类中所有值的均值得到，每个类用聚类中心来描述。对于给定的一个包含n个d维数据点的数据集X以及要分得的类别K,选取欧式距离作为相似度指标，聚类目标是使得各类的聚类平方和最小，聚类中心为对应类别中各数据点的平均值。

五、实验结果

5.1 实验代码：

要求必要时对代码进行注释。

第一题：

x=[[1,0];[0,1];[1,1];[2,1];[6,6]]; %样本集

J = [];

for i = 1: 5 %遍历五种情况

M(1, :) = sum(x);

M(1, :) = (M(1, :) - x(i, :)) / 4;

M(2, :) = x(i, :);

y = x;

y(i, :) = [];

J(i) = sum(sum((y - M(1, :)).^2));%聚类准则函数

end

disp(J);

第二题：

NearNeighborsCluster.m

function [result, C] = NearNeighborsCluster(X, Z, t) %近邻聚类算法( x:样本,z:聚类中心,t:阈值)

result = Z;%初始聚类中心加入result矩阵

C = [1, 1];

%将聚类中心从样本中去除

for i = 1: size(X, 1)

if isequal(X(i, :), Z)

X(i, :) = [];

break;

end

end

while size(X, 1) > 0

flag = 1; %判断是否创建新的聚类中心

x = X(1, :); %样本向量

d = []; %欧氏距离

for i = 1: size(C, 1) %遍历聚类中心

z = result(C(i, 1), :);

if norm(x - z) <= t %该样本属于第i类

d = [d; [i, norm(x - z)]];

flag = 0;

end

end

if flag

result = [result; x];

row = C(end) + 1;

C = [C; [row, row]];

else

[value, index] = min(d(:, 2));

index = d(index, 1); %找到最小欧式距离所在的类

%将样本插入result数组

row = C(index, 2); %待插入的行

result = [result(1: row, :); x; result(row + 1 : end, :)];

C(index, 2) = C(index, 2) + 1;

for i = index + 1: size(C, 1)

C(i, 1) = C(i, 1) + 1;

C(i, 2) = C(i, 2) + 1;

end

end

X(1, :) = [];

end

T2.m

X = [[0, 0]; [3, 8]; [2, 2]; [1, 1]; [5, 3]; [4, 8]; [6, 3]; [5, 4]; [6, 4]; [7, 5]]; %样本集

Z = [3,8]; %任取样本作为聚类中心初始值

[result, class] = NearNeighborsCluster(X, Z, 1);

%colormap;

% a three-column matrix of RGB triplets for each interval:

color = [1 0 0;

0 1 0;

0 0 1;

0.8500 0.3250 0.0980;

1 0 1;

1 1 0;

0 0 0;

0.4940 0.1840 0.5560;

0 0.4470 0.7410;

0 1 1;

];

for i = 1: size(class, 1) %可视化

Start = class(i, 1);

End = class(i, 2);

x = result(Start: End, 1);

y = result(Start: End, 2);

plot(x, y, 'color', color(i, :), 'Marker', '.', 'LineStyle', 'none', 'MarkerSize', 20);

hold on;

end

第三题：

T3.m

X1=[0,3,1,2,0];

X2=[1,3,0,1,0];

X3=[3,3,0,0,1];

X4=[1,1,0,2,0];

X5=[3,2,1,2,1];

X6=[4,1,1,1,0];

X = [X1;X2;X3;X4;X5;X6];

T = 2.3;

result = HierarchicalCluster(X,T);

for i = 1:length(result)

disp(i)

disp(result{i})

disp(X(result{i},:))

end

HierarchicalCluster.m

function result = HierarchicalCluster(X,T)

%最短距离

n = length(X);

D = zeros(n,n);

for i = 1:n

for j = 1:n

if i>=j

D(i,j) = norm(X(i,:)-X(j,:));

end

end

end

D(D==0) = inf;

result = cell(1,n);

for i = 1:n

result{i}= [i];

end

for i = 1:n

min\_D = min(min(D));

if(min\_D>T)

break;

end

[r,c]=find(D==min\_D);

D(r,:) = [];

result{r} = [result{c},result{r}];

result(c) = [];

D(:,c) = min(D(:,c),D(:,r));

D(:,r) = [];

end

end

第四题

C\_means.m

function [result,C] = C\_means(X,c)

clear new\_C;

n = length(X);

if c>length(X)

C = X;

result = [1:c];

return;

end

% 在样本集合中选择C个点作为初始类中心；

% 在剩下的样本点中选择一个，计算其到各个中心点的距离，选取距离最短者将其归为那个类别；

for i = 1:c

C(i,:) = X(i,:);

end

new\_C = [];

%选择下一个样本，重复直到计算完所有样本

while ~isequal(new\_C,C)

if ~isempty(new\_C)

C = new\_C;

new\_C = [];

end

for i = 1:n

for j = 1:c

d(j) = norm(X(i,:)-C(j,:));

end

[~,index] = min(d);

result(i) = index;

end

for i =1:c

new\_C(i,:) = mean(X((result==i),:),1);

end

end

end

T4.m

X1=[[2.4,3.8,4.9,4.7]; [3.2,10.2,100.4,12.4]; [4.6,7.8,8.9,9.0]; [10.2,11.2,23.4,24.6]];

X2=[[1.2,3.4,5.6,7.8]; [2.3,4.5,6.7,7.8]; [11.2,12.3,12.4,14.5]; [2.3,4.5,5.6,7.8]];

for c = 1:10 %计算完所有样本

[result,C] = C\_means(X1,c);

disp(result);

disp(C);

[result,C] = C\_means(X2,c);

disp(result);

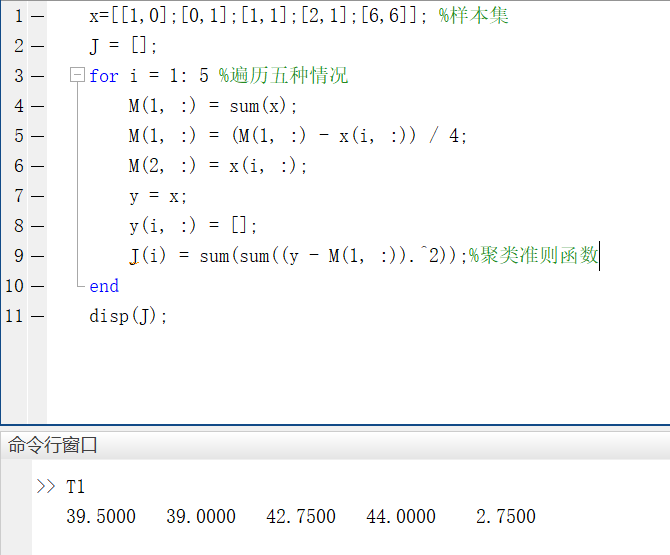
disp(C);

end

5.2 结果展示：

实验结果包括输出数据、截图等，将实验文档截图附在此处。

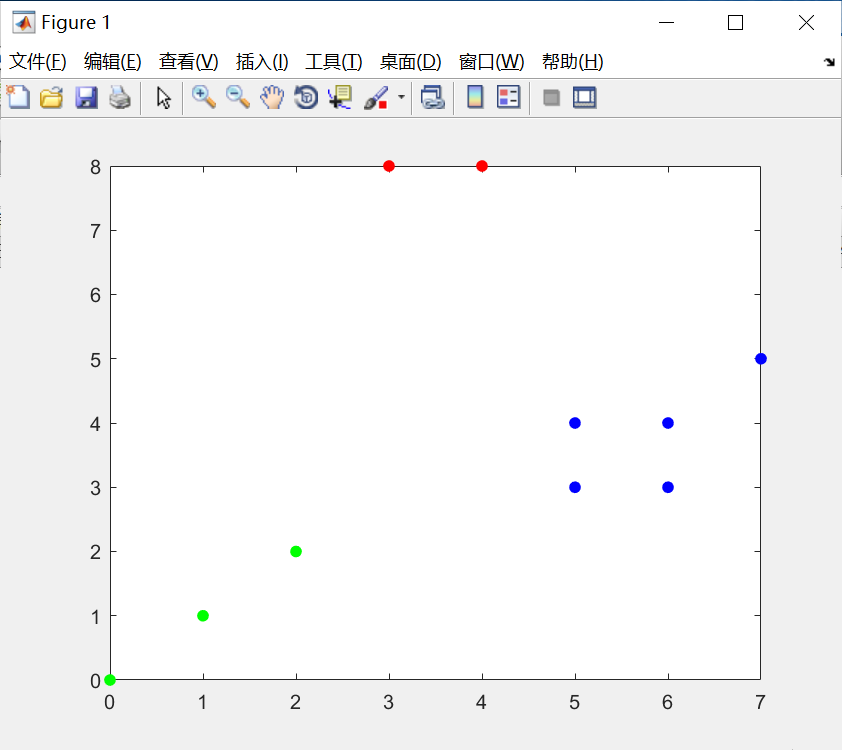
第一题：



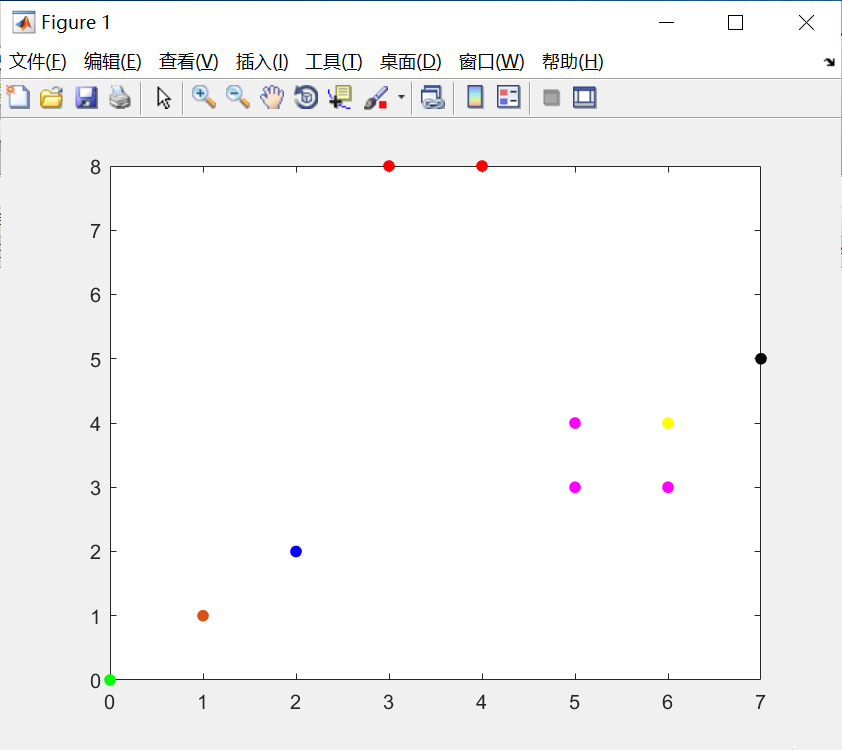
误差平方和最小为2.75，即*ω1* 为[6,6]，*ω2* 为[1,0],[0,1],[1,1],[2,1]时误差平方和最小，聚类效果最好

第二题

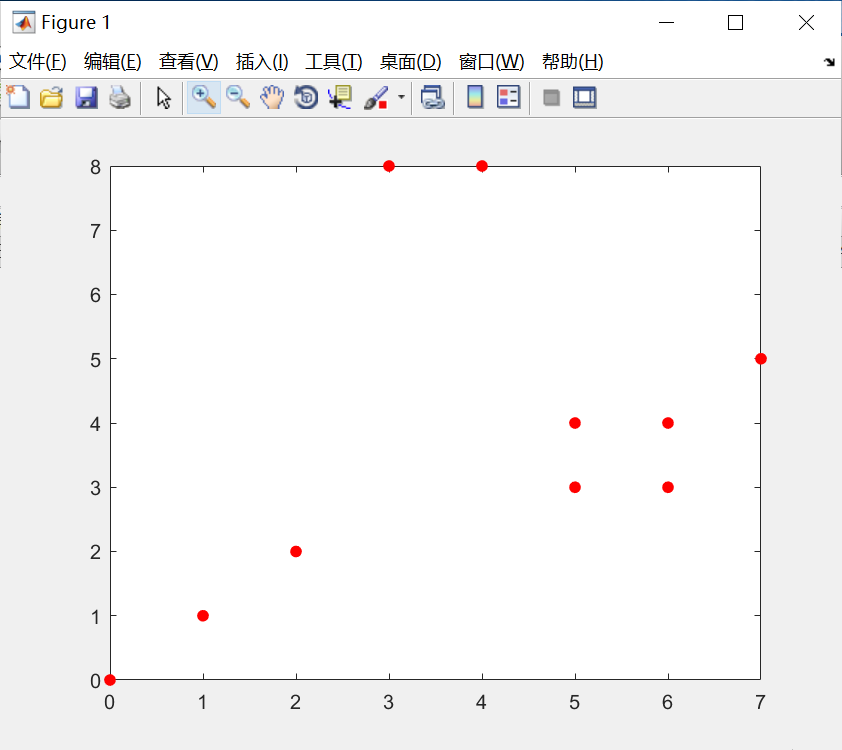
阈值t为3、4、5时聚类效果最好



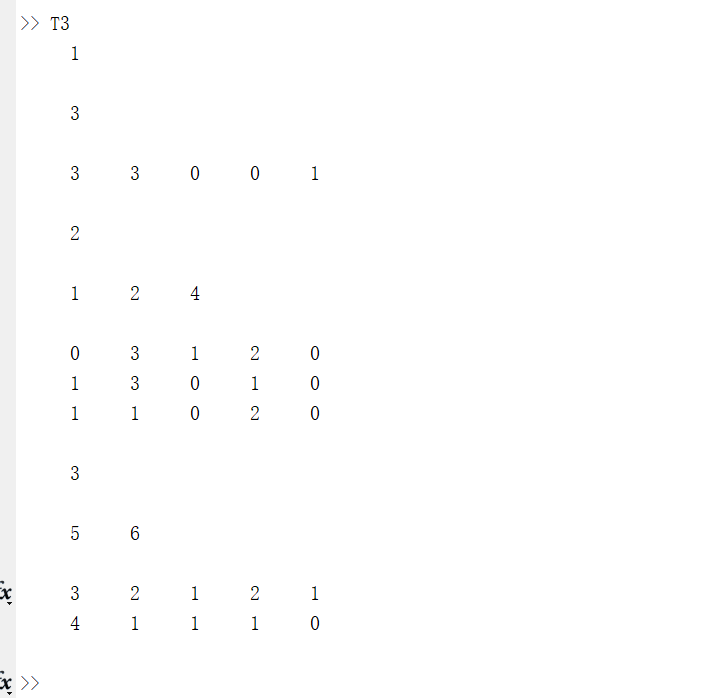
t=1时



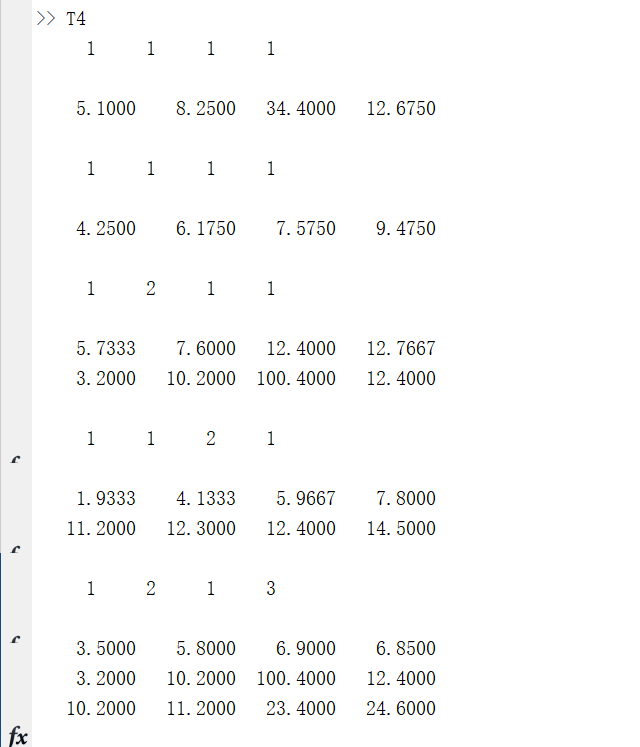
t=10时

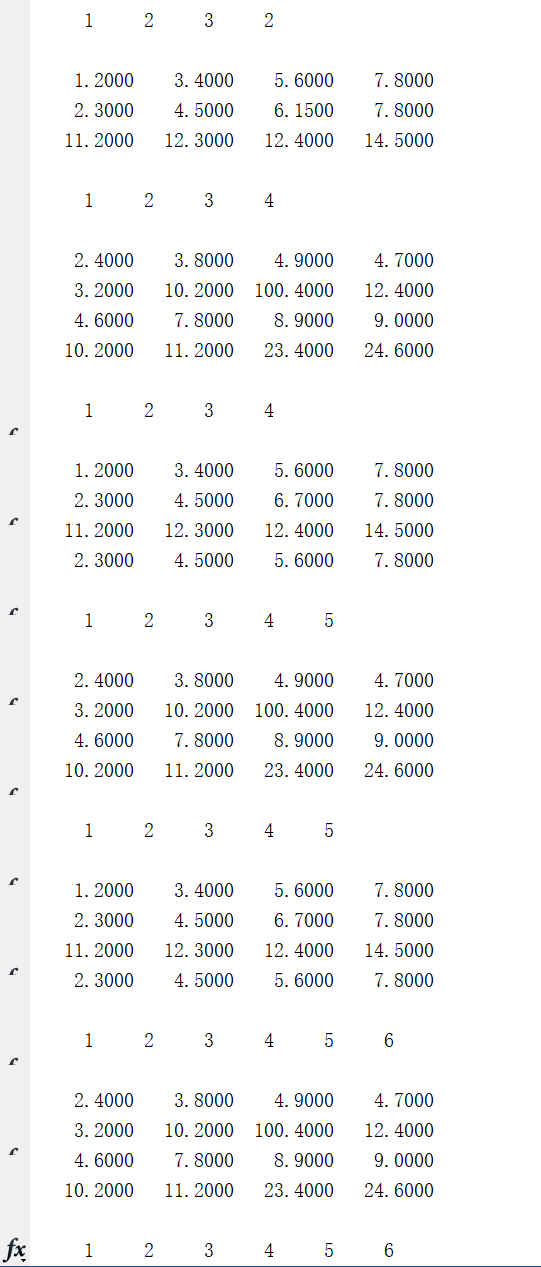


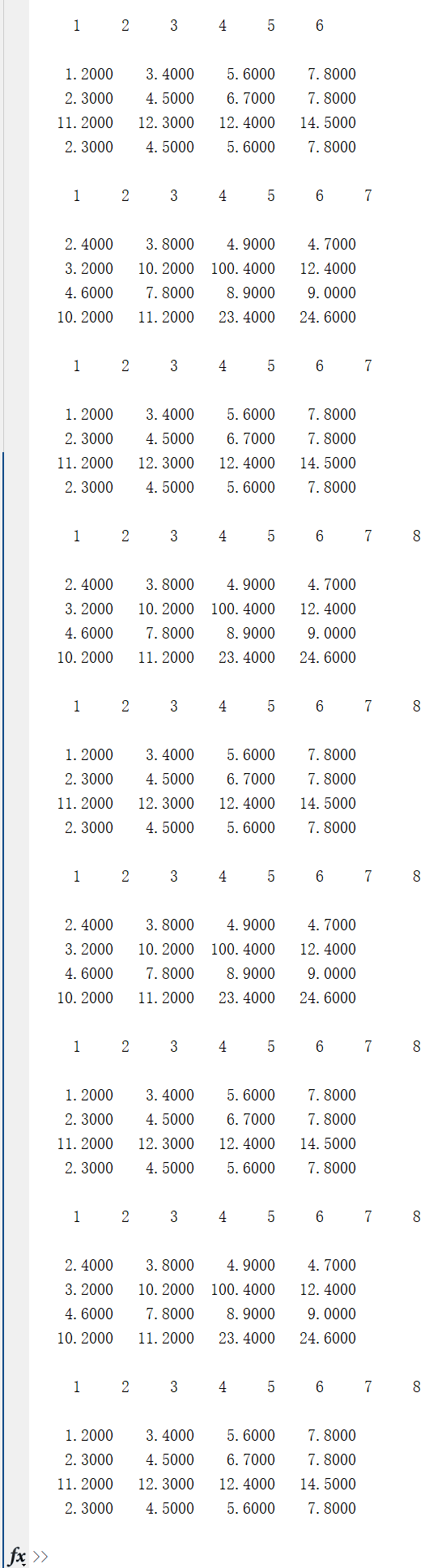
第3题



第4题







六、实验总结

本次实验主要涉及聚类算法实验的操作/内容。

6.1) 在第三题中，我碰到了对于用matlab作图不熟悉问题，通过查找matlab官网资料，找到colormap函数进行解决。

6.2) 此次实验让我对于聚类的概念更加的清晰聚类就是按照某个特定标准(如距离准则)把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。即聚类后同一类的数据尽可能聚集到一起，不同数据尽量分离。

6.3) 聚类算法一般有五种方法，最主要的是划分方法和层次方法两种。层次聚类由不同层次的分割聚类组成，层次之间的分割具有嵌套的关系。它不需要输入参数，这是它优于分割聚类 算法的一个明显的优点，其缺点是终止条件必须具体指定。