厦门大学信息学院 本科选修课
2021-2022 第二学期

模 式 识 别
Pattern Recognition

主 讲：王 程

第二章 聚类分析
2.2　动态聚类法

2.2 动态聚类法

- **动态聚类: 聚类过程中，*类心*和*类别*可变**
  - 技术要点

(1) 确定模式和聚类的距离测度

(2) 建立评估聚类质量的准则函数

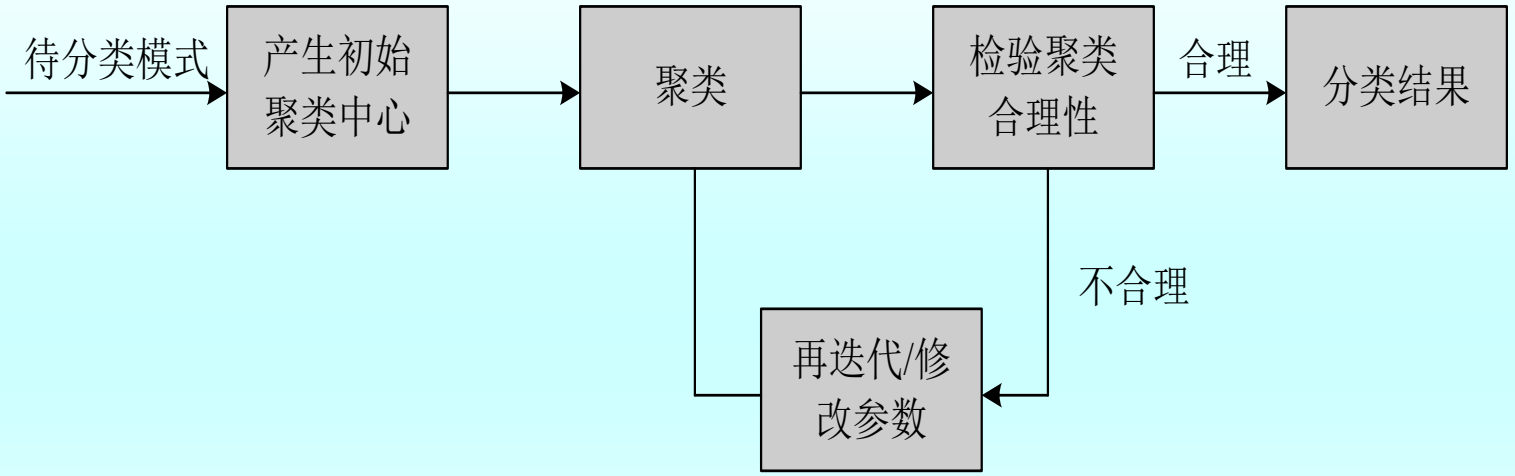
(3) 确定分划、聚类合并或分裂的规则和策略

  - 基本步骤

- 选取中心和有关参数进行初始聚类
- 计算模式和聚类的距离，调整模式类别
- 计算聚类参数，删除、合并或分裂聚类
- 运用迭代算法动态改变模式类别和聚类中心，直至达到聚类结束条件

2.2 动态聚类法

- 原理框图



- 代表算法

C-均值法 ISODATA算法

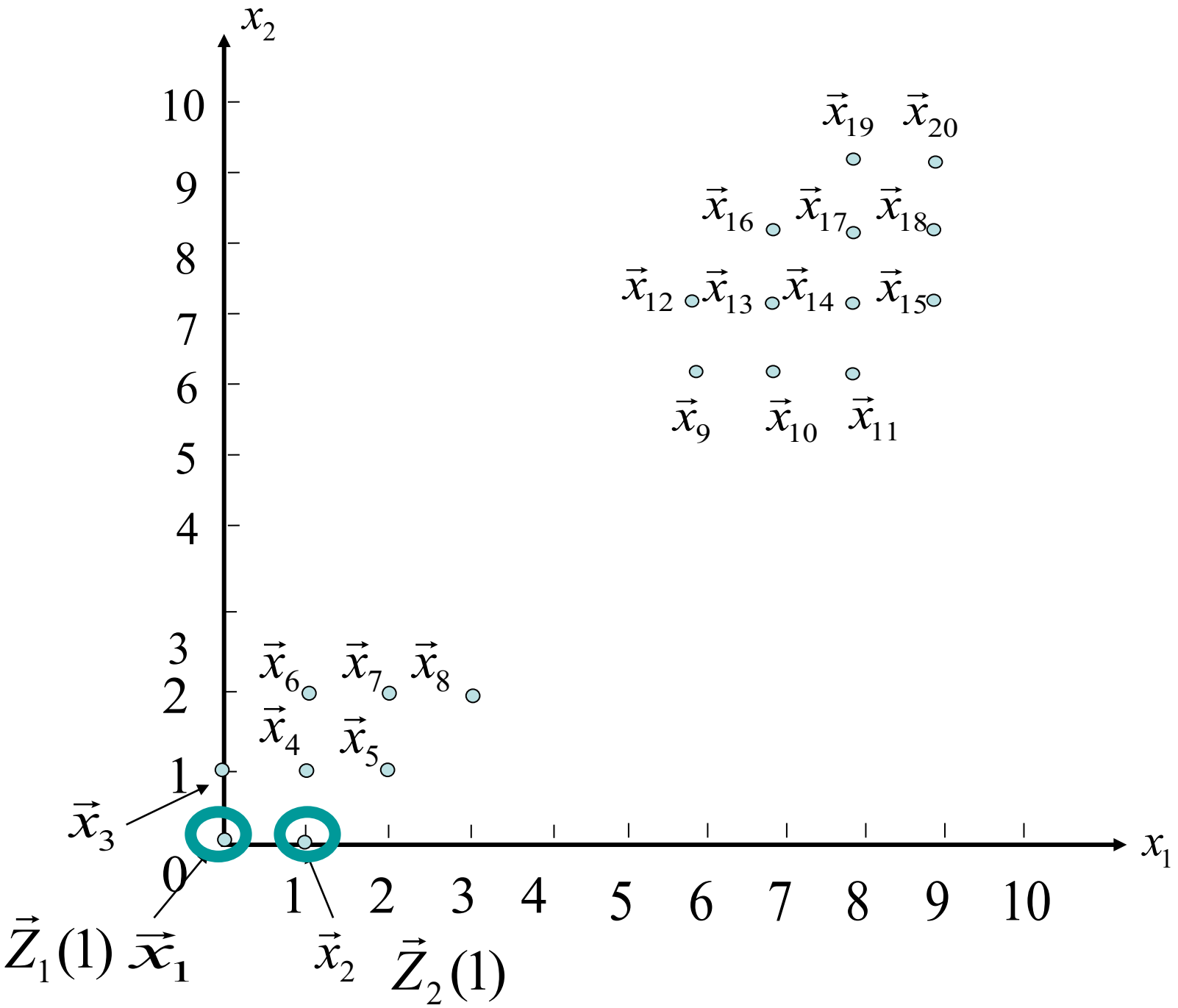
（一）C-均值法(C-Means)

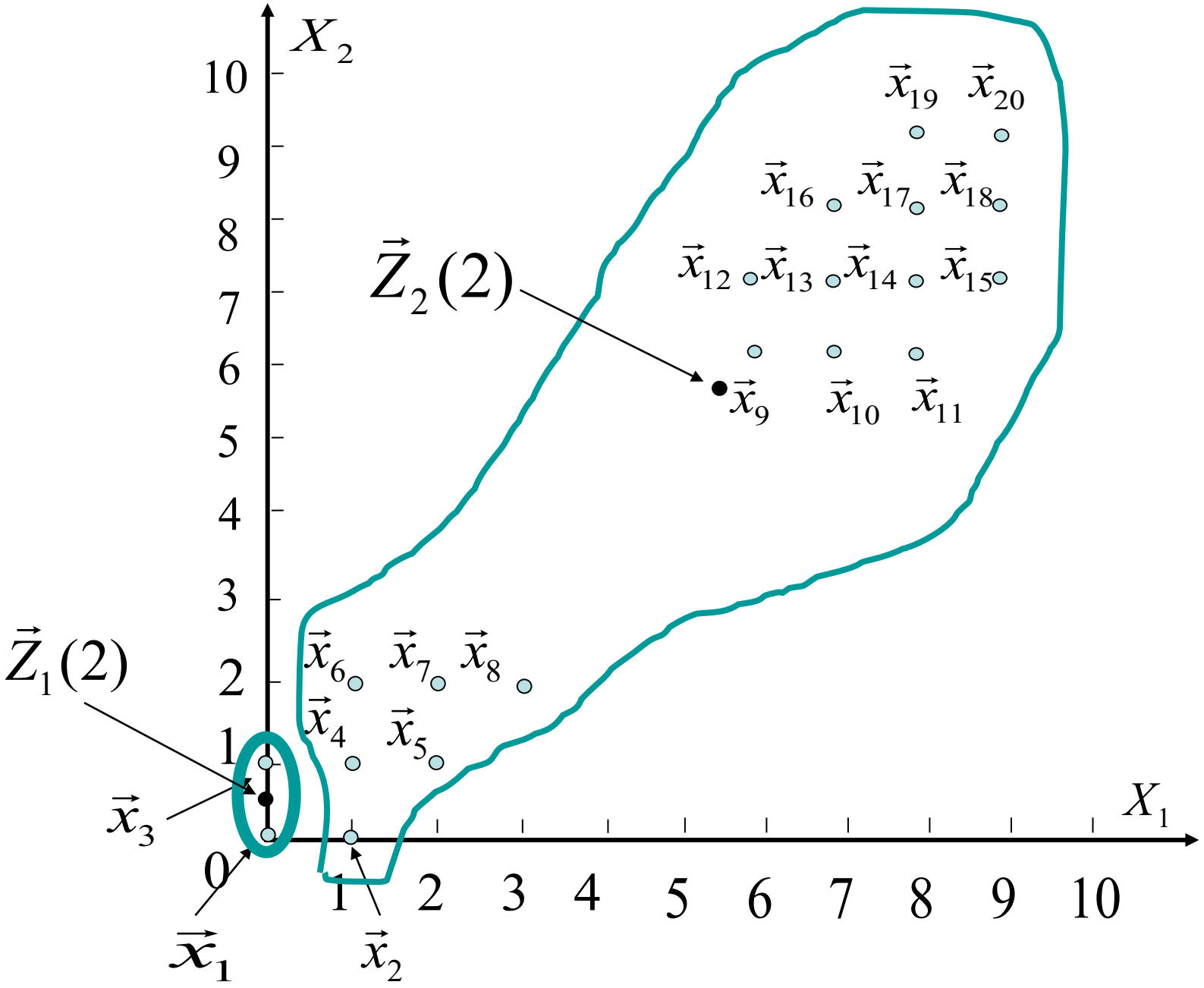
- 条件和约定

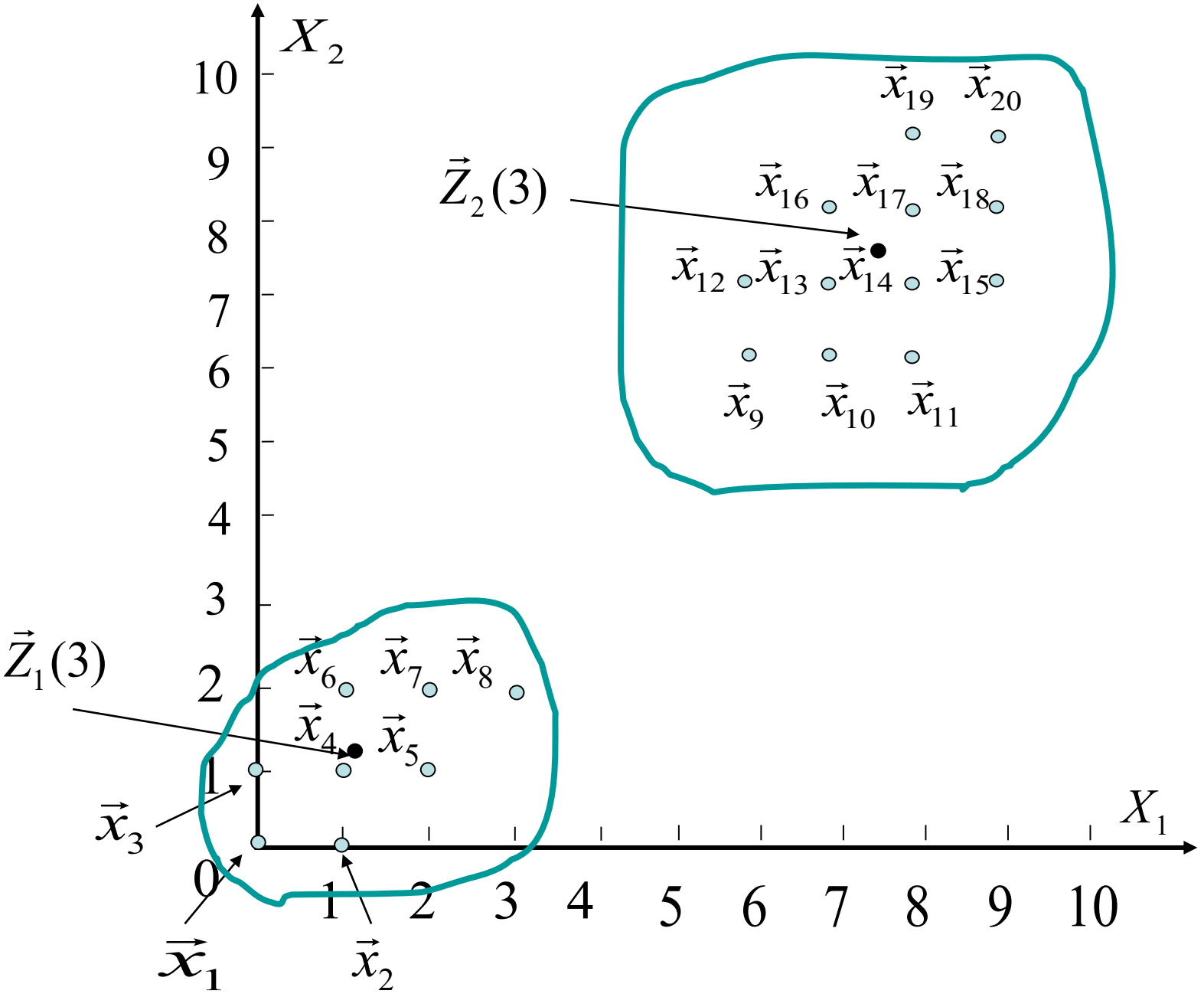
设待分类的模式特征矢量为 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$
类数 c 事先取定。

- 基本思想

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别分配到各类；不断地计算类心和调整各模式的类别，直至各模式到其判属类别中心的距离平方和最小。


x_2
10
9
8
7
6
5
4
3
2
1
0
$\vec{x}_{19}$ $\vec{x}_{20}$
$\vec{x}_{16}$ $\vec{x}_{17}$ $\vec{x}_{18}$
$\vec{x}_{12}$ $\vec{x}_{13}$ $\vec{x}_{14}$ $\vec{x}_{15}$
$\vec{x}_9$ $\vec{x}_{10}$ $\vec{x}_{11}$
$\vec{x}_6$ $\vec{x}_7$ $\vec{x}_8$
$\vec{x}_4$ $\vec{x}_5$
$\vec{x}_3$
x_1
1 2 3 4 5 6 7 8 9 10
$\vec{Z}_1(1)$ $\vec{x}_1$
$\vec{x}_2$ $\vec{Z}_2(1)$


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


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

3. 算法步骤

C-均值法

- 算法步骤

（1）任选 c 个模式特征矢量作为初始聚类中心。

（2）将待分类模式逐个按最小距离准则分划到c类

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

$$\text{则判} \quad \boldsymbol{x}_i \in \omega_l^{(k+1)}$$

式中，$d_{ij}^{(k)}$ 表示 $\boldsymbol{x}_i$ 和 $\omega_j^{(k)}$ 的中心 $\boldsymbol{z}_j^{(k)}$ 的距离，上标表示迭代次数。

3. 算法步骤

C-均值法

- 算法步骤

（3）计算重新分类后的类心

$$z_j^{(k+1)} = \frac{1}{n_j^{(k+1)}} \sum_{x_i \in \omega_j^{(k+1)}} x_i$$

$n_j^{(k+1)}$表示 $\omega_j^{(k+1)}$ 类中的模式个数。

（4）如果 $z_j^{(k+1)} = z_j^{(k)} \ (j = 1, 2, \cdots, c)$ 则结束；否则，

$k = k+1$ ，转至步骤（2）。

例2.5.2

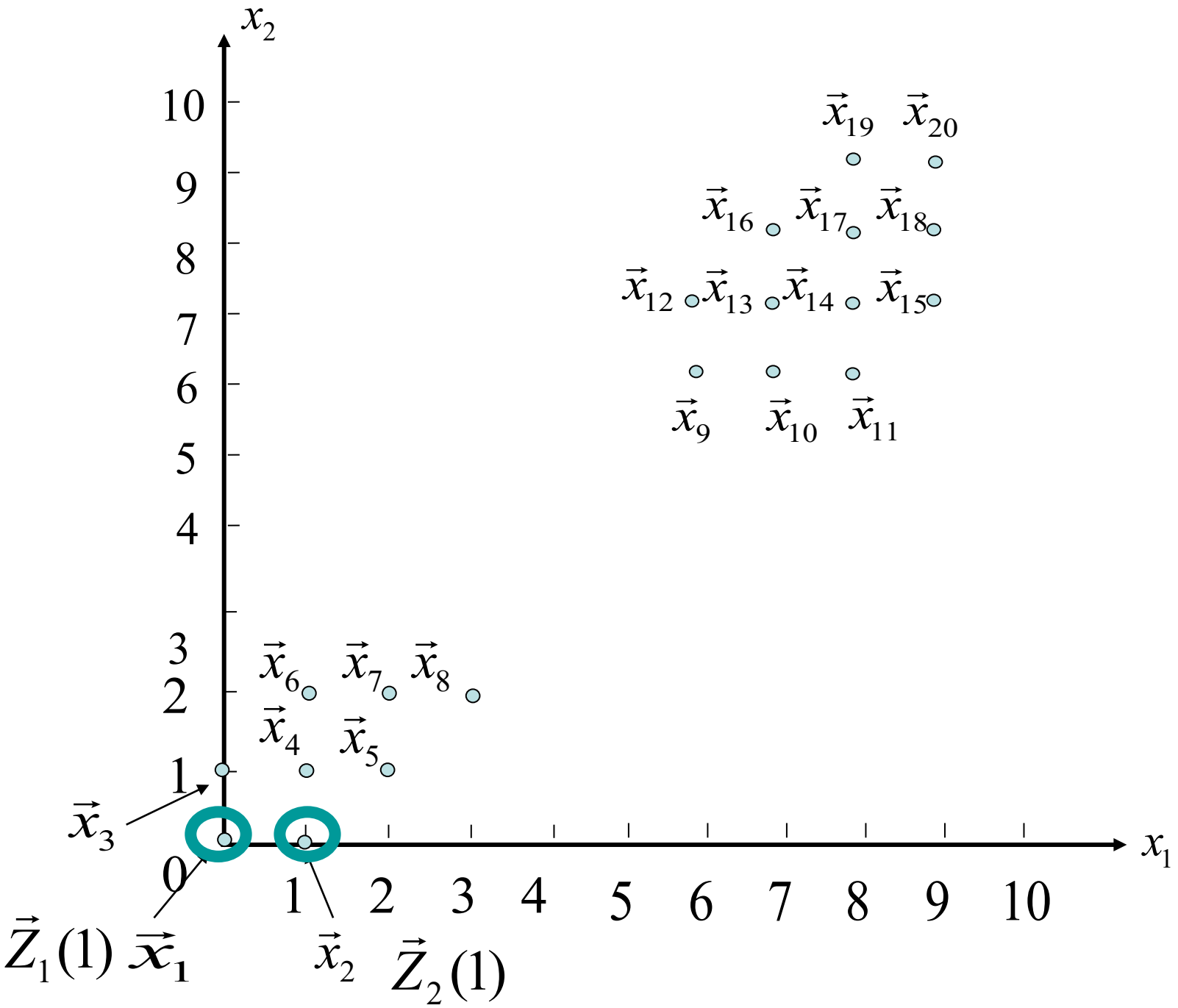
- 已知有20个样本，每个样本有2个特征，数据分布如下图，使用C－均值法实现样本分类（C=2）。

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

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
10
9
8
7
6
5
4
3
2
1
0
$\vec{x}_{19}$ $\vec{x}_{20}$
$\vec{x}_{16}$ $\vec{x}_{17}$ $\vec{x}_{18}$
$\vec{x}_{12}$ $\vec{x}_{13}$ $\vec{x}_{14}$ $\vec{x}_{15}$
$\vec{x}_9$ $\vec{x}_{10}$ $\vec{x}_{11}$
$\vec{x}_6$ $\vec{x}_7$ $\vec{x}_8$
$\vec{x}_4$ $\vec{x}_5$
$\vec{x}_3$
x_1
1 2 3 4 5 6 7 8 9 10
$\vec{Z}_1(1)$ $\vec{x}_1$
$\vec{x}_2$ $\vec{Z}_2(1)$

第二步

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

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

因为 $\| \vec{x}_1 - \vec{Z}_1(1) \| < \| \vec{x}_1 - \vec{Z}_2(1) \|$ **所以** $\vec{x}_1 \in \omega_1^{(1)}$

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

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

因为 $\| \vec{x}_2 - \vec{Z}_1(1) \| > \| \vec{x}_2 - \vec{Z}_2(1) \|$ **所以** $\vec{x}_2 \in \omega_2^{(1)}$

第二步

同理

$$\because \| \vec{x}_3 - \vec{Z}_1(1) \| = 1 < \| \vec{x}_3 - \vec{Z}_2(1) \| = 2 \quad \therefore \vec{x}_3 \in \omega_1^{(1)}$$

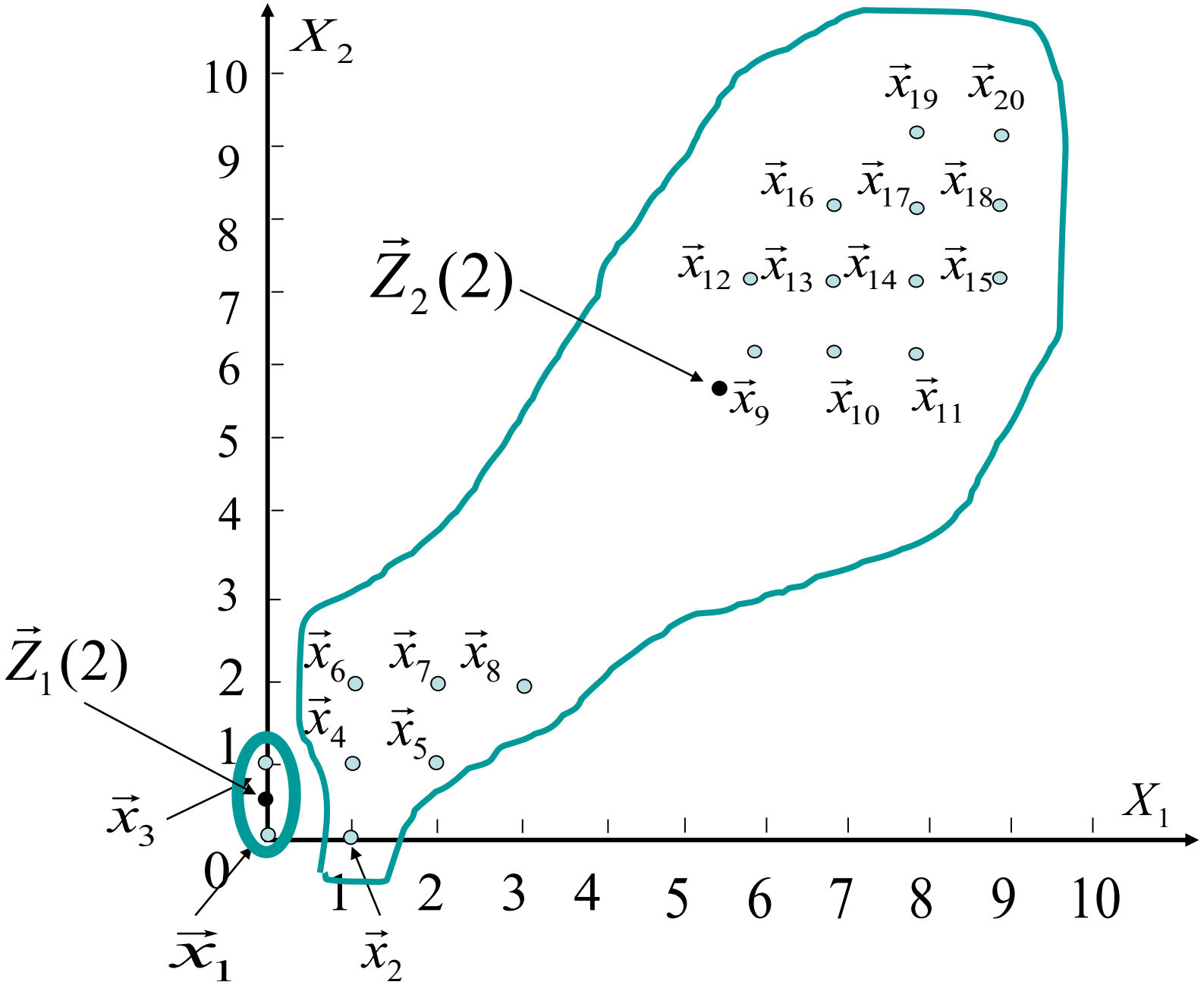
$$\because \| \vec{x}_4 - \vec{Z}_1(1) \| = 2 > \| \vec{x}_4 - \vec{Z}_2(1) \| = 1 \quad \therefore \vec{x}_4 \in \omega_2^{(1)}$$

$$\vec{x}_5, \vec{x}_6, \cdots, \vec{x}_{20} \in \omega_2^{(1)}$$

由此得到两类：

$$\omega_1^{(1)} = \{\vec{x}_1, \vec{x}_3\}, N_1 = 2$$

$$\omega_2^{(1)} = \{\vec{x}_2, \vec{x}_4, \vec{x}_5, \cdots, \vec{x}_{20}\}, N_2 = 18$$


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

第三步

$$\vec{Z}_1(2) = \frac{1}{N_1}\sum_{\vec{x}_i \in \omega_1^{(1)}} \vec{x}_i = \frac{1}{2}(\vec{x}_1 + \vec{x}_3) = \frac{1}{2}\left[\begin{pmatrix}0\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}\right]$$

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

$$\vec{Z}_2(2) = \frac{1}{N_2}\sum_{\vec{x}_i \in \omega_2^{(1)}} \vec{x}_i = \frac{1}{18}(\vec{x}_2 + \vec{x}_4 + \vec{x}_5 + \ldots + \vec{x}_{20})$$

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

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，

$$\omega_1^{(2)} = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_8), N_1 = 8$$

$$\omega_2^{(2)} = (\vec{x}_9, \vec{x}_{10}, \ldots, \vec{x}_{20}), N_2 = 12$$

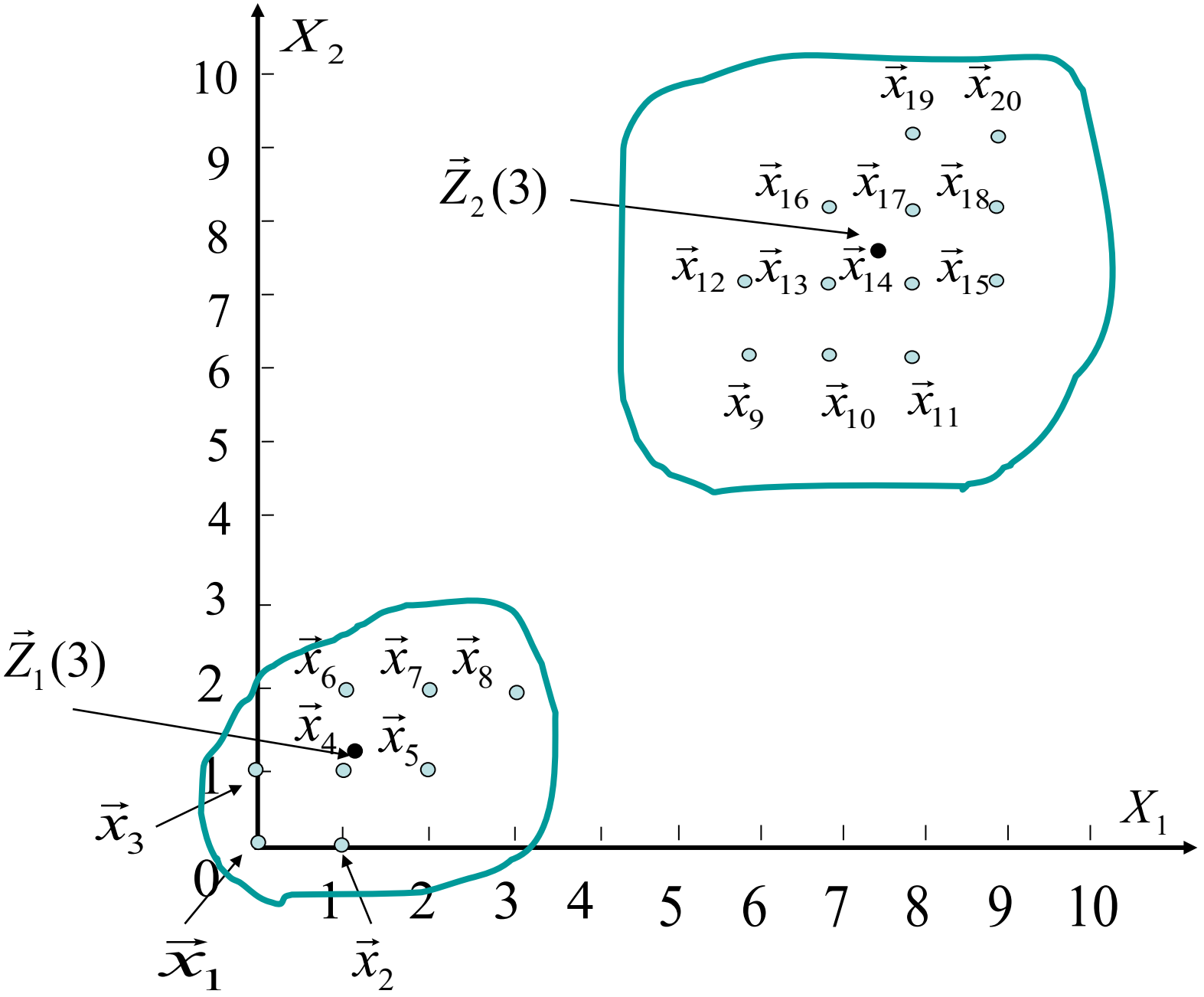
第三步

$$\vec{Z}_1(3) = \frac{1}{N_1} \sum_{\vec{x}_i \in \omega_1^{(2)}} \vec{x}_i = \frac{1}{8}(\vec{x}_1 + \vec{x}_2 + \vec{x}_3 + \ldots + \vec{x}_8)$$

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

$$\vec{Z}_2(3) = \frac{1}{N_2} \sum_{\vec{x}_i \in \omega_2^{(2)}} \vec{x}_i = \frac{1}{12}(\vec{x}_9 + \vec{x}_{10} + \ldots + \vec{x}_{20})$$

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


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

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$\vec{x}_1, \vec{x}_2, ..., \vec{x}_{20}$到$\vec{Z}_1(3), \vec{Z}_2(3)$的距离，

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

重新分为二类$\omega_1^{(3)} = (\vec{x}_1, \vec{x}_2, ..., \vec{x}_8)$

$\omega_2^{(3)} = (\vec{x}_9, \vec{x}_{10}, ..., \vec{x}_{20}), N_1 = 8, N_2 = 12$

第三步

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

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

计算结束。

4. 算法收敛性分析

C-均值法

- 收敛性分析

在分类过程中不断地计算新分划的类心并按最小距离原则归类可使类内距离平方和极小。

- 算法性能

- 是使模式到其类心距离平方和最小的最佳聚类
- 受类别数目和初始中心的影响，局部最优
- 方法简单，结果尚可

4. 算法收敛性分析

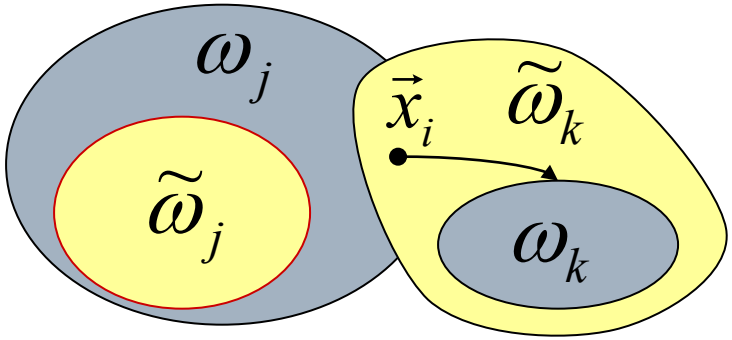
$$J^{(k)}=\sum_{j=1}^{c}\sum_{\vec{x}_i\in\omega_j^{(k)}}\left\|\vec{x}_i-\vec{z}_j^{(k)}\right\|^2$$

算法可使准则函数$J^{(k)}$变小，

- 设某样本 $\vec{x}_i$ 从聚类ω_j移至聚类ω_k中，移出后的集合记为$\tilde{\omega}_j$，移入后的集合记为$\tilde{\omega}_k$。可以证明

$$\tilde{J}_j=J_j-\frac{n_j}{n_j-1}\left\|\vec{x}_i-\vec{m}_j\right\|^2$$

$$\tilde{J}_k=J_k+\frac{n_k}{n_k+1}\left\|\vec{x}_i-\vec{m}_k\right\|^2$$



$$\because \left\|\vec{x}_i-\vec{m}_k\right\|^2<\left\|\vec{x}_i-\vec{m}_j\right\|^2,\quad (\vec{x}_i\text{距离}\omega_k\text{更近})$$

$$\frac{n_k}{n_k+1}\left\|\vec{x}_i-\vec{m}_k\right\|^2<\frac{n_j}{n_j-1}\left\|\vec{x}_i-\vec{m}_j\right\|^2$$

>1

$$\therefore \tilde{J}_j+\tilde{J}_k<J_j+J_k$$

距离平方和准则函数的变化关系证明

$$\because \vec{m} = \frac{1}{n}\sum_{l=1}^{n}\vec{x}_l = \frac{1}{n}\left(\sum_{l=1}^{n-1}\vec{x}_l + \vec{x}_i\right) = \frac{n-1}{n}\tilde{\vec{m}} + \frac{1}{n}\vec{x}_i \quad (1)$$

$$\therefore \vec{x}_i - \tilde{\vec{m}} = \frac{n}{n-1}(\vec{x}_i - \vec{m}) \quad (2)$$

$$J = \sum_{l=1}^{n}\left\|\vec{x}_l - \vec{m}\right\|^2 = \sum_{l=1}^{n-1}\left\|\vec{x}_l - \vec{m}\right\|^2 + \left\|\vec{x}_i - \vec{m}\right\|^2$$

$$= \sum_{l=1}^{n-1}\left\|\vec{x}_l - \frac{n-1}{n}\tilde{\vec{m}} - \frac{1}{n}\vec{x}_i\right\|^2 + \left\|\vec{x}_i - \vec{m}\right\|^2 = \sum_{l=1}^{n-1}\left\|\vec{x}_l - \tilde{\vec{m}} - \frac{1}{n}(\vec{x}_i - \tilde{\vec{m}})\right\|^2 + \left\|\vec{x}_i - \vec{m}\right\|^2$$

$$= \sum_{l=1}^{n-1}[(\vec{x}_l - \tilde{\vec{m}}) - \frac{1}{n-1}(\vec{x}_i - \vec{m})]'[(\vec{x}_l - \tilde{\vec{m}}) - \frac{1}{n-1}(\vec{x}_i - \vec{m})] + \left\|\vec{x}_i - \vec{m}\right\|^2$$

$$= \sum_{l=1}^{n-1}[\left\|\vec{x}_l - \tilde{\vec{m}}\right\|^2 - \frac{2}{n-1}(\vec{x}_l - \tilde{\vec{m}})'(\vec{x}_i - \vec{m}) + \frac{1}{(n-1)^2}(\vec{x}_i - \vec{m})'(\vec{x}_i - \vec{m})] + \left\|\vec{x}_i - \vec{m}\right\|^2$$

$$= \sum_{l=1}^{n-1}\left\|\vec{x}_l - \tilde{\vec{m}}\right\|^2 - 2[(\frac{1}{n-1}\sum_{l=1}^{n-1}\vec{x}_l) - \tilde{\vec{m}}]'(\vec{x}_i - \vec{m}) + \frac{1}{(n-1)}\left\|\vec{x}_i - \vec{m}\right\|^2 + \left\|\vec{x}_i - \vec{m}\right\|^2$$

$$= \tilde{J} - 2(\tilde{\vec{m}} - \tilde{\vec{m}})'(\vec{x}_i - \vec{m}) + \frac{n}{n-1}\left\|\vec{x}_i - \vec{m}\right\|^2$$

$$= \tilde{J} + \frac{n}{(n-1)}\left\|\vec{x}_i - \vec{m}\right\|^2$$

$$\therefore \tilde{J}_j = J_j - \frac{n_j}{(n_j-1)}\left\|\vec{x}_i - \vec{m}_j\right\|^2$$

注：上述推导过程中省略了下标*j*

C均值的计算复杂度

一般流程：

- 先随机设定k个聚类中心（此处的k决定了最终的k类）然后计算n个点到k个点的距离，单个点与点距离计算的复杂度为d
- 聚类完成之后，我们需要进一步迭代。首先计算聚的k个类的聚类中心，然后用找到的k个聚类中心进行再次聚类。
- 当聚类中心不再变动时，聚类停止。
- 设此时迭代次数为t次。

因此可以看出，k-means聚类复杂度为 O=k*n*d*t

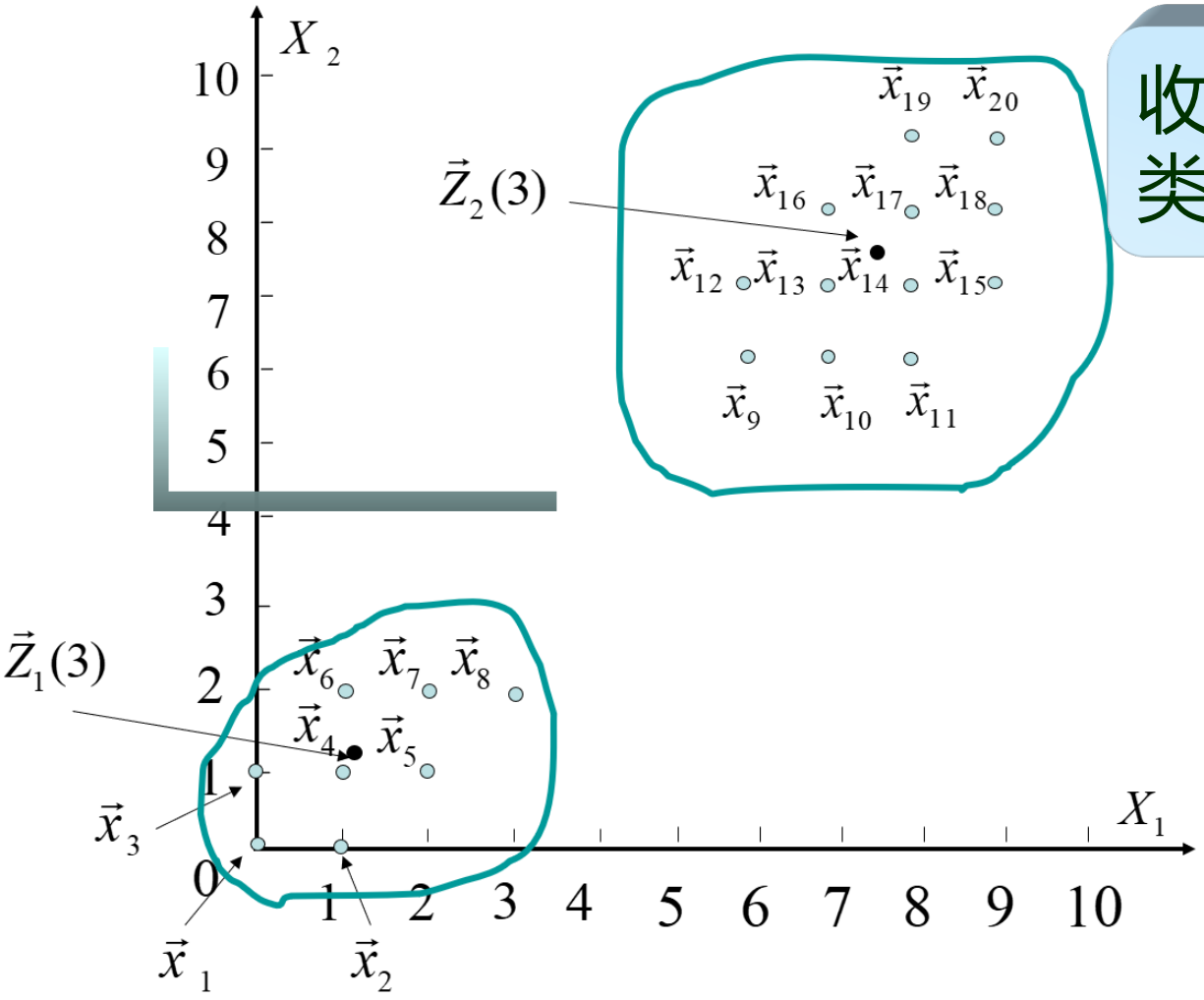
5. 性能

C-均值法

- 算法性能

➢是使模式到其类心距离平方和最小的最佳聚类
➢以确定的类数c、模式输入次序、及选定的初始聚类中心为前提，受此限制结果只是局部最优
➢算法简单，收敛。如模式呈现类内团状分布，效果很好，故应用较多。

小结



收敛特性
类内距离平方和极小。

- 算法性能

作业

1. 编写C均值算法程序

End