模式识别作业 6

非监督聚类算法

姓名：罗雁天

院系：清华大学电子系

学号：2018310742

日期：May 18, 2019

目录

第 1 章 Kmeans 聚类

1.1 问题描述

给定数据集 “testSet.txt”，包含 60 行 2 维数据，每行代表一个样本点，分布如图1.1所示。

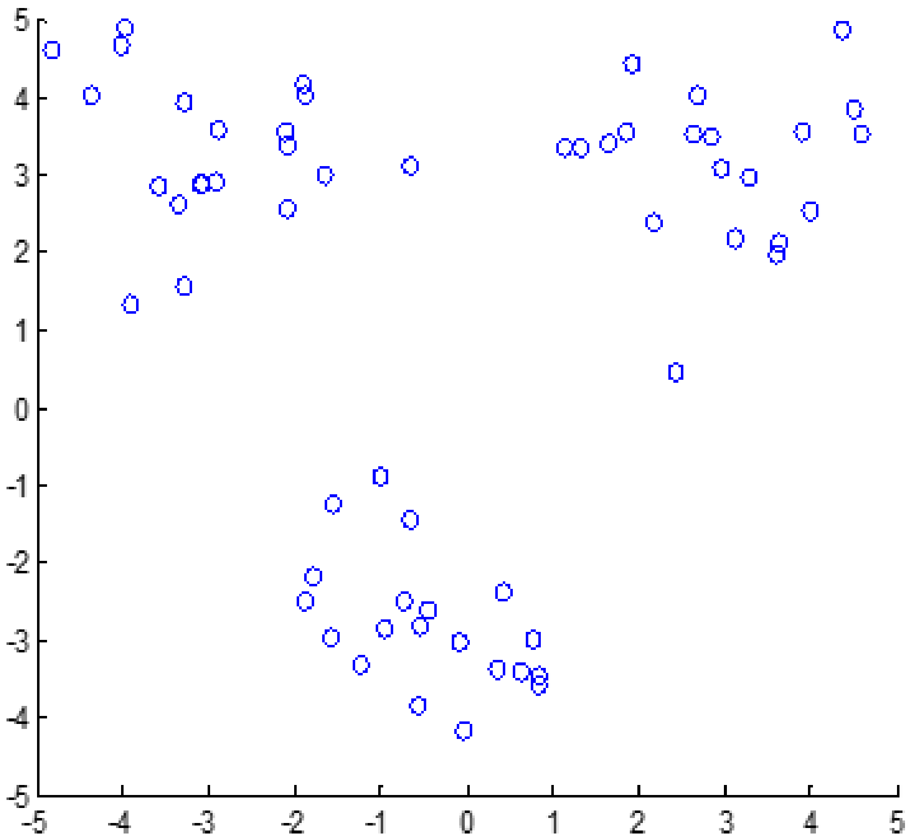


图 1.1

对这组数据进行 kmeans 聚类，令 $k = 2,3,4$。画出聚类结果及每类的中心点，观察聚类结果。记录使用不同初始点时的聚类结果，收敛迭代次数及误差平方和。

- $k = 3$ 时，用给出几组初始点进行聚类:
  - 初始点组 1：[-4.822 4.607;-0.7188 -2.493;4.377 4.864]
  - 初始点组 2：[-3.594 2.857;-0.6595 3.111;3.998 2.519]

  - 初始点组 3：[-0.7188 -2.493;0.8458 -3.59;1.149 3.345]
  - 初始点组 4：[-3.276 1.577;3.275 2.958;4.377 4.864]
- k = 2或4 时，自行给出初始点并聚类，观察聚类结果.

1.2 实验结果

1.2.1 K=3 时

初始点组 1　使用 △ 表示初始聚类中心，○ 表示最终聚类中心，不同颜色表示各样本的聚类结果。首先我们采用欧氏距离作为度量，得到实验结果如图1.2所示。从结果来看，迭代两次便得到了最终的聚类结果，收敛很快。

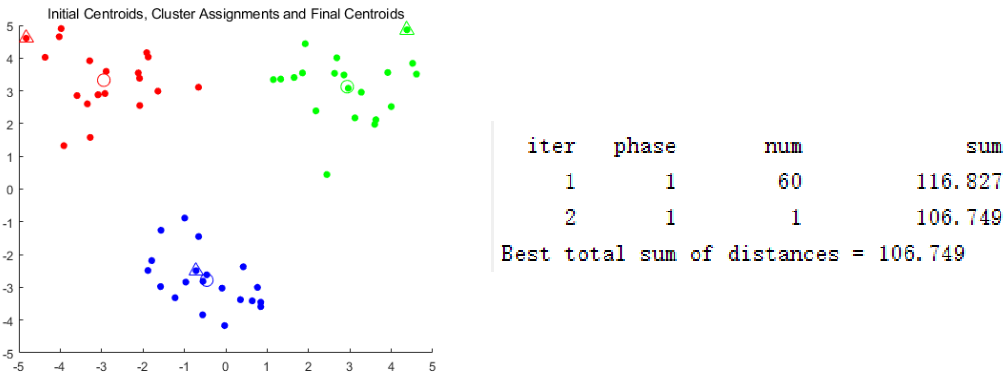


图 1.2: K=3，使用初始点组 1 聚类结果 (欧氏距离作为度量)

与之进行对比，我们使用曼哈顿距离作为度量，再次进行实验，结果如图1.3所示。可以看出，聚类结果有微小的差距，迭代次数同样是 2 次。由此可以看出，在这种初始化条件下，使用欧氏距离和曼哈顿距离进行 Kmeans 聚类的结果相近，并且迭代次数都很少，说明这种初始条件效果较好。

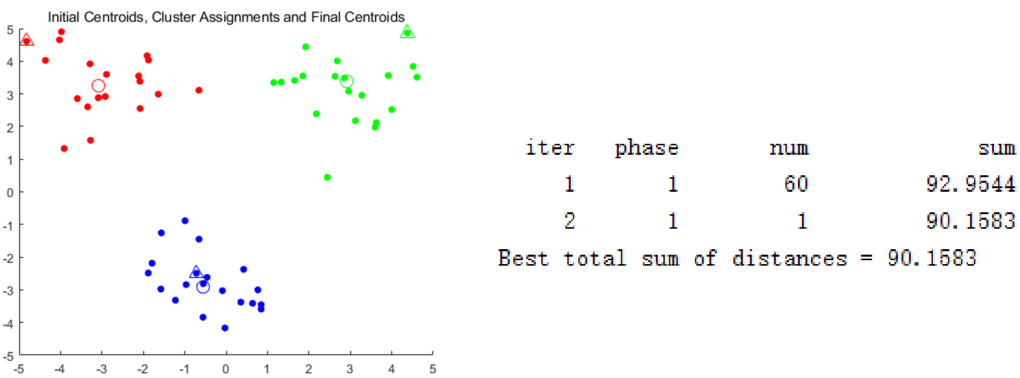
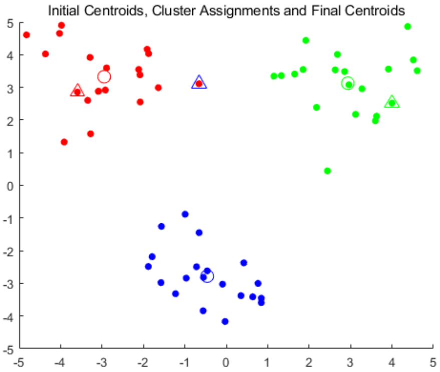


图 1.3: K=3，使用初始点组 1 聚类结果 (曼哈顿距离作为度量)

初始点组 2 使用 △ 表示初始聚类中心，∘ 表示最终聚类中心，不同颜色表示各样本的聚类结果。首先我们采用欧氏距离作为度量，得到实验结果如图1.4所示。从结果来看，迭代两次便得到了最终的聚类结果，收敛很快。

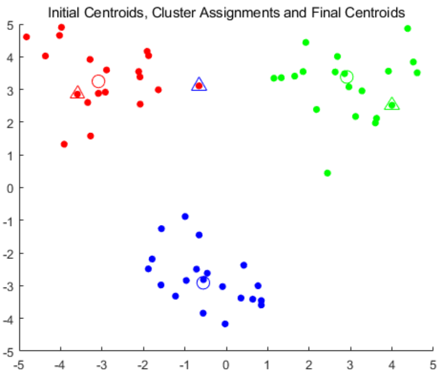


iter	phase	num	sum
1	1	60	409.256
2	1	12	106.749

Best total sum of distances = 106.749

图 1.4: K=3，使用初始点组 2 聚类结果 (欧氏距离作为度量)

与之进行对比，我们使用曼哈顿距离作为度量，再次进行实验，结果如图1.5所示。可以看出，聚类结果有微小的差距，迭代次数同样是 2 次。由此可以看出，在这种初始化条件下，使用欧氏距离和曼哈顿距离进行 Kmeans 聚类的结果相近，并且迭代次数都很少。



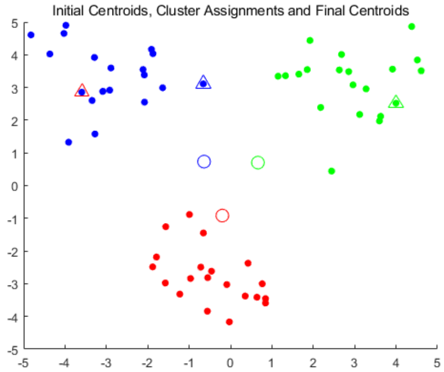
iter	phase	num	sum
1	1	60	154.616
2	1	11	90.1583

Best total sum of distances = 90.1583

图 1.5: K=3，使用初始点组 2 聚类结果 (曼哈顿距离作为度量)

然后我们又尝试使用余弦相似度作为度量进行聚类，结果如图**??**所示，从结果来看，最后聚类结果还是很不错的，但是迭代次数相比于欧氏距离和曼哈顿距离多了一次，因此在此初始条件下，余弦相似度作为度量的聚类效果不如欧氏距离和曼哈顿距离好。

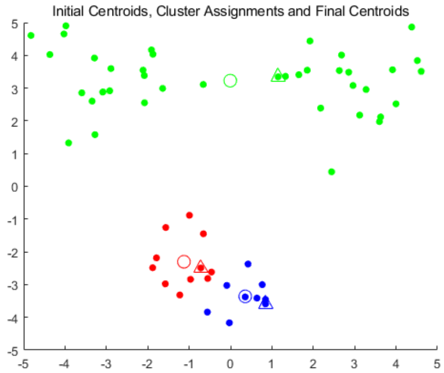
初始点组 3 同样，使用 △ 表示初始聚类中心，∘ 表示最终聚类中心，不同颜色表示各样本的聚类结果。首先我们采用欧氏距离作为度量，得到实验结果如图1.7所示。从结果来看，迭代两次便得到了最终的聚类结果，收敛很快，但是聚类结果显然没有之前两组好，由此也可以说明 kmeans 聚类是初始化敏感的，初始化对最终聚类结果的影响是较大的。



iter	phase	num	sum
1	1	60	22.7772
2	1	20	9.75537
3	1	10	2.47058

Best total sum of distances = 2.47058

图 1.6: K=3，使用初始点组 2 聚类结果 (余弦相似度作为度量)

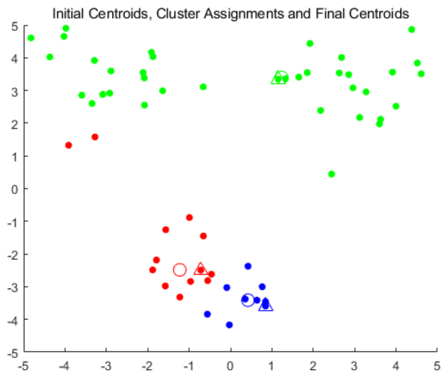


iter	phase	num	sum
1	1	60	442.011
2	1	4	436.63

Best total sum of distances = 436.63

图 1.7: K=3，使用初始点组 3 聚类结果 (欧氏距离作为度量)

与之进行对比，我们使用曼哈顿距离作为度量，再次进行实验，结果如图1.8所示。可以看出，迭代次数是 3 次，比欧氏距离多了 1 次，说明收敛较慢。同样，聚类效果相比于之前两组初始条件不好。



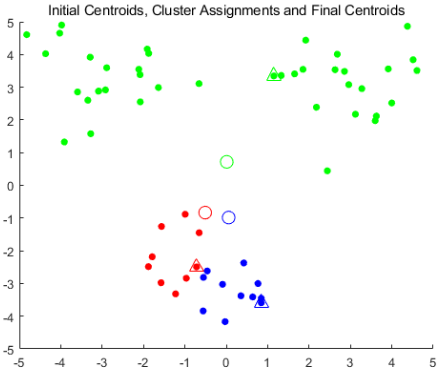
iter	phase	num	sum
1	1	60	164.844
2	1	3	163.897
3	1	1	163.776

Best total sum of distances = 163.776

图 1.8: K=3，使用初始点组 3 聚类结果 (曼哈顿距离作为度量)

然后我们又尝试使用余弦相似度作为度量进行聚类，结果如图1.9所示，从结果来看，迭代次数更多了，而且聚类效果也不如之前初始化条件下的好。

初始点组 4 使用 △ 表示初始聚类中心，○ 表示最终聚类中心，不同颜色表示各样本的聚类结果。首先我们采用欧氏距离作为度量，得到实验结果如图1.10所示。

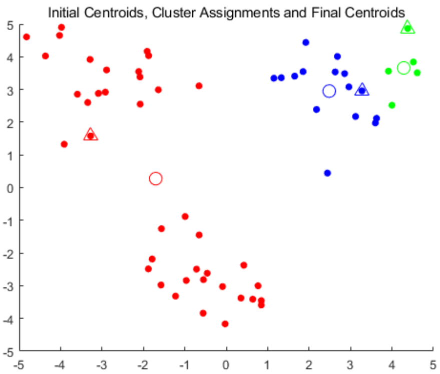


iter	phase	num	sum
1	1	60	11.7457
2	1	1	11.6809
3	1	1	11.666
4	1	1	11.6506
5	1	1	11.629

Best total sum of distances = 11.629

图 1.9: K=3，使用初始点组 3 聚类结果 (余弦相似度作为度量)

从结果来看，得到的最终结果也没有初始条件 1 和 2 的效果好，但是收敛速度还是挺快的。

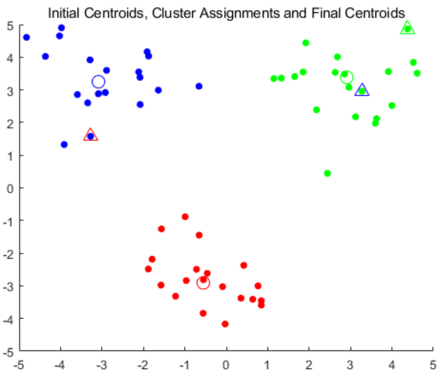


iter	phase	num	sum
1	1	60	529.506
2	1	1	527.901
3	1	1	527.304

Best total sum of distances = 527.304

图 1.10: K=3，使用初始点组 4 聚类结果 (欧氏距离作为度量)

与之进行对比，我们使用曼哈顿距离作为度量，再次进行实验，结果如图1.11所示。可以看出，此时的聚类效果和初始条件 1 和 2 的类似，比使用欧氏距离得到的结果较好，但是迭代次数稍微多了一点。所以对于此种初始化条件下来看，曼哈顿距离比欧氏距离具有更好的聚类效果。



iter	phase	num	sum
1	1	60	203.29
2	1	6	198.716
3	1	2	196.892
4	1	1	196.3
5	1	11	175.284
6	1	11	124.144
7	1	10	90.1583

Best total sum of distances = 90.1583

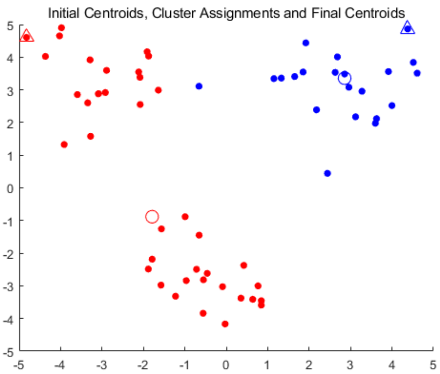
图 1.11: K=3，使用初始点组 4 聚类结果 (曼哈顿距离作为度量)

1.2.2 K=2

从之前 $K = 3$ 部分的讨论来看，初始的聚类中心对实验结果有较大的影响，因此我们观察数据，设置较好的聚类中心进行实验。并且从之前的讨论来看，使用曼哈顿距离聚类效果更好一些，因此在本部分我们选择曼哈顿距离作为度量。

观察数据可以看出，数据主要集中在三部分，要聚类成两类最终结果大概是把其中两类聚在了一起，因此，我们可以构造不同的初始条件，将不同的两类聚类到一起。

初始条件 1 我们选取初始条件为:[-4.822 4.607;4.377 4.864],实验结果如图1.12所示。从图中可以看出，我们将左上角和下面的两类聚成了一类，右上角聚成了一类。

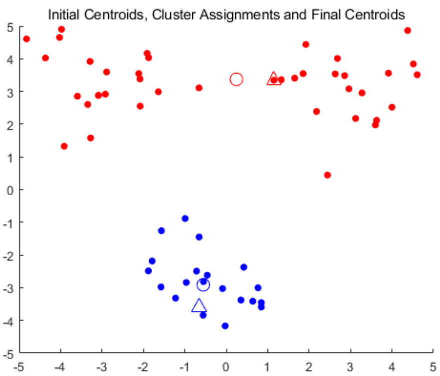


iter	phase	num	sum
1	1	60	213.892
2	1	2	211.885
3	1	2	210.888
4	1	4	205.73
5	1	1	204.365

Best total sum of distances = 204.365

图 1.12: K=2 聚类结果 (曼哈顿距离作为度量)

初始条件 2 我们选取初始条件为：[1.149 3.345;-0.6595 -3.59]，实验结果如图1.13所示。从图中可以看出，我们将左上角和右上角的两类聚成了一类，下面聚成了一类。

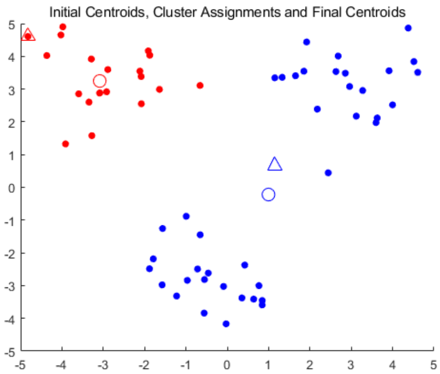


iter	phase	num	sum
1	1	60	174.038

Best total sum of distances = 174.038

图 1.13: K=2 聚类结果 (曼哈顿距离作为度量)

初始条件 3 我们选取初始条件为：[1.149 3.345;-0.6595 -3.59]，实验结果如图1.14所示。从图中可以看出，我们将下面和右上角的两类聚成了一类，左上角聚成了一类。

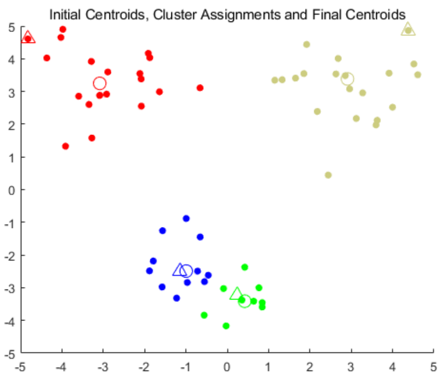



```
  iter    phase      num          sum
     1        1       60      219.034
     2        1        1      217.563
Best total sum of distances = 217.563
```

图 1.14: K=2 聚类结果 (曼哈顿距离作为度量)

1.2.3 K=4

同样的，我们我们选择曼哈顿距离作为度量，观察数据设计初始聚类中心为：[-4.822 4.607;-1.149 -2.493;0.232, -3.222;4.377 4.864]，实验结果如图1.15所示。从图中可以看出，聚类效果还是很不错的，并且迭代两次便收敛了，收敛速度也很快。




```
  iter    phase      num          sum
     1        1       60      82.7963
     2        1        1      80.4009
Best total sum of distances = 80.4009
```

图 1.15: K=4 聚类结果 (曼哈顿距离作为度量)

1.3 实验结论

从结果来看，最终的聚类效果与初始条件的设置有很大的关系，因此，在聚类之前观察数据的分布，选择较好的类别数和聚类中心能够得到更好的聚类效果。

第 2 章 分层聚类

有可用高斯分布近似的两个样本集：

$$\begin{aligned}\omega_1 &= \{(2,0),(2,2),(2,4),(3,3)\} \\ \omega_2 &= \{(0,3),(-2,2),(-1,-1),(1,-2),(3,-1)\}\end{aligned} \tag{2.1}$$

- 求：用最小错误概率分类时的识别界面
- 令 $\omega = \omega_1 \cup \omega_2$，距离取最远距离 $d_{max}(S_i, S_j) = \max\limits_{X_i \in S_i, X_j \in S_j} ||X_i - X_j||$，使用分层聚类法聚类并作图。

2.1 用最小错误概率分类时的识别界面

使用最小错误概率分类算法已在上次作业中实现，因此本次直接调用了上次高斯分布时判决的函数代码进行实验。首先计算两个类的类均值：

$$M_1 = [2.25, 2.25], M_2 = [0.2, 0.2] \tag{2.2}$$

根据无偏估计的协方差矩阵计算方法：

$$\Sigma_i = \frac{1}{N-1}(\omega_i - M_i)^T(\omega_i - M_i) \quad i = 1,2 \tag{2.3}$$

计算两个类的协方差矩阵：

$$\Sigma_1 = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 2.9167 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 3.7 & -2.05 \\ -2.05 & 4.7 \end{bmatrix} \tag{2.4}$$

计算正态分布时贝叶斯判别准则所需要的参数如下：

$$
\begin{aligned}
\mathbf{W}_1 &= -\frac{1}{2}\Sigma_1^{-1} = \begin{bmatrix} -2.1875 & 0.1875 \\ 0.1875 & -0.1875 \end{bmatrix} \\
\mathbf{W}_2 &= -\frac{1}{2}\Sigma_2^{-1} = \begin{bmatrix} -0.1782 & -0.0777 \\ -0.0777 & -0.1403 \end{bmatrix} \\
\mathbf{w}_1 &= \Sigma_1^{-1} M_1 = [9,0]^T; \\
\mathbf{w}_2 &= \Sigma_2^{-1} M_2 = [0.1024, 0.0872]^T; \\
w_{10} &= -\frac{1}{2} M_1^T \Sigma_1^{-1} M_1 - \frac{1}{2}\ln|\Sigma_1| + \ln P(\omega_1) = -10.6154 \\
w_{20} &= -\frac{1}{2} M_2^T \Sigma_1^{-1} M_2 - \frac{1}{2}\ln|\Sigma_2| + \ln P(\omega_2) = -2.0017
\end{aligned}
\tag{2.5}
$$

对任意数据点 $\mathbf{x} = [x_1, x_2]^T$，计算两类的识别函数如下：

$$
\begin{aligned}
d_1(\mathbf{x}) &= \mathbf{x}^T \mathbf{W}_1 \mathbf{x} + \mathbf{w}_1^T \mathbf{x} + w_{10} \\
&= -2.1875x_1^2 - 0.1875x_2^2 + 9x_1 - 10.6154 \\
d_2(\mathbf{x}) &= \mathbf{x}^T \mathbf{W}_2 \mathbf{x} + \mathbf{w}_2^T \mathbf{x} + w_{20} \\
&= -0.1782x_1^2 - 0.1403x_2^2 + 0.1024x_1 + 0.0872x_2 - 2.0017
\end{aligned}
\tag{2.6}
$$

则判别函数如下：

$$
f(\mathbf{x}) = \begin{cases} x \in \text{class1} & \text{if} \quad d_1(\mathbf{x}) > d_2(\mathbf{x}) \\ x \in \text{class2} & \quad else \end{cases}
\tag{2.7}
$$

计算识别界面如下：

$$
d_1(\mathbf{x}) = d_2(\mathbf{x}) \Rightarrow 2.0093x_1^2 + 0.0472x_2^2 - 0.5305x_1x_2 + 9.1024x_1 + 0.0872x_2 - 12.6172 = 0
\tag{2.8}
$$

由此可以看出，分类界面在此种情况下是椭圆。

绘制出两个二维高斯分布的曲面如图2.1所示，识别界面如图2.2所示

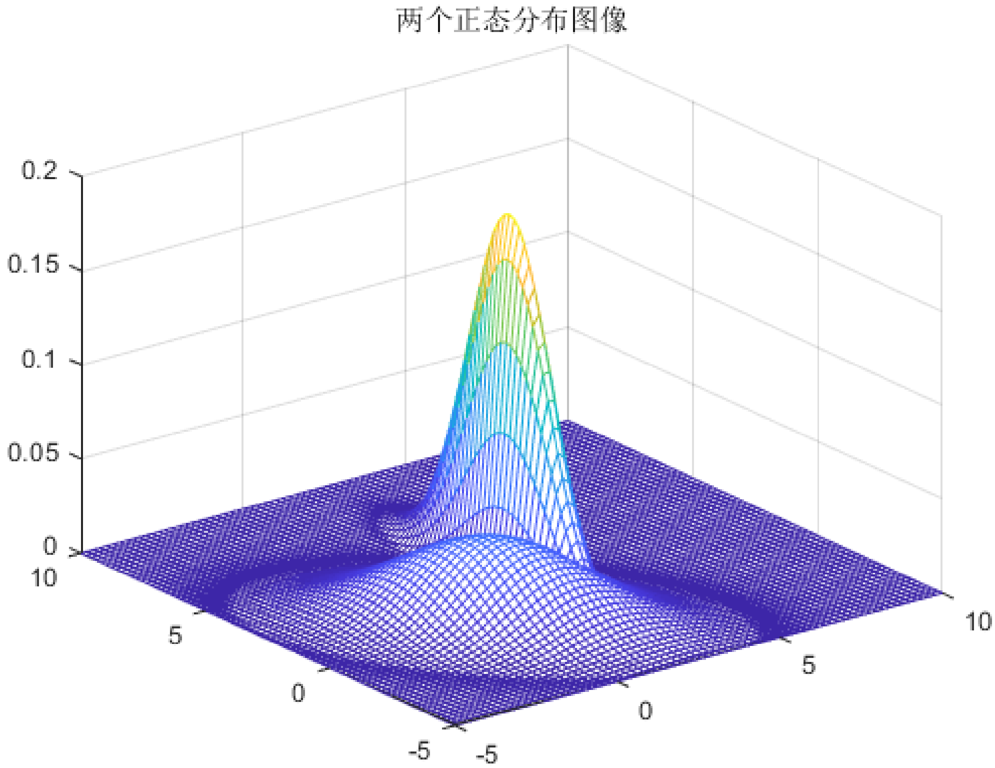


图 **2.1:** 二维高斯分布密度函数曲面

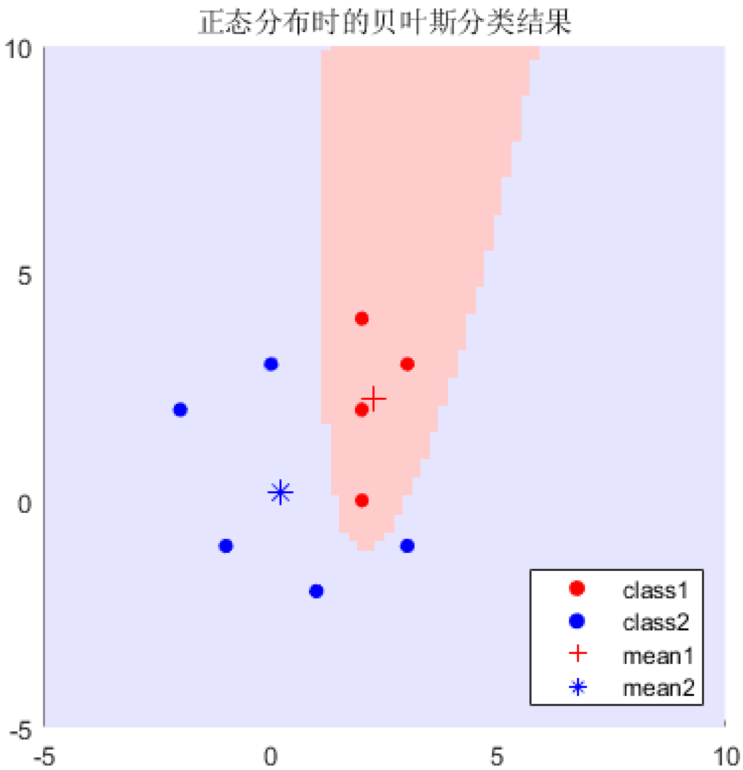


图 **2.2:** 识别界面

2.2 分层聚类

使用最远距离作为度量，层次聚类的结果如图2.3所示，其中每层聚合图如图2.4所示。

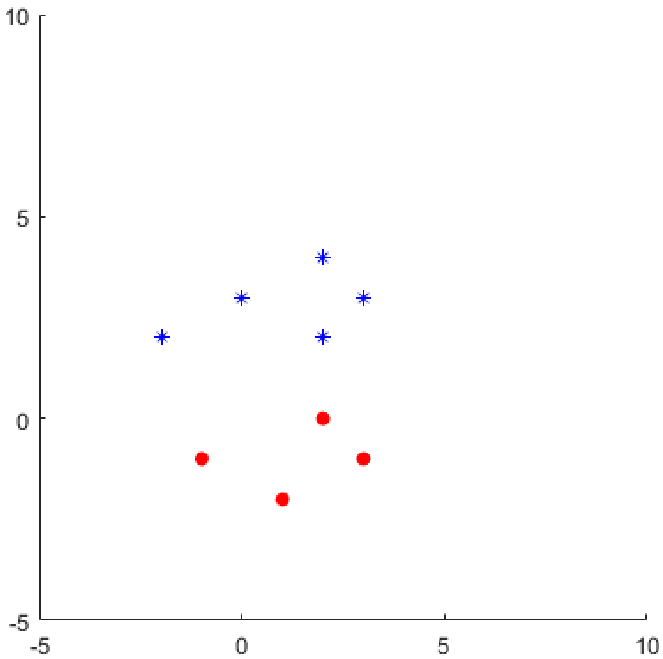


图 2.3: 层次聚类结果图

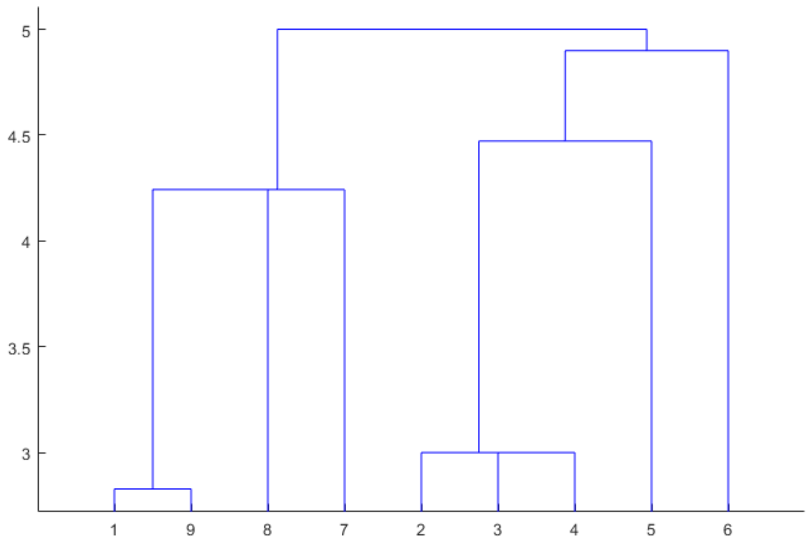


图 2.4: 层次聚类树

作为对比，我们使用闵科夫斯基距离 (式2.9) 在不同 p 值下进行层次聚类，我们使用了 $p = 2, 5, 10, 50, 100$ 进行实验，($p = 2$ 时为欧氏距离，$p = +\infty$ 时为最

远距离) 结果如图2.5,2.6,2.7,2.8,2.9所示，从图中可以看出，层次聚类的过程有轻微的差别，但是最终结果都是一致的。

$$d_{ij} = \left(\sum_{k=1}^{m} \left| x_{kj} - x_{ki} \right|^{p} \right)^{\frac{1}{p}} \tag{2.9}$$

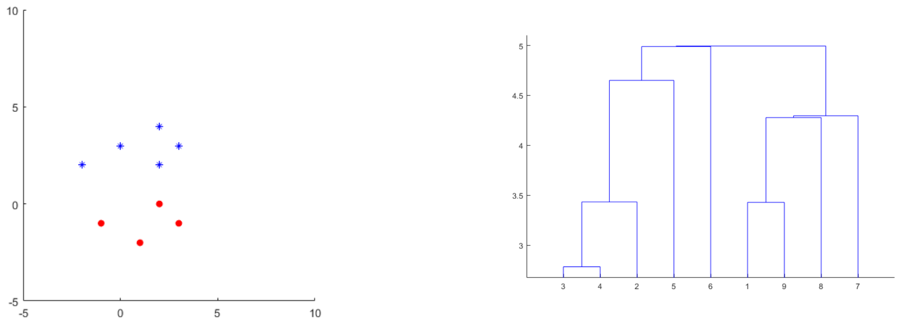


图 2.5: p=2 时聚类结果 (即为欧式距离)

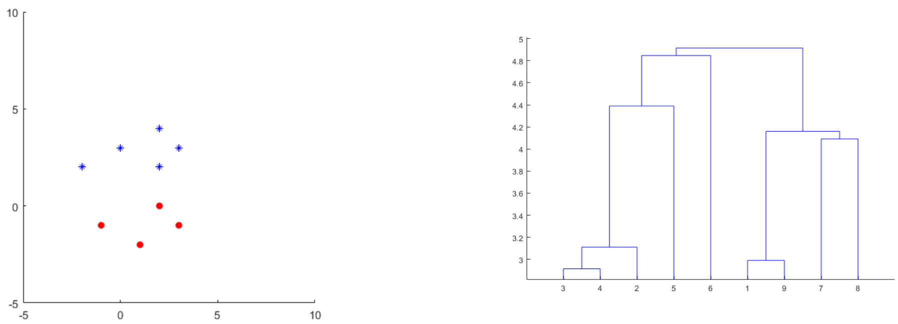


图 2.6: p=5 时聚类结果

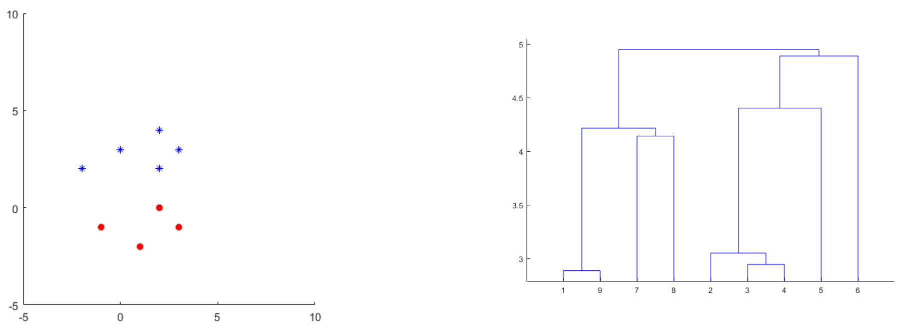


图 2.7: p=10 时聚类结果

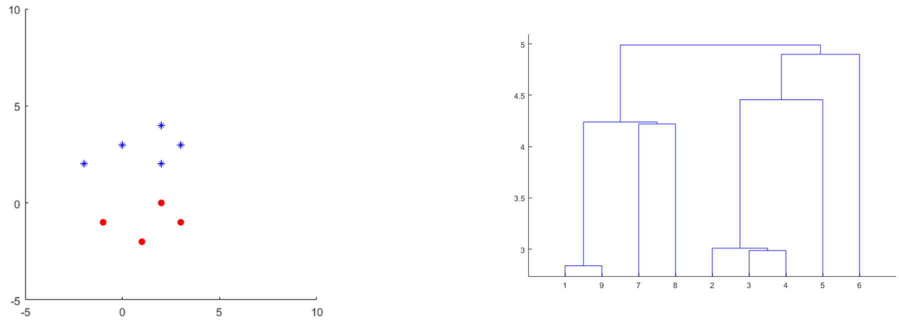


图 2.8: p=50 时聚类结果

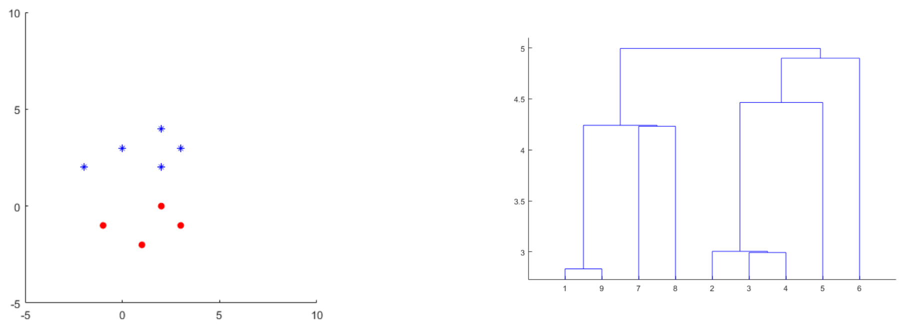


图 2.9: p=100 时聚类结果

第 3 章 代码说明

本次实验使用 Matlab 语言编写，所有代码放置在“code/”文件夹下:

- kmeansmain.m: 使用 kmeans 聚类讨论的主程序，直接执行即可；
- Bayes_Gauss.m: 使用高斯分布时的贝叶斯判别准则绘制分类界面的函数，与上次作业一致；
- gauss_main.m: 使用最小错误率绘制识别界面的主函数，直接执行即可；
- hierarchical_cluster.m: 使用层次聚类的主程序，直接执行即可。