

模式识别大作业

--k 均值

班级：1602051

学号：16020510072

姓名：王方颖

一、题目

已知：

数据：sonar, iris

方法：k 均值

任务：编程实现采用 k 均值对 sonar 数据，和 iris 数据聚类。

二、k 均值聚类

1、动态聚类算法：

对数据聚类的优化过程是从“不合理的”划分到“最佳”划分，是一个动态的迭代过程。

要点：

选定某种距离度量作为样本间的相似性度量。

确定样本合理的初始分类，包括聚类中心的选择等。

确定某种评价聚类结果质量的准则函数，用以调整初始分类直至达到该准则函数的极值。

2、k-均值聚类模型

将 N 个样本 $\{x_1, \dots, x_N\}$ 划分到 k 个类 $\{C_1, \dots, C_k\}$ 中， k 为一正整数

目标：使得各个数据与其对应聚类中心点的误差平方和最小

$$J = \sum_{i=1}^k J_i = \sum_{i=1}^k \sum_{x \in C_i} \|x - m_i\|^2$$

- J_i 为第 i 类聚类的目标函数

- k 为聚类个数

- x 是划分到类 C_i 的样本

- m_1, \dots, m_k 是类 C_1, \dots, C_k 的质心（均值向量）

$$m_i = \frac{1}{N_i} \sum_{x \in C_i} x$$

3、K 均值聚类算法流程：

不合理-----> “最佳” 划分

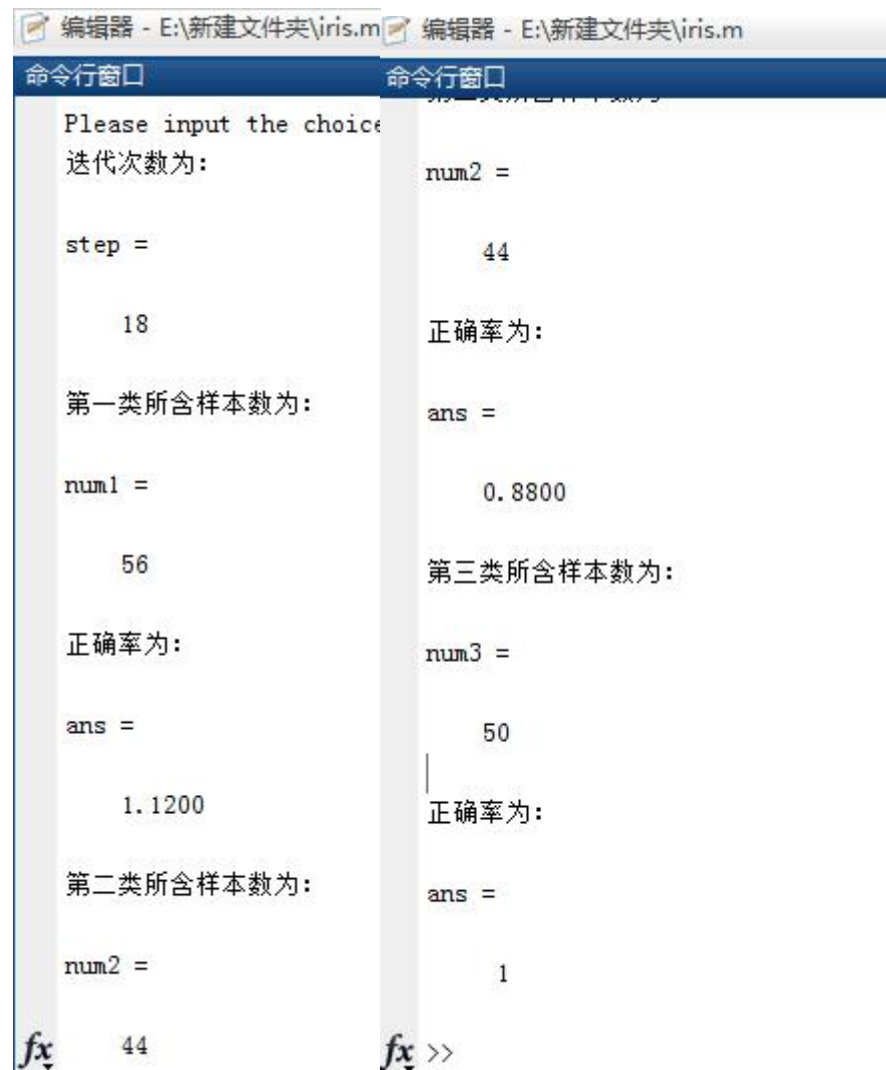
Step 1: 初始化：随机选择 k 个样本点，并将其视为各聚类的初始中心 m_1, \dots, m_k ；

Step 2: 按照最小距离法则逐个将样本 x 划分到以聚类中心 m_1, \dots, m_k 为代表的 k 个类 C_1, \dots, C_k 中；

Step 3: 计算聚类准则函数 J ，重新计算 k 个类的聚类中心 m_1, \dots, m_k ；

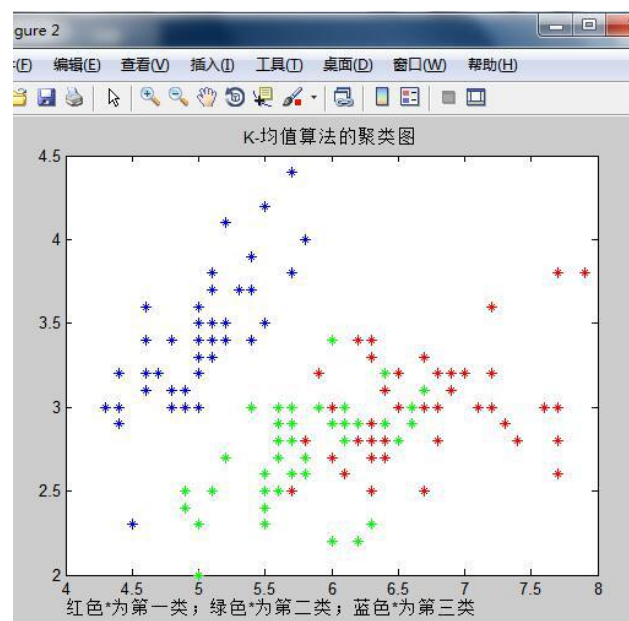
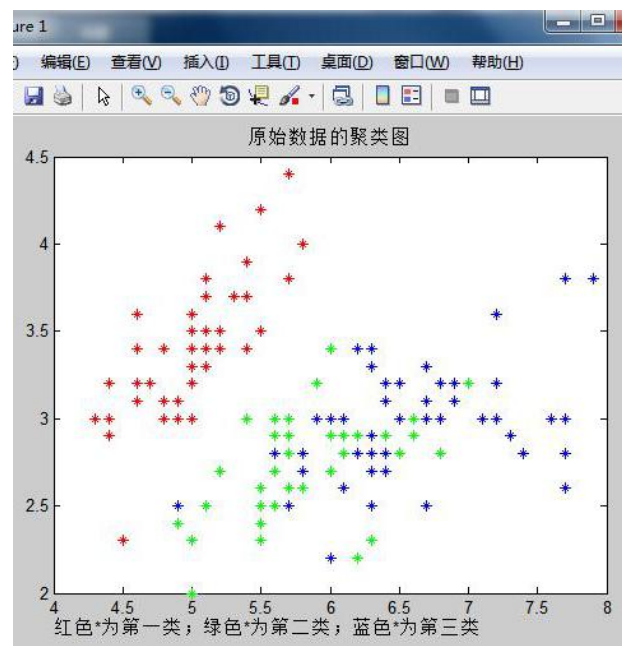
Step 4: 重复 Step 2 和 Step 3 直到聚类中心 m_1, \dots, m_k 无改变或目标函数 J 不减小。

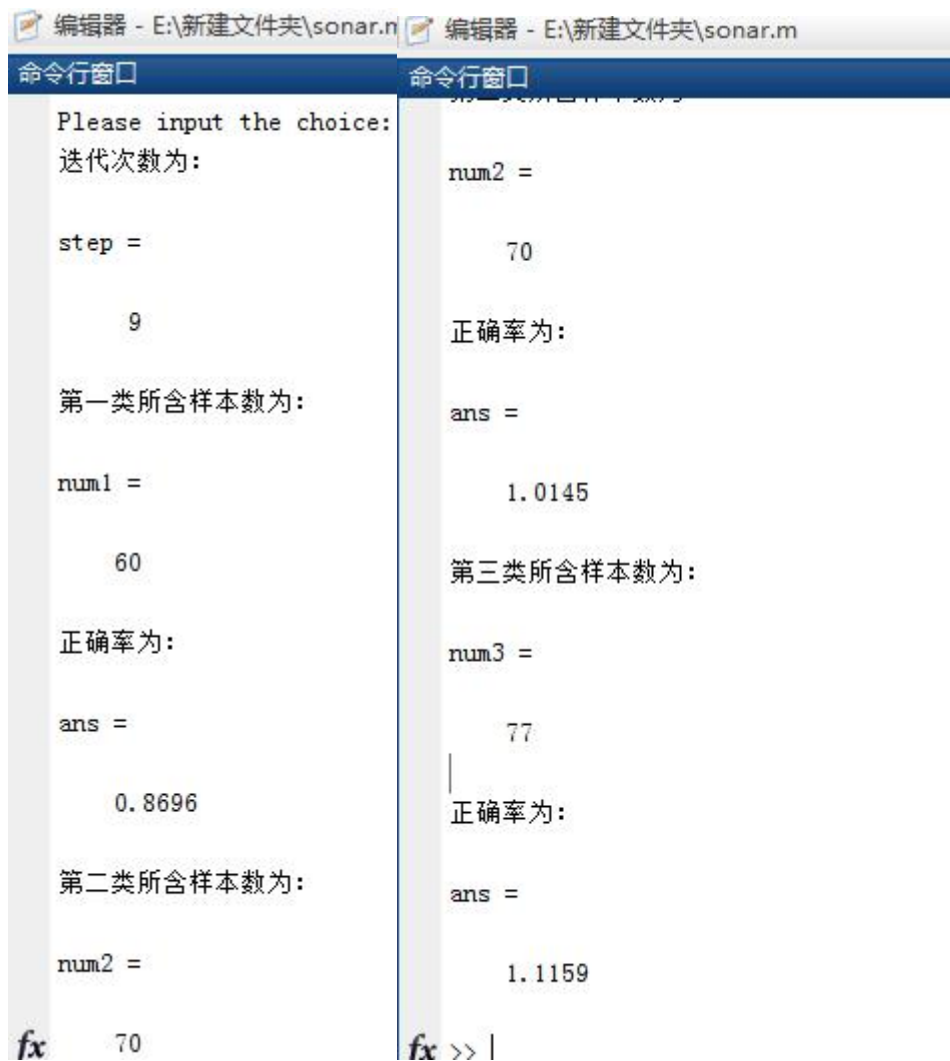
三、实验结果



The image displays two side-by-side MATLAB command window screenshots. The left window shows the script's execution flow with prompts and user input. The right window shows the corresponding MATLAB assignments and their results.

Left Window (Prompts/Inputs)	Right Window (MATLAB Code/Outputs)
Please input the choice	
迭代次数为:	num2 =
step =	44
18	正确率为:
第一类所含样本数为:	ans =
num1 =	0.8800
56	第三类所含样本数为:
正确率为:	num3 =
ans =	50
1.1200	正确率为:
第二类所含样本数为:	ans =
num2 =	1
44	





四、代码

Sonar 数据:

```
clc;  
clear;  
close all;  
%读入 sonar 数据  
x1 = xlsread('111.xlsx');  
x2 = xlsread('222.xlsx');  
x3 = xlsread('333.xlsx');  
X = [x1;x2;x3];
```

```
choice = input('Please input the choice: (1 为聚类中心在第一类, 2 为聚类中心在第二类, 3 为聚类中心在第三类, 4 为聚类中心分别在三类中) ');
```

```
index = randperm(69);% 生成随机序号
```

```
% 初始聚类中心的选择
```

```
switch(choice)
```

```
    case(1)      % 聚类中心在第一类中,并选择聚类中心
```

```
        Z1 = x1(index(1),:);
```

```
        Z2 = x1(index(2),:);
```

```
        Z3 = x1(index(3),:);
```

```
    case(2)      % 聚类中心在第二类中
```

```
        Z1 = x2(index(1),:);
```

```
        Z2 = x2(index(2),:);
```

```
        Z3 = x2(index(3),:);
```

```
    case(3)      % 聚类中心在第三类中
```

```
        Z1 = x3(index(1),:);
```

```
        Z2 = x3(index(2),:);
```

```
        Z3 = x3(index(3),:);
```

```
    case(4)      % 聚类中心分别在三类中
```

```
        Z1 = x1(index(1),:);
```

```
        Z2 = x2(index(2),:);
```

```
        Z3 = x3(index(3),:);
```

```
end
```

```
% K-均值算法
```

```
step = 0;
```

```
while(1)
```

```
    step = step + 1; % 记录迭代次数
```

```
    num1 = 0;
```

```
    num2 = 0;
```

```
    num3 = 0;
```

```
    for i=1:207
```

```
        dist1 = 0; % 样本与第一类聚类中心的距离
```

```
        dist2 = 0; % 样本与第二类聚类中心的距离
```

```
        dist3 = 0; % 样本与第三类聚类中心的距离
```

```
        dist1 = sum(( X(i,:) - Z1 ).^2);
```

```
        dist2 = sum(( X(i,:) - Z2 ).^2);
```

```
        dist3 = sum(( X(i,:) - Z3 ).^2);
```

```
        % 判断样本属于哪一类, 并把样本存入该类中
```

```
        if dist1 <= dist2 && dist1 <= dist3
```

```
            num1 = num1 + 1;
```

```
            class1(num1,:) = X(i,:);
```

```
        end
```

```
    if dist2 <= dist1 && dist2 <= dist3
```

```
        num2 = num2 + 1;
```

```

        class2(num2,:) = X(i,:);
    end

    if dist3 <= dist1 && dist3 <= dist2
        num3 = num3 + 1;
        class3(num3,:) = X(i,:);
    end

end

temp1 = mean(class1);
temp2 = mean(class2);
temp3 = mean(class3);
%重新调整当前类别的聚类中心
if sum(temp1-Z1)==0 && sum(temp2-Z2)==0 && sum(temp3-Z3)==0
    break;
else
    Z1 = temp1;
    Z2 = temp2;
    Z3 = temp3;
end
end
disp('迭代次数为: '),step
disp('第一类所含样本数为: '),num1
disp('正确率为: '),num1/69
disp('第二类所含样本数为: '),num2
disp('正确率为: '),num2/69
disp('第三类所含样本数为: '),num3
disp('正确率为: '),num3/69

```

iris 数据:

```

clc;
clear;
close all;
%读入 Iris 数据
x1 = xlsread('data1.xls');
x2 = xlsread('data2.xls');
x3 = xlsread('data3.xls');
X = [x1;x2;x3];
choice = input('Please input the choice: (1 为聚类中心在第一类, 2 为聚类中心在第二类, 3 为聚类中心在第三类, 4 为聚类中心分别在三类中) ');
index = randperm(50);% 生成随机序号
% 初始聚类中心的选择

```

```

switch(choice)
    case(1)      % 聚类中心在第一类中,并选择聚类中心
        Z1 = x1(index(1),:);
        Z2 = x1(index(2),:);
        Z3 = x1(index(3),:);
    case(2)      % 聚类中心在第二类中
        Z1 = x2(index(1),:);
        Z2 = x2(index(2),:);
        Z3 = x2(index(3),:);
    case(3)      % 聚类中心在第三类中
        Z1 = x3(index(1),:);
        Z2 = x3(index(2),:);
        Z3 = x3(index(3),:);
    case(4)      % 聚类中心分别在三类中
        Z1 = x1(index(1),:);
        Z2 = x2(index(2),:);
        Z3 = x3(index(3),:);
end
% K-均值算法
step = 0;
while(1)
    step = step + 1; % 记录迭代次数
    num1 = 0;
    num2 = 0;
    num3 = 0;
    for i=1:150
        dist1 = 0; % 样本与第一类聚类中心的距离
        dist2 = 0; % 样本与第二类聚类中心的距离
        dist3 = 0; % 样本与第三类聚类中心的距离
        dist1 = sum((X(i,:) - Z1).^2);
        dist2 = sum((X(i,:) - Z2).^2);
        dist3 = sum((X(i,:) - Z3).^2);
        % 判断样本属于哪一类, 并把样本存入该类中
        if dist1 <= dist2 && dist1 <= dist3
            num1 = num1 + 1;
            class1(num1,:) = X(i,:);
        end

        if dist2 <= dist1 && dist2 <= dist3
            num2 = num2 + 1;
            class2(num2,:) = X(i,:);
        end

        if dist3 <= dist1 && dist3 <= dist2

```



```

        num3 = num3 + 1;
        class3(num3,:) = X(i,:);
    end

end

temp1 = mean(class1);
temp2 = mean(class2);
temp3 = mean(class3);
%重新调整当前类别的聚类中心
if sum(temp1-Z1)==0 && sum(temp2-Z2)==0 && sum(temp3-Z3)==0
    break;
else
    Z1 = temp1;
    Z2 = temp2;
    Z3 = temp3;
end
end
disp('迭代次数为: '),step
disp('第一类所含样本数为: '),num1
disp('正确率为: '),num1/50
disp('第二类所含样本数为: '),num2
disp('正确率为: '),num2/50
disp('第三类所含样本数为: '),num3
disp('正确率为: '),num3/50
% 画图程序
% 以第一个特征作 x 轴和第二个特征作 y 轴，画图
%原始数据的聚类图
for i=1:150
    if (i<=50)
        plot(X(i,1),X(i,2),'r*')
        hold on
    end
    if (i>50 &&i<=100)
        plot(X(i,1),X(i,2),'g*')
        hold on
    end
    if (i>100&&i<=150)
        plot(X(i,1),X(i,2),'b*')
        hold on
    end
end
end
title('原始数据的聚类图');
strt=['红色*为第一类； 绿色*为第二类； 蓝色*为第三类'];
text(4,1.8,strt);

```

```
%K-均值算法分类的聚类图
figure;
for i=1:num1
    plot(class1(i,1),class1(i,2),'r*');
    hold on;
end
for j=1:num2
    plot(class2(j,1),class2(j,2),'g*');
    hold on;
end
for k=1:num3
    plot(class3(k,1),class3(k,2),'b*');
    hold on;
end
title('K-均值算法的聚类图');
str1=['红色*为第一类； 绿色*为第二类； 蓝色*为第三类'];
text(4,1.8,str1);
```