第二章　聚类分析

2.1　聚类分析的概念

2.2　模式相似性测度

2.3　类的定义与类间距离

2.4 准则函数

2.5　聚类的算法

2.5 聚类的算法

最大最小距离和层次聚类算法的一个共同特点是某个模式一旦划分到某一类之后，在后继的算法过程中就不改变了，而简单聚类算法中类心一旦选定后在后继算法过程中也不再改变了。因此，这些方法效果一般不会太理想。

2.5.4 动态聚类法

1. 确定模式和聚类的距离测度。当采用欧氏距离时，是计算此模式和该类中心的欧氏距离；为能反映出类的模式分布结构，可采用马氏距离，设该类的均矢为 $\vec{\mu}$，协方差阵为 Σ，则模式 $\vec{x}$ 和该类的距离平方为 $\vec{x}$ 与该类均矢 $\vec{\mu}$ 的马氏距离：

$$d^2 = (\vec{x} - \vec{\mu})'\Sigma^{-1}(\vec{x} - \vec{\mu})$$

2. 确定评估聚类质量的准则函数。

3. 确定模式分划及聚类合并或分裂的规则。

2.5 聚类的算法

2.5.4 动态聚类法

基本步骤：

1. 建立初始聚类中心，进行初始聚类；

2.5 聚类的算法

2.5.4 动态聚类法

基本步骤：

1. 建立初始聚类中心，进行初始聚类；

2. 计算模式和类的距离，调整模式的类别；

2.5 聚类的算法

2.5.4 动态聚类法

基本步骤：

1. 建立初始聚类中心，进行初始聚类；
2. 计算模式和类的距离，调整模式的类别；
3. 计算各聚类的参数，删除、合并或分裂一些聚类；

2.5 聚类的算法

2.5.4 动态聚类法

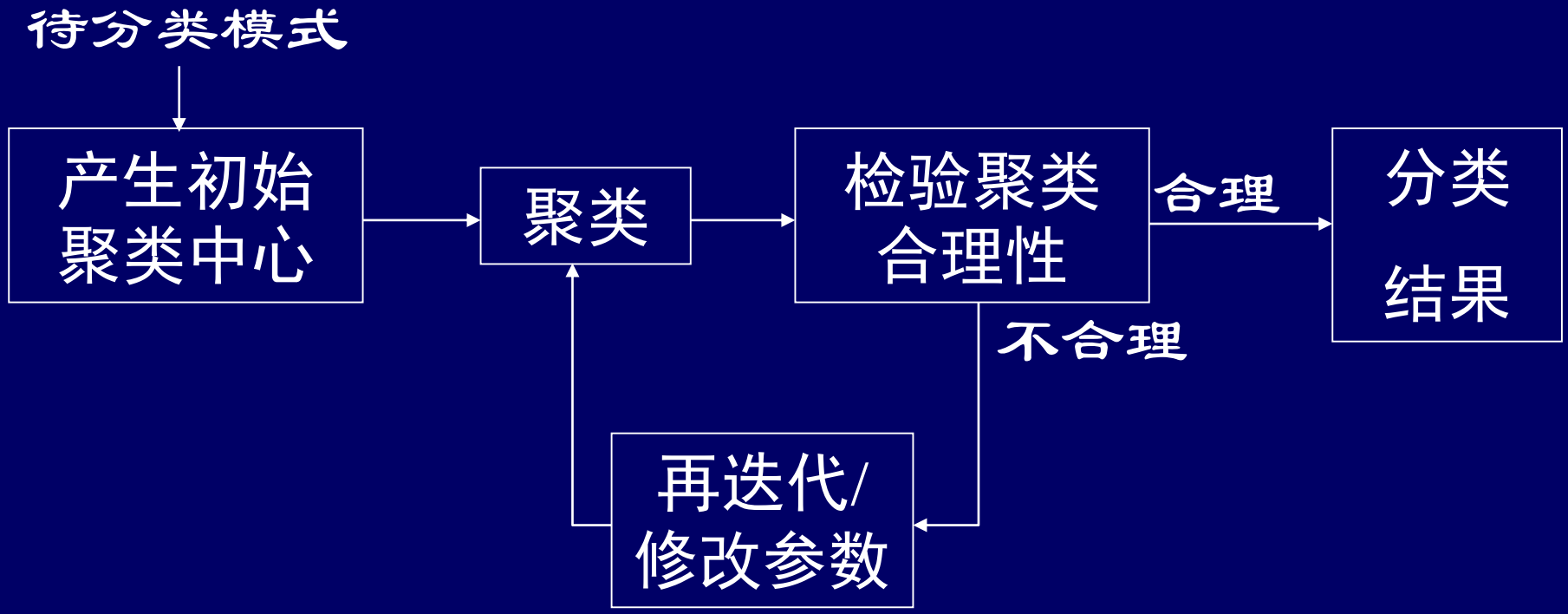
基本步骤：

1. 建立初始聚类中心，进行初始聚类；
2. 计算模式和类的距离，调整模式的类别；
3. 计算各聚类的参数，删除、合并或分裂一些聚类；
4. 从初始聚类开始，运用迭代算法动态地改变模式的类别和聚类的中心使准则函数取得极值或设定的参数达到设计要求时停止。

2.5 聚类的算法

2.5.4 动态聚类法

动态聚类基本框图



2.5 聚类的算法

2.5.4 动态聚类法　　C-均值法

1. 条件及约定

设待分类的模式特征矢量集为：$\{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}$

类的数目C是事先取定的。

2. 算法思想

该方法**取定 C个类别**和选取 **C个初始聚类中心**，按最小距离原则将各模式分配到 C类中的某一类，之后不断地**计算类心和调整各模式**的类别，最终使各模式到其判属类别中心的距离平方之和最小。

2.5.4 动态聚类法　　C-均值法

3. 算法原理步骤

(1) 任选C个模式特征矢量作为初始聚类中心：

$\vec{z}_1^{(0)}, \vec{z}_2^{(0)}, \cdots, \vec{z}_c^{(0)}$ ，令 $k=0$ 。

(2) 将待分类的模式特征矢量集 $\{\vec{x}_i\}$ 中的模式逐个按最小距离原则分划给C类中的某一类，即：

如果 $d_{il}^{(k)} = \min_j [d_{ij}^{(k)}] \qquad (i=1,2,\cdots,N)$

则 $\vec{x}_i \in \omega_l^{(k+1)}$ ，式中 $d_{ij}^{(k)}$ 表示 $\vec{x}_i$ 和 $\omega_j^{(k)}$ 的中心 $\vec{z}_j^{(k)}$ 的距离，上角标表示迭代次数。于是产生新聚类 $\omega_j^{(k+1)} \quad (j=1,2\cdots,c)$ 。

2.5.4 动态聚类法　　C-均值法

3. 算法原理步骤

(3) 计算重新分类后的各类心

$$\vec{z}_j^{(k+1)} = \frac{1}{n_j^{(k+1)}} \sum_{\vec{x}_i \in \omega_j^{(k+1)}} \vec{x}_i, \quad (j = 1,2,\cdots,c)$$

式中 $n_j^{(k+1)}$ 为类 $\omega_j^{(k+1)}$ 中所含模式的个数。

(4) 如果 $\vec{z}_j^{(k+1)} = \vec{z}_j^{(k)}$, $j = 1,2,\cdots,c$, 则结束, 否则 $k = k+1$, 转至(2)。

2.5 聚类的算法

例：已知有20个样本，每个样本有2个特征，数据分布如下图，使用C—均值法实现样本分类（C=2）。

样本序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
特征x_1	0	1	0	1	2	1	2	3	6	7
特征x_2	0	0	1	1	1	2	2	2	6	6

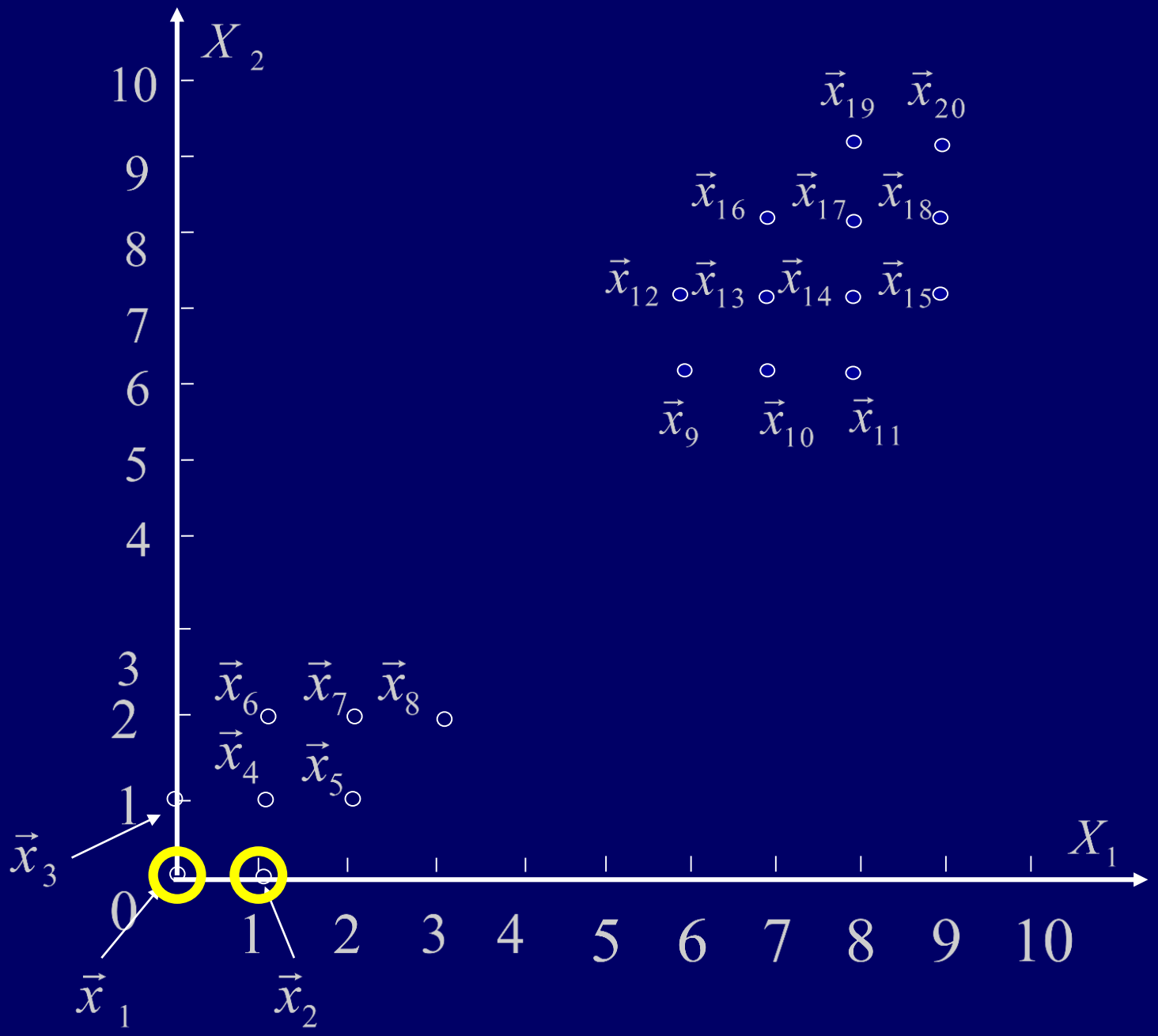
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
8	6	7	8	9	7	8	9	8	9
6	7	7	7	7	8	8	8	9	9

2.5 聚类的算法

2.5.4 动态聚类法　　C-均值法

第一步：令C=2，选初始聚类中心为

$$\vec{Z}_1(1) = \vec{x}_1 = (0,0)^T ; \vec{Z}_2(1) = \vec{x}_2 = (1,0)^T$$


X_2
X_1
$\vec{x}_1$
$\vec{x}_2$
$\vec{x}_3$
$\vec{x}_4$
$\vec{x}_5$
$\vec{x}_6$
$\vec{x}_7$
$\vec{x}_8$
$\vec{x}_9$
$\vec{x}_{10}$
$\vec{x}_{11}$
$\vec{x}_{12}$
$\vec{x}_{13}$
$\vec{x}_{14}$
$\vec{x}_{15}$
$\vec{x}_{16}$
$\vec{x}_{17}$
$\vec{x}_{18}$
$\vec{x}_{19}$
$\vec{x}_{20}$

2.5 聚类的算法

第二步： $\|\vec{x}_1 - \vec{Z}_1(1)\| = \left\| \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\| = 0$

$\|\vec{x}_1 - \vec{Z}_2(1)\| = \left\| \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\| = 1$

因为 $\|\vec{x}_1 - \vec{Z}_1(1)\| < \|\vec{x}_1 - \vec{Z}_2(1)\|$

所以 $\vec{x}_1 \in \vec{Z}_1(1)$

$\|\vec{x}_2 - \vec{Z}_1(1)\| = \left\| \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\| = 1$

2.5 聚类的算法

$$\| \vec{x}_2 - \vec{Z}_2(1) \| = \left\| \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\| = 0$$

因为 $\| \vec{x}_2 - \vec{Z}_1(1) \| > \| \vec{x}_2 - \vec{Z}_2(1) \|$ ， 所以 $\vec{x}_2 \in \vec{Z}_2(1)$

同理 $\| \vec{x}_3 - \vec{Z}_1(1) \| = 1 < \| \vec{x}_3 - \vec{Z}_2(1) \| = 2, \therefore \vec{x}_3 \in \vec{Z}_1(1)$

$\| \vec{x}_4 - \vec{Z}_1(1) \| = 2 > \| \vec{x}_4 - \vec{Z}_2(1) \| = 1, \therefore \vec{x}_4 \in \vec{Z}_2(1)$

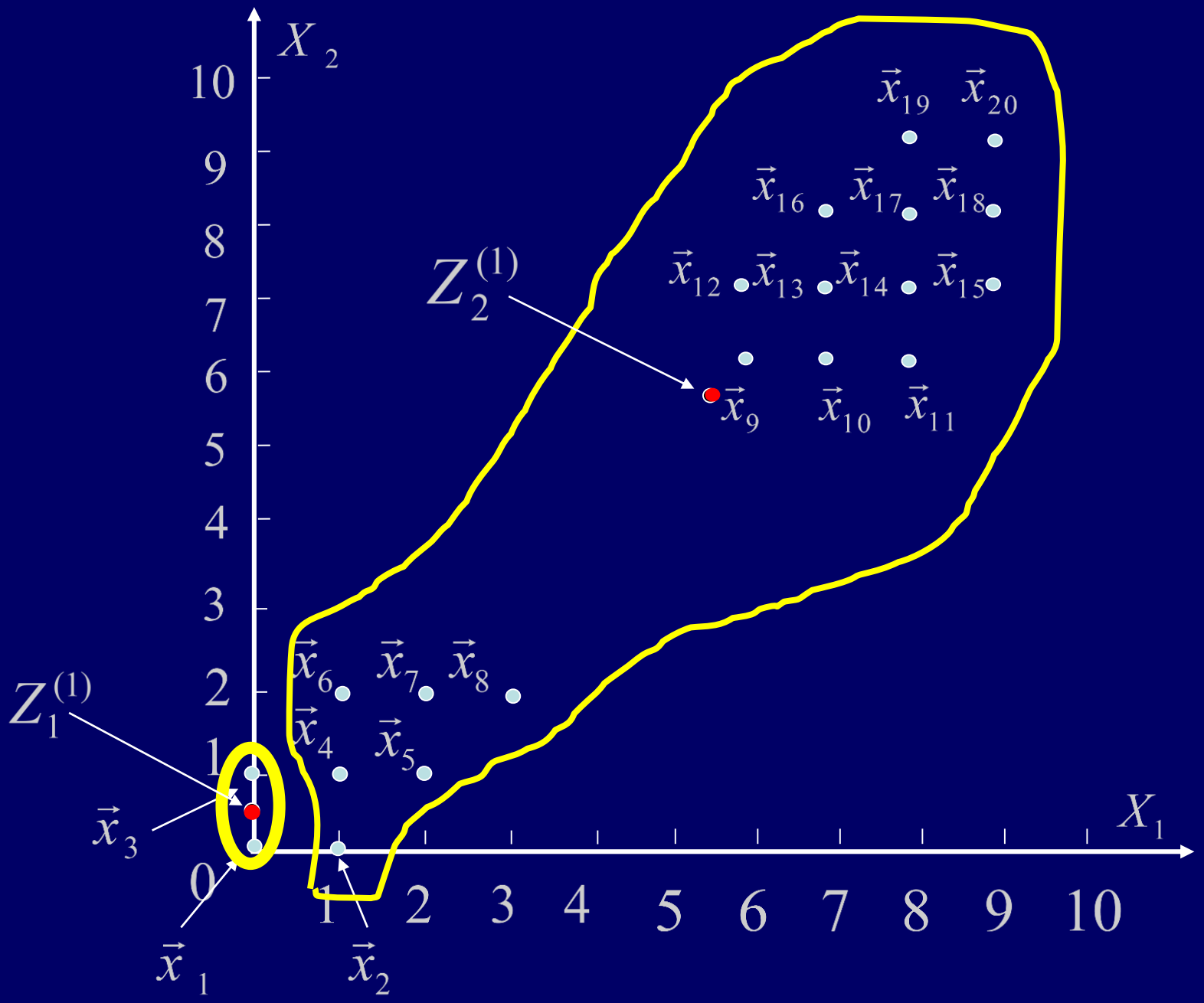
同样把所有 $\vec{x}_5$、 $\vec{x}_6$、… $\vec{x}_{20}$ 与第二个聚类中心的距离计算出来，判断 $\vec{x}_5$、 $\vec{x}_6$、… $\vec{x}_{20}$ 都属于 $\vec{Z}_2(1)$

因此分为两类：

一、 $G_1(1) = (\vec{x}_1, \vec{x}_3)$,

二、 $G_2(1) = (\vec{x}_2, \vec{x}_4, \vec{x}_5, \cdots \vec{x}_{20})$

$N_1 = 2, N_2 = 18$


X_2
X_1
$Z_2^{(1)}$
$Z_1^{(1)}$
$\vec{x}_1$
$\vec{x}_2$
$\vec{x}_3$
$\vec{x}_4$
$\vec{x}_5$
$\vec{x}_6$
$\vec{x}_7$
$\vec{x}_8$
$\vec{x}_9$
$\vec{x}_{10}$
$\vec{x}_{11}$
$\vec{x}_{12}$
$\vec{x}_{13}$
$\vec{x}_{14}$
$\vec{x}_{15}$
$\vec{x}_{16}$
$\vec{x}_{17}$
$\vec{x}_{18}$
$\vec{x}_{19}$
$\vec{x}_{20}$

2.5 聚类的算法

第三步：根据新分成的两类建立新的聚类中心

$$\vec{Z}_1(2)=\frac{1}{N_1}\sum_{\vec{x}_i\in G_1(1)}\vec{x}_i=\frac{1}{2}(\vec{x}_1+\vec{x}_3)=\frac{1}{2}\left[\begin{pmatrix}0\\0\end{pmatrix}+\begin{pmatrix}0\\1\end{pmatrix}\right]$$

$$=\frac{1}{2}\begin{pmatrix}0\\1\end{pmatrix}=(0,0.5)^T$$

$$\vec{Z}_2(2)=\frac{1}{N_2}\sum_{\vec{x}_i\in G_2(1)}\vec{x}_i=\frac{1}{18}(\vec{x}_2+\vec{x}_4+\vec{x}_5+\ldots+\vec{x}_{20})$$

$$=(5.67,5.33)^T$$

2.5 聚类的算法

第四步：

因为 $\vec{Z}_j(2) \neq \vec{Z}_j(1), j = 1, 2$(新旧聚类中心不等)

转第二步。

第二步：重新计算 $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{20}$ 到 $\vec{Z}_1(2)$，$\vec{Z}_2(2)$的距离，把它们归为最近聚类中心，重新分为两类

2.5 聚类的算法

$$G_1(2) = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_8), N_1 = 8$$

$$G_2(2) = (\vec{x}_9, \vec{x}_{10}, \ldots, \vec{x}_{20}), N_2 = 12$$

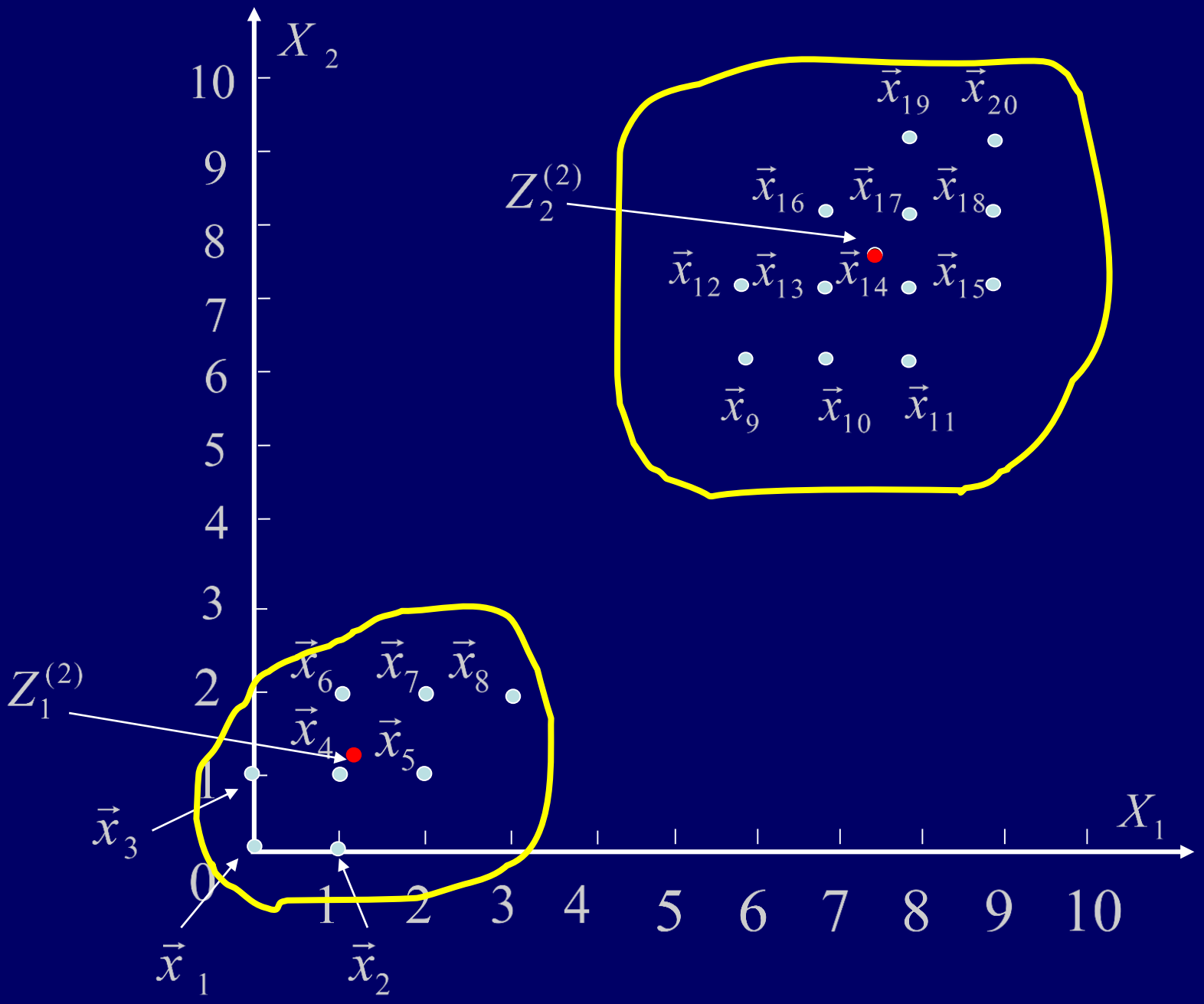
第三步：更新聚类中心

$$\vec{Z}_1(3) = \frac{1}{N_1} \sum_{\vec{x}_i \in G_1(2)} \vec{x}_i = \frac{1}{8}(\vec{x}_1 + \vec{x}_2 + \vec{x}_3 + \ldots + \vec{x}_8)$$

$$= (1.25, 1.13)^T$$

$$\vec{Z}_2(3) = \frac{1}{N_2} \sum_{\vec{x}_i \in G_2(2)} \vec{x}_i = \frac{1}{12}(\vec{x}_9 + \vec{x}_{10} + \ldots + \vec{x}_{20})$$

$$= (7.67, 7.33)^T$$


X_2
X_1
$Z_2^{(2)}$
$Z_1^{(2)}$
$\vec{x}_{19}$ $\vec{x}_{20}$
$\vec{x}_{16}$ $\vec{x}_{17}$ $\vec{x}_{18}$
$\vec{x}_{12}$ $\vec{x}_{13}$ $\vec{x}_{14}$ $\vec{x}_{15}$
$\vec{x}_9$ $\vec{x}_{10}$ $\vec{x}_{11}$
$\vec{x}_6$ $\vec{x}_7$ $\vec{x}_8$
$\vec{x}_4$ $\vec{x}_5$
$\vec{x}_3$
$\vec{x}_1$ $\vec{x}_2$
0 1 2 3 4 5 6 7 8 9 10

2.5 聚类的算法

第四步： 因$\vec{Z}_j(3) \neq \vec{Z}_j(2), j=1,2$,转第二步

第二步： 重新计算$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{20}$到$\vec{Z}_1(3), \vec{Z}_2(3)$的距离，

分别把$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{20}$归于最近的那个聚类中心

重新分为二类$G_1(4) = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_8)$

$G_2(4) = (\vec{x}_9, \vec{x}_{10}, \ldots, \vec{x}_{20}), N_1 = 8, N_2 = 12$

第三步： 更新聚类中心

$Z_1(4) = Z_1(3) = (1.25, 1.13)^T$

$Z_2(4) = Z_2(3) = (7.67, 7.33)^T$

计算结束。

2.5 聚类的算法

2.5.4 动态聚类法　　C-均值法

C均值算法的分类结果受到取定的类别数和初始聚类中心的影响，通常结果只是局部最优的，但其方法简单，结果尚令人满意，故应用较多。

也可以对C均值算法进行如下改进。

2.5.4 动态聚类法　　C-均值法

C-均值算法的改进

- **类数c的调整**

  - 利用先验知识选取
  - 让类数递增重复进行聚类，选取 $J-C$ 曲线曲率变化最大点所对应的类数（J 为准则函数）

2.5 聚类的算法

2.5.4 动态聚类法　　C-均值法

C-均值算法的改进

- **初始聚类中心的选取可采取的方式**
  - 凭经验选取
  - 将模式随机分为 c 类计算每类类心
  - 按密度大小选取
  - 选取相距最远的 c 个特征点
  - 随机地从 N 个模式中取出部分用谱系聚类法聚成 c 类，以各类类心作为初始类心
  - 由 $c-1$ 类问题得出 $c-1$ 个类心，再找出最远点

- **基于核函数的C-均值算法**

1. 对给定的待分类模式集 $\{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}$ 进行初始划分产生c类；
2. 计算各聚类 ω_j 所含模式数 n_j 均值矢量 $\vec{\mu}_j$ 和协方差阵 Σ_j；
3. 将各模式 $\vec{x}_i$ 按最小距离原则分划到某一聚类中。这里采用最小误判概率准则下正态分布情况的判决规则，计算模式 $\vec{x}$ 到 ω_j 的距离。

$$d(\vec{x},\omega_j) = \ln|\Sigma_j| + (\vec{x}-\vec{\mu}_j)'\Sigma_j^{-1}(\vec{x}-\vec{\mu}_j) - 2\ln\frac{n_j}{N}$$

如果 $d(\vec{x},\omega_k) = \min_j[d(\vec{x},\omega_j)]$ 则 $\vec{x} \in \omega_k$

4. 如果没有模式改变其类别，则停止算法；否则转至2.。

2.5 聚类的算法

2.5.4 动态聚类法

进一步的动态聚类可以采用ISODATA算法，其基本思想是每轮迭代时，样本重新调整类别之后计算类内及类间有关参数，通过和门限比较确定两类的合并或分裂，不断地“自组织”，在满足设计参数条件下，使模式到类心的距离平方和最小。

作业

上机实验：P67：2.5

谢 谢！