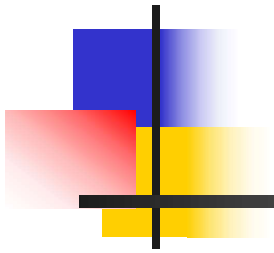


非监督学习方法（聚类分析）

- 基本概念
- 分级聚类法
- C均值聚类法
- 实验介绍

6.1 基本概念

- 监督学习(Supervised Classification)

  利用已知类别的样本进行训练

- 非监督学习(Nonsupervised Classification)

  所用样本没有类别标志（本章主要研究）

- 聚类分析

  按照物以类聚的思想，对未知类别的样本集根据样本之间的相似程度分类，相似的归为一类。

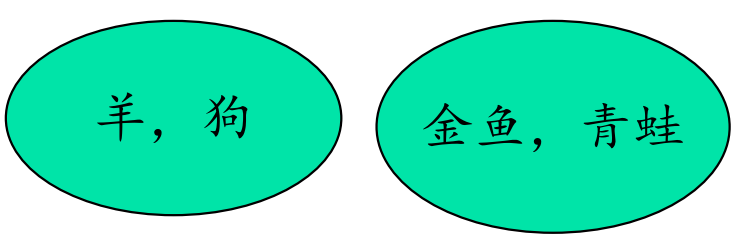
6.1 基本概念

- 什么叫两个样本相似？

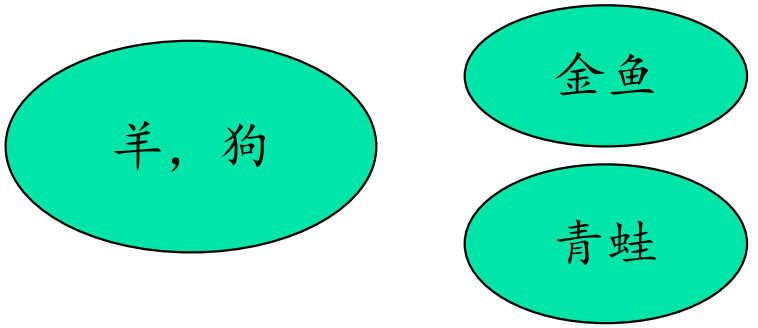
  模式的相似性测度

- 相似到什么程度归为一类？

  聚类准则，聚类准则不同，分类结果也不同



以胎生为准则聚类



以生存环境为准则聚类

6.1 基本概念

- 模式相似性测度

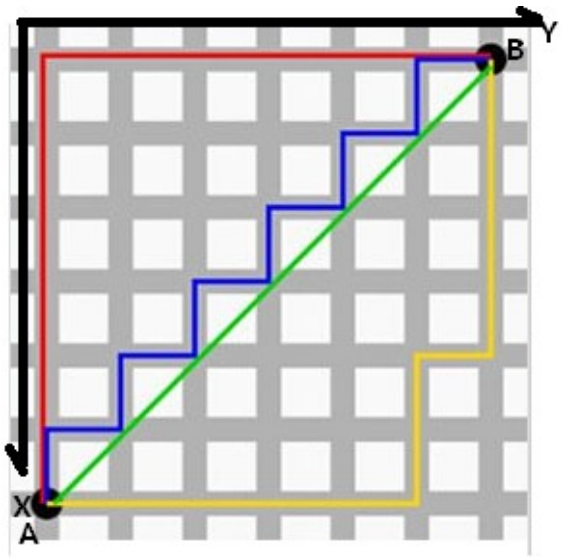
度量同一类样本间的类似性和不属于同一类样本间的差异性。

1、距离度量

欧氏距离：设x^i, x^j为两个样本，则

$$D(\mathbf{x}^i,\mathbf{x}^j)=\left\|\mathbf{x}^i-\mathbf{x}^j\right\|=\left(\sum_{k=1}^{d}(x_k^i-x_k^j)^2\right)^{1/2}$$

街坊（曼哈顿）距离：$D(\mathbf{x}^i,\mathbf{x}^j)=\sum_{k=1}^{d}\left|x_k^i-x_k^j\right|$



6.1 基本概念

- 欧氏距离

  欧氏距离在使用中反映各特征分量的样本个数应均衡。

  例如，取5个样本，其中有4个反映对分类有意义的特征A，只有1个对分类有意义的特征B，欧氏距离的计算结果则主要体现特征A。

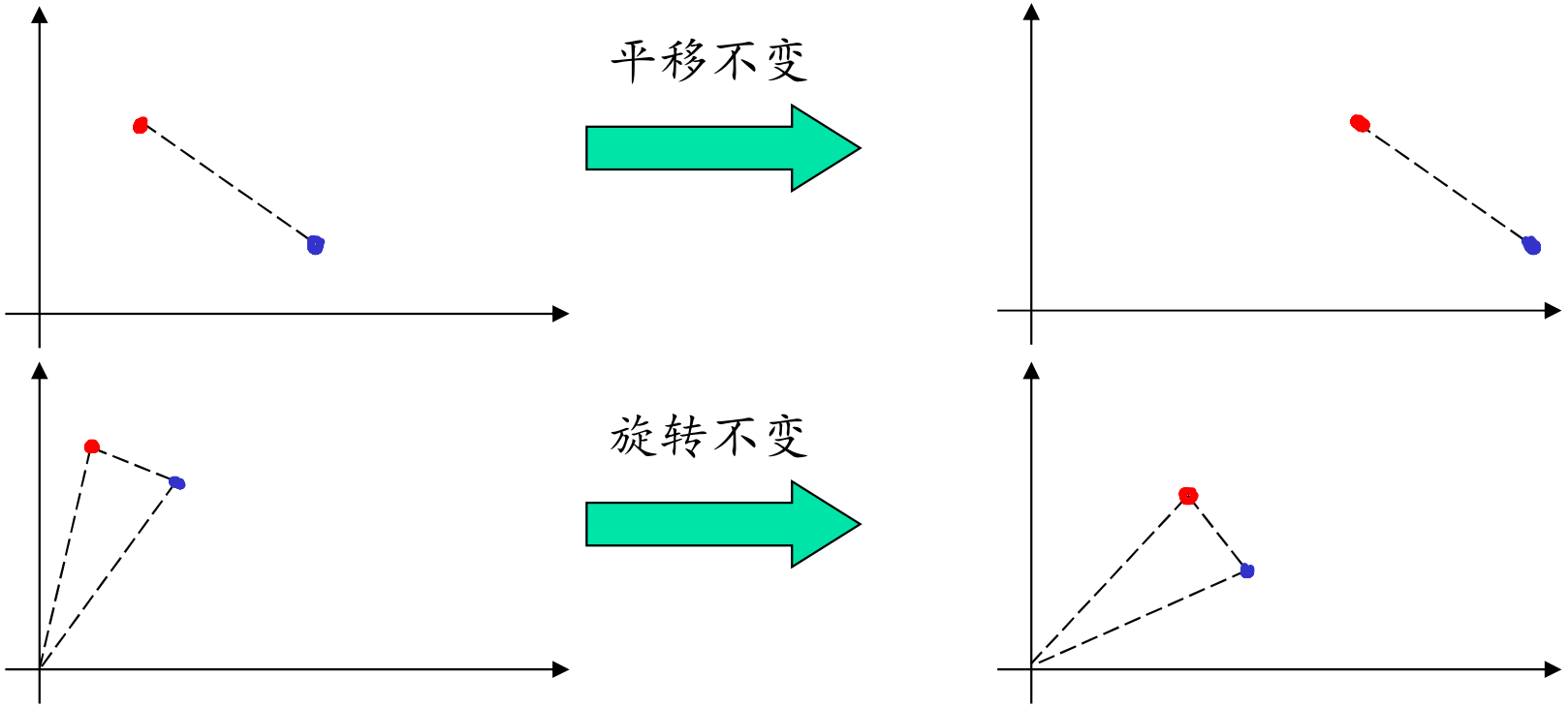
- 马氏距离

  马氏距离排除了不同特征之间相关性的影响。

6.1 基本概念

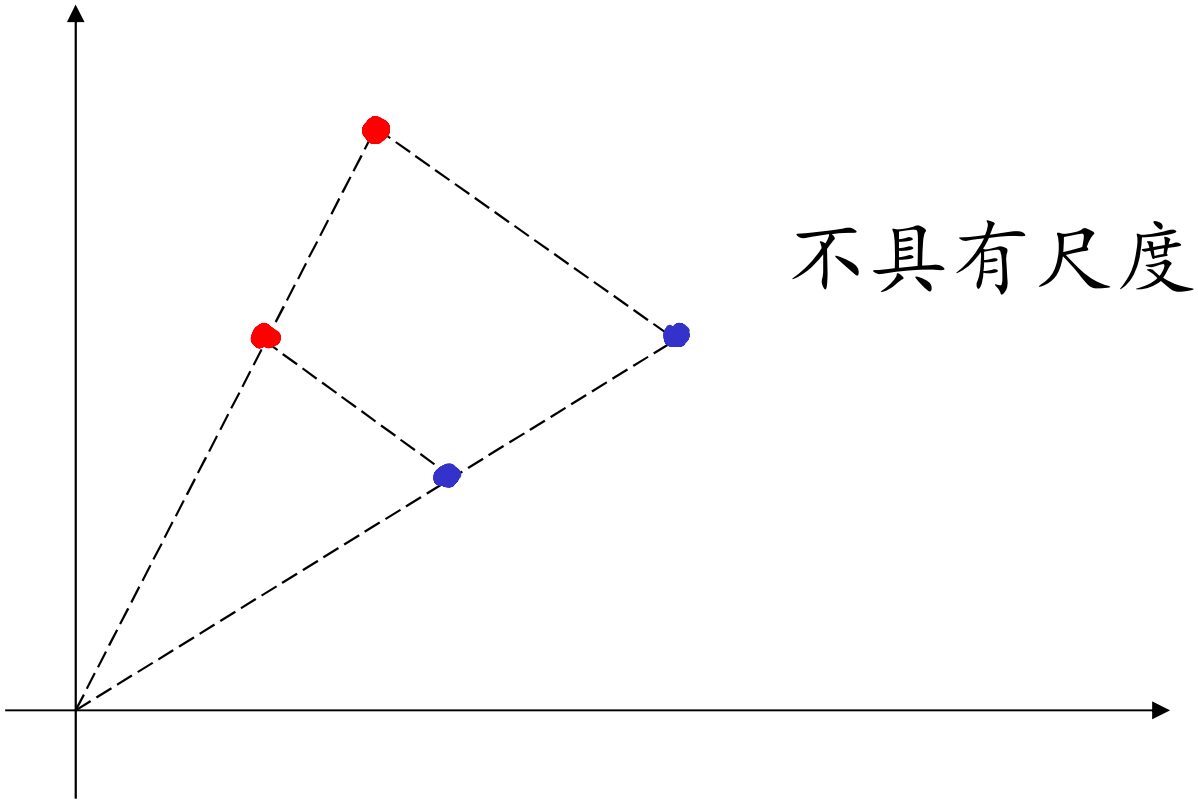
距离测度的优缺点

以距离作为测度具有平移不变性和旋转不变性，但不具尺度不变性。



6.1 基本概念

距离测度的优缺点



不具有尺度不变性

6.1 基本概念

2、角度测度

由内积公式

$$\left(\mathbf{x}^i\right)^T\left(\mathbf{x}^j\right)=\left\|\mathbf{x}^i\right\|\cdot\left\|\mathbf{x}^j\right\|\cos\theta$$

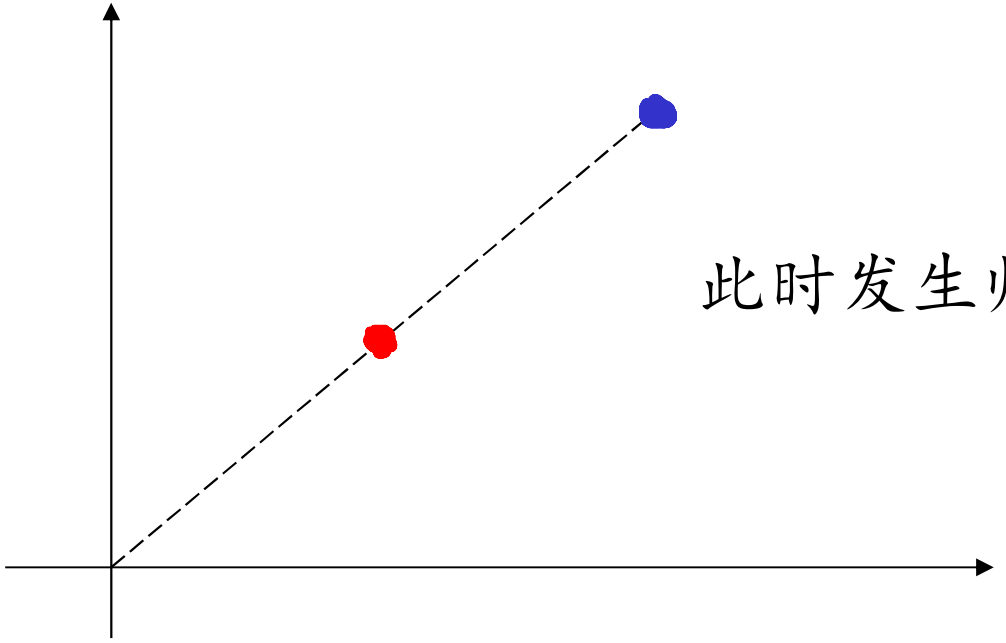
可定义

$$\mathrm{S}\left(\mathbf{x}^i,\ \mathbf{x}^j\right)=\frac{\left(\mathbf{x}^i\right)^T\left(\mathbf{x}^j\right)}{\left\|\mathbf{x}^i\right\|\cdot\left\|\mathbf{x}^j\right\|}=\left(\frac{\mathbf{x}^i}{\left\|\mathbf{x}^i\right\|}\right)^T\left(\frac{\mathbf{x}^j}{\left\|\mathbf{x}^j\right\|}\right)=\cos\theta$$

6.1 基本概念

角度测度具有尺度不变性和旋转不变性，不具有平移不变性。

缺点：



此时发生归为一类的错误

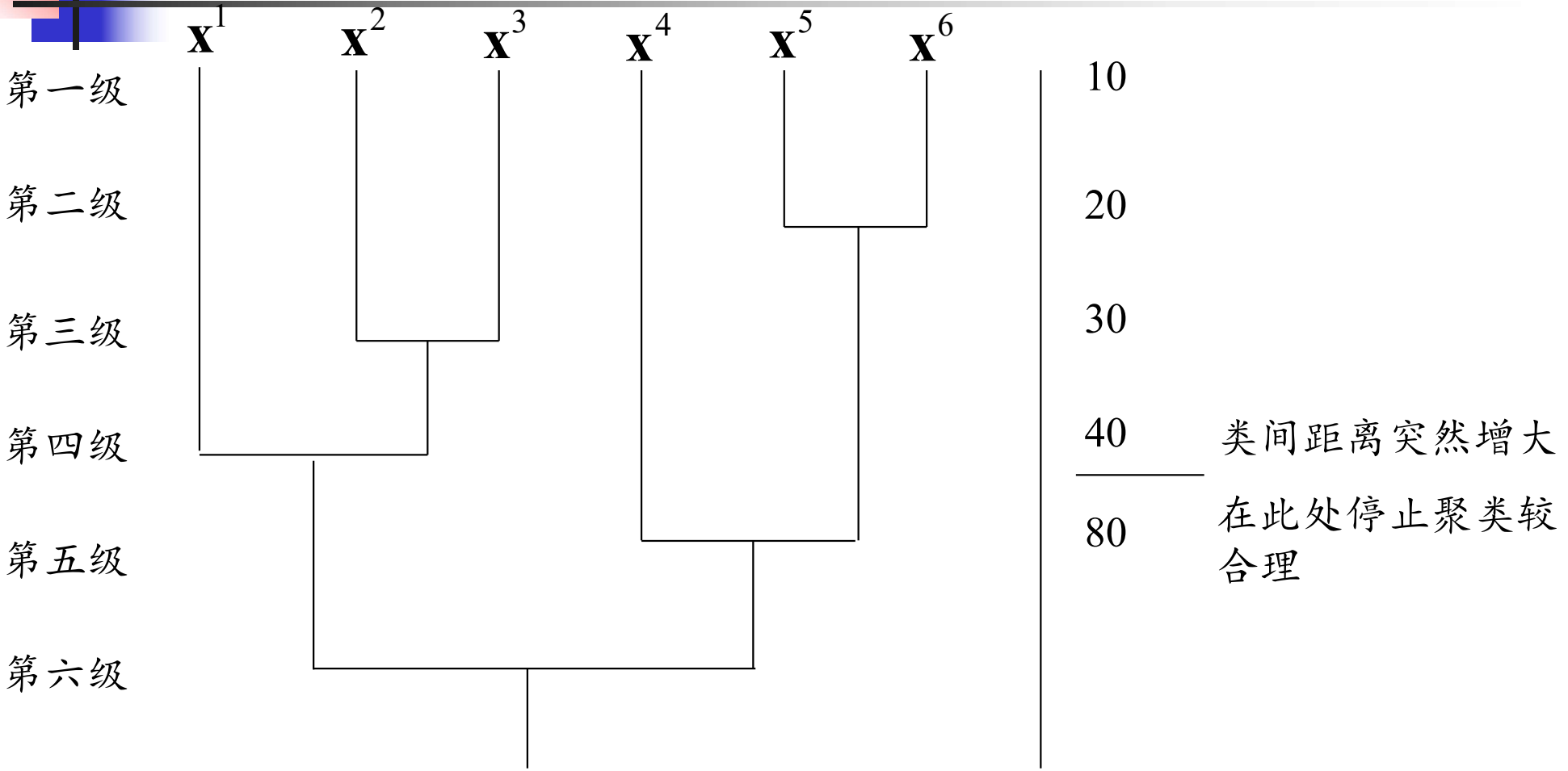
6.2 分级聚类法（系统聚类）

聚类算法分类 { 非迭代的分级聚类算法
迭代的动态聚类算法

聚类分析是把N个没有类别标志的样本分成若干类

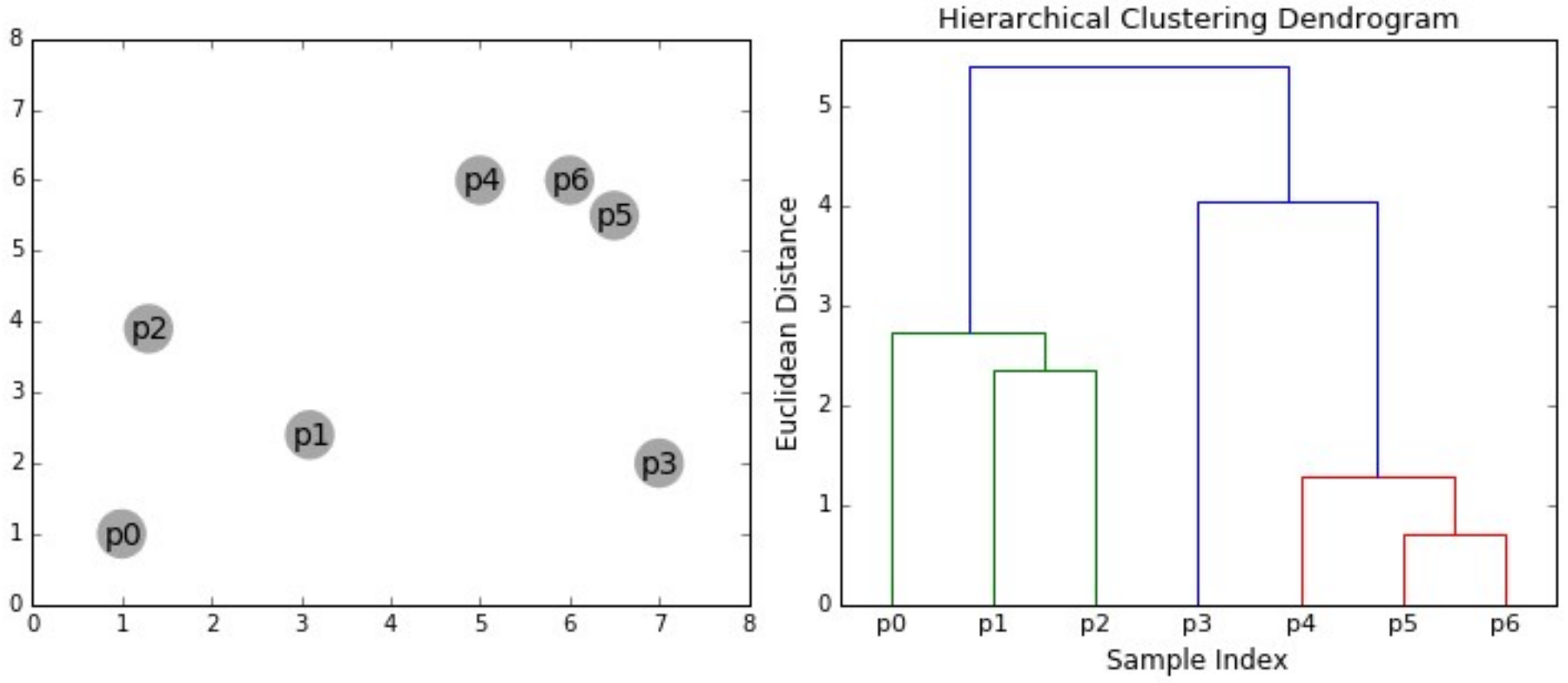
两种极端情况 { N个样本分成N类
⋮ ⇐ 划分序列
N个样本分成1类

6.2 分级聚类法（系统聚类）



缺陷：某一样本若在某级划分中归入了某一类，则在后面的划分中，它永远属于该类。

分级聚类法演示（系统聚类）



6.2 分级聚类法（系统聚类）

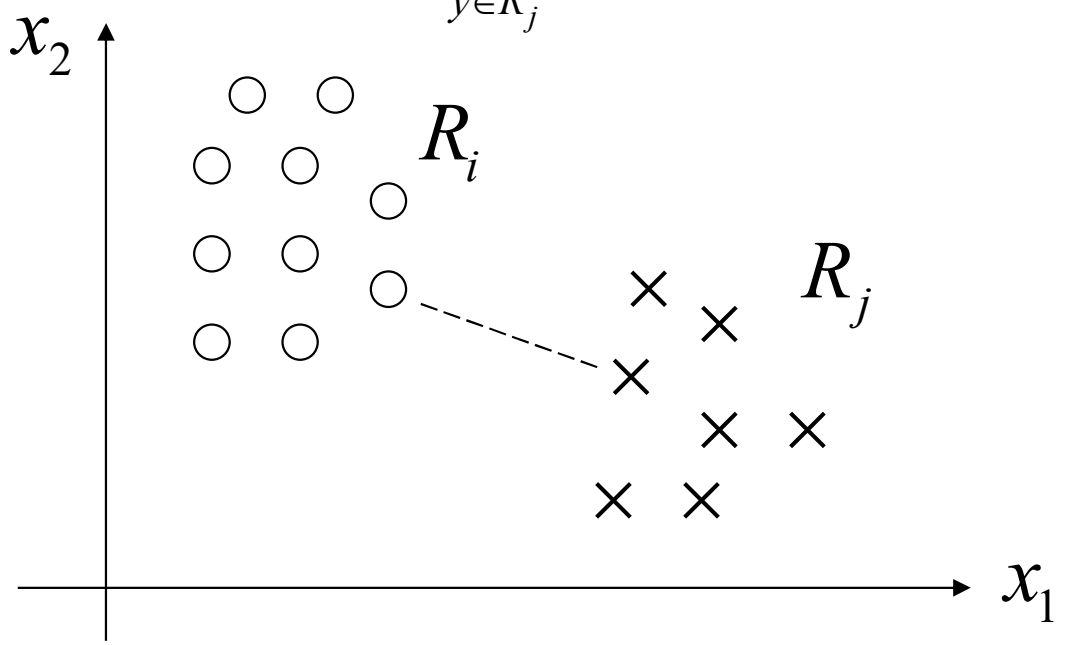
- 分级聚类需要解决的两个问题
  - 如何定义类间距离？
  - 何时停止聚类？
- 聚类停止的两个可选条件
  - 设定聚类类别数，达到则停止—类别数已知情况
  - 类间距离超过预定的阈值—类别数未知情况

6.2 分级聚类法（系统聚类）

- 类间距离定义（区别于样本间距离）

（1）最近距离

$$\Delta(R_i, R_j) = \min_{\substack{x \in R_i \\ y \in R_j}} \{D(x, y)\}$$



6.2 分级聚类法（系统聚类）

- 类间距离定义（区别于样本间距离）

（2）最远距离

$$\Delta(R_i, R_j) = \max_{\substack{x \in R_i \\ y \in R_j}} \{D(x, y)\}$$

（3）均值距离

$$\Delta(R_i, R_j) = D(m_i, m_j)$$

选择不同的距离，可得到不同的聚类的结果，用不同的方法试一下，选一个合理的距离定义。

6.2 分级聚类法（系统聚类）

(1) 初始时，设置 $R_j = \{\mathbf{x}^j\}, \forall j \in I, I = \{1,2,\cdots,N\}$

(2) 在集合 $\{R_j \mid j \in I\}$ 中找到一对满足条件

$$\Delta(R_i, R_k) = \min\{\Delta(R_j, R_l)\}$$ 的聚类 R_i 和 R_k

(3) 把 R_i 并入 R_k，去掉 R_i

(4) 把 i 从指标集 I 中删除，若 I 的基数等于 C，算法终止，否则转(2)

对应于聚为 C 类的情况

6.2 分级聚类法（系统聚类）

例：给出六个五维模式样本如下，按最小距离准则进行系统聚类分类并画出聚类过程树。最小距离准则指类间距离采用最小距离定义。

$$\mathbf{x}^1:0,3,1,2,0 \quad \mathbf{x}^2:1,3,0,1,0 \quad \mathbf{x}^3:3,3,0,0,1$$

$$\mathbf{x}^4:1,1,0,2,0 \quad \mathbf{x}^5:3,2,1,2,1 \quad \mathbf{x}^6:4,1,1,1,0$$

6.2 分级聚类法（系统聚类）

解：1.将每个样本看成一类，则有六类模式样本，即

$$
\begin{array}{c|cccccc}
 & G_1^{(0)} & G_2^{(0)} & G_3^{(0)} & G_4^{(0)} & G_5^{(0)} & G_6^{(0)} \\
G_1^{(0)} & 0 & & & & & \\
G_2^{(0)} & \sqrt{3}^{*} & 0 & & & & \\
G_3^{(0)} & \sqrt{15} & \sqrt{6} & 0 & & & \\
G_4^{(0)} & \sqrt{6} & \sqrt{5} & \sqrt{13} & 0 & & \\
G_5^{(0)} & \sqrt{11} & \sqrt{8} & \sqrt{6} & \sqrt{7} & 0 & \\
G_6^{(0)} & \sqrt{21} & \sqrt{14} & \sqrt{8} & \sqrt{11} & \sqrt{4} & 0
\end{array}
$$

6.2 分级聚类法（系统聚类）

2.上表最小距离为 $\sqrt{3}$，是$G_1^{(0)}$与$G_2^{(0)}$之间距离，合为一类，得

	$G_1^{(1)}$	$G_2^{(1)}$	$G_3^{(1)}$	$G_4^{(1)}$	$G_5^{(1)}$
$G_1^{(1)}$	0				
$G_2^{(1)}$	$\sqrt{6}$	0			
$G_3^{(1)}$	$\sqrt{5}$	$\sqrt{13}$	0		
$G_4^{(1)}$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0	
$G_5^{(1)}$	$\sqrt{14}$	$\sqrt{8}$	$\sqrt{11}$	$\sqrt{4}^{*}$	0

$$G_1^{(1)}=\{x^1,x^2\},G_2^{(1)}=\{x^3\}$$

$$\min\{\Delta(x^1,x^3)=\sqrt{15},\Delta(x^2,x^3)=\sqrt{6}\}$$

6.2 分级聚类法（系统聚类）

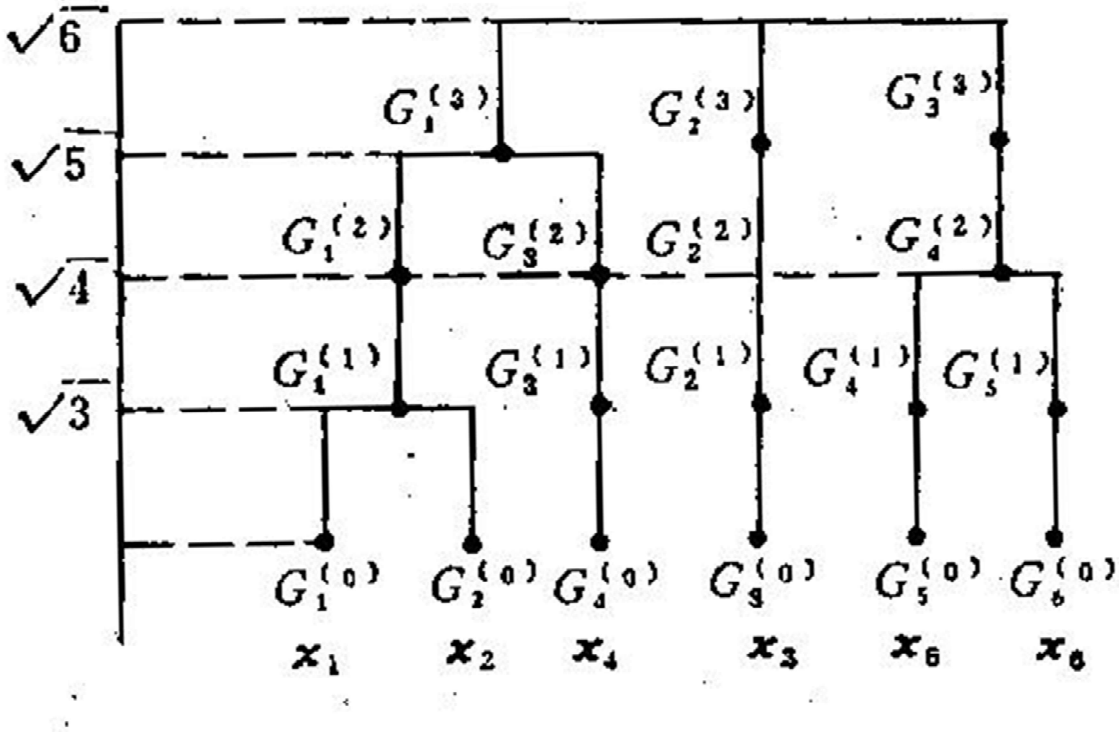
3.上表最小距离为 $\sqrt{4}$，合并$G_4^{(1)}$与$G_5^{(1)}$，得

	$G_1^{(2)}$	$G_2^{(2)}$	$G_3^{(2)}$	$G_4^{(2)}$
$G_1^{(2)}$	0			
$G_2^{(2)}$	$\sqrt{6}$	0		
$G_3^{(2)}$	$\sqrt{5}^{*}$	$\sqrt{13}$	0	
$G_4^{(2)}$	$\sqrt{8}$	$\sqrt{6}$	$\sqrt{7}$	0

6.2 分级聚类法（系统聚类）

4.上表最小距离$\sqrt{5}$，合并$G_1^{(2)}$与$G_3^{(2)}$，得

	$G_1^{(3)}$	$G_2^{(3)}$	$G_3^{(3)}$
$G_1^{(3)}$	0		
$G_2^{(3)}$	$\sqrt{6}^*$	0	
$G_3^{(3)}$	$\sqrt{7}$	$\sqrt{6}^*$	0

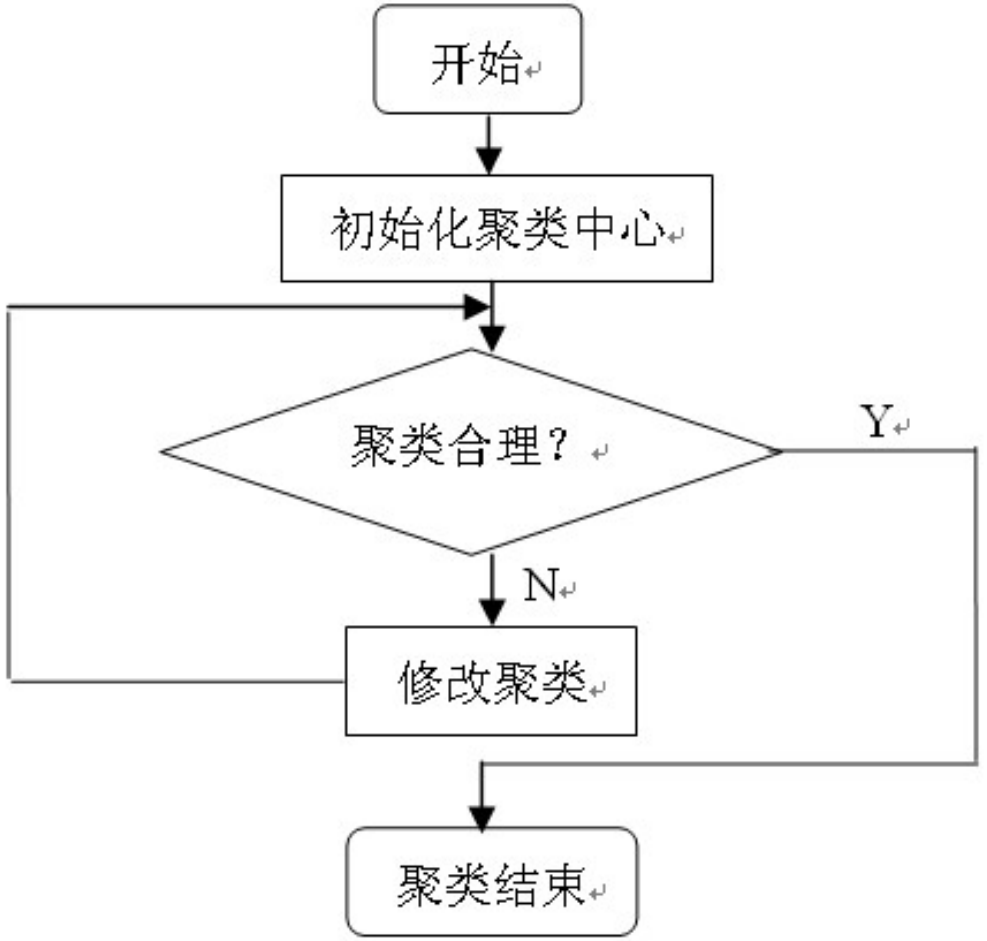


系统聚类树

6.2 C均值算法（动态聚类法）

- 系统聚类缺陷

系统聚类中如样本被划入某一类，则在以后各级划分中，样本始终属于该类，动态聚类可动态调整类属，是普遍采用的方法。



动态聚类流程

6.2 C均值算法（动态聚类法）

- 初始化聚类中心

(1)根据具体问题，凭经验从样本集中选出C个比较合适的样本作为初始聚类中心。

(2)用前C个样本作为初始聚类中心

(3)将全部样本随机地分成C类，计算每类的样本均值，将样本均值作为初始聚类中心。

6.2 C均值算法（动态聚类法）

- 初始聚类

(1)按就近原则将样本归入各聚类中心所代表的类中。(批处理)

(2)取一样本，将其归入与其最近的聚类中心的那一类中，重新计算样本均值，更新聚类中心。然后取下一样本，重复操作，直至所有样本归入相应类中。（单样本）

6.2 C均值算法（动态聚类法）

- 聚类准则函数

误差平方和准则函数

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in \Gamma_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

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

最小方差划分

6.2 C均值算法（动态聚类法）

算法步骤：

已知样本集为

（1）给定类别数C和允许误差Emax, k←1

（2）初始化聚类中心 $\mathbf{m}_i(k), i=1,2,\cdots,c$

（3）修正 $d_{ji}=\begin{cases}1 & \left\|\mathbf{x}^j-\mathbf{m}_i(k)\right\|^2=\min\limits_l\left\{\left\|\mathbf{x}^j-\mathbf{m}_l(k)\right\|^2\right\}\\ 0 & \text{其它}\end{cases}$

$i=1,2,\cdots,c;\ j=1,2,\cdots,N$

（4）修正聚类中心 $\mathbf{m}_i(k+1)=\sum_{j=1}^{N}d_{ji}\mathbf{x}^j/\sum_{j=1}^{N}d_{ji}$

（5）计算误差 $e=\sum_{i=1}^{c}\left\|\mathbf{m}_i(k+1)-\mathbf{m}_i(k)\right\|^2$,如e<Emax, 结束,否则转（3）

6.2 C均值算法（动态聚类法）

- 聚类结果

$$若 d_{ji}=1, \mathbf{x}^j \in \omega_i$$

- 单样本改进

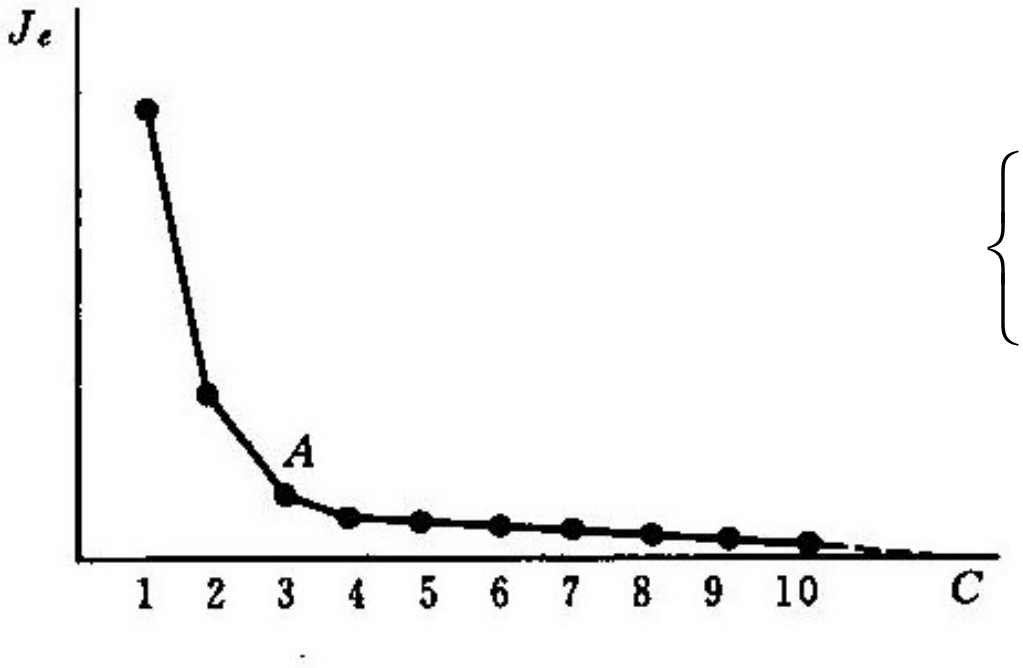
每调整一个样本的类别就重新计算一次各类的聚类中心

$$d_{ji}=\begin{cases}1 & \left\|\mathbf{x}^j-\mathbf{m}_i(k)\right\|^2=\min_l\left\{\left\|\mathbf{x}^j-\mathbf{m}_l(k)\right\|^2\right\} \\ 0 & 其它\end{cases}$$

$i=1,2,\cdots,c$　　只调整一个样本

6.2 C均值算法（动态聚类法）

- 聚类类别数的确定

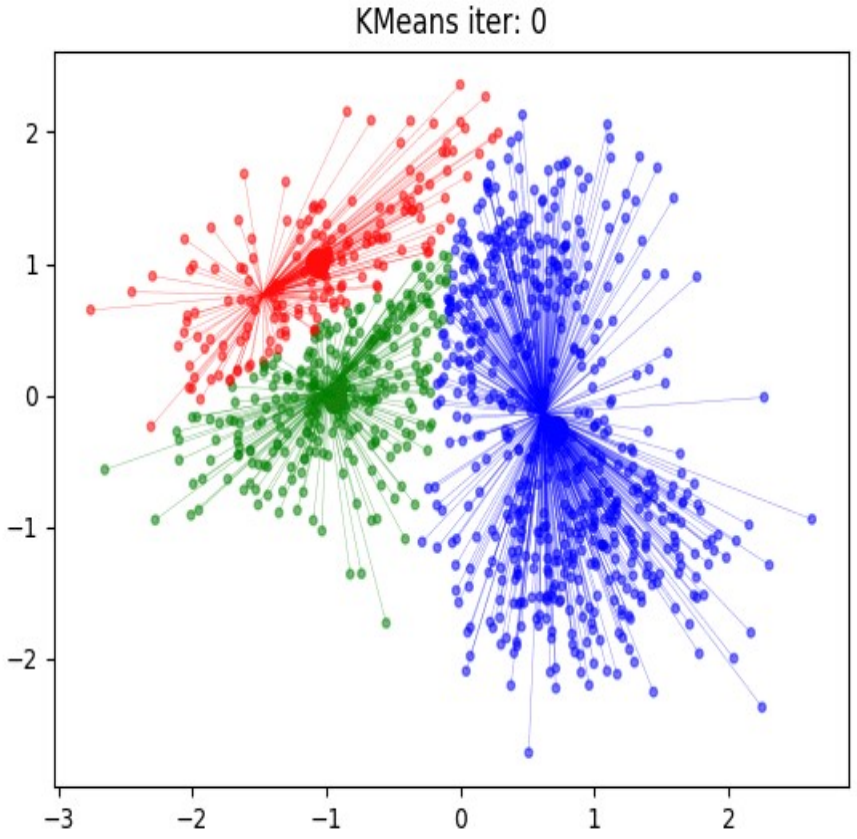


实验方法确定类别数

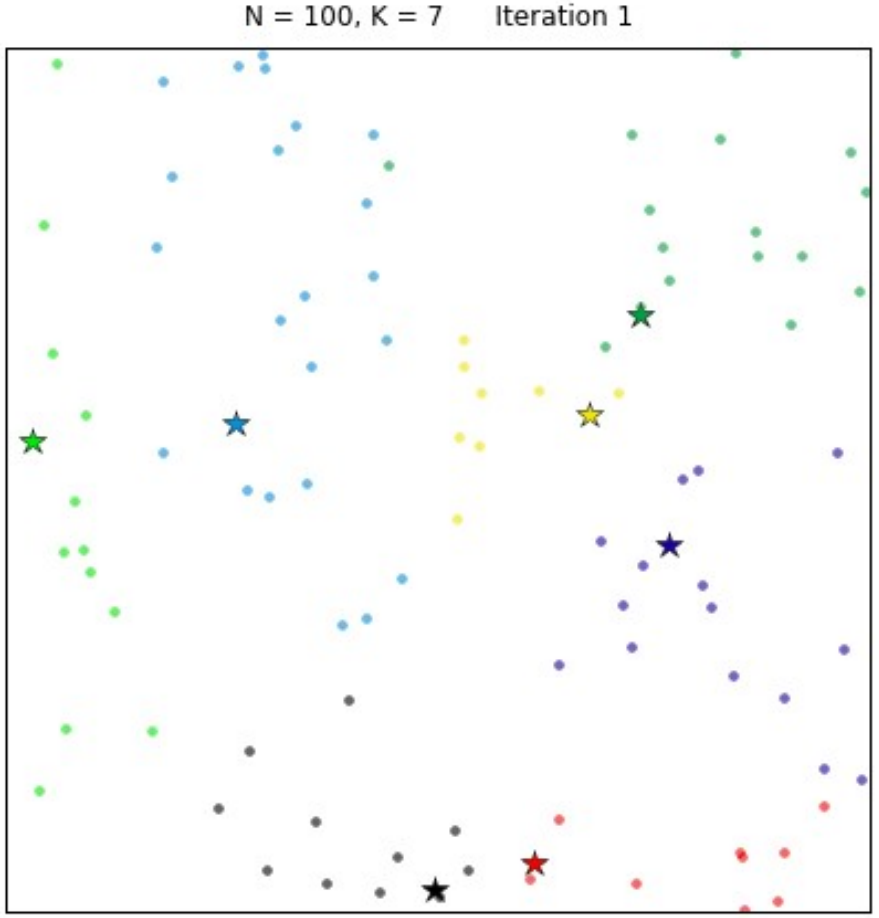
如无拐点，此法失效

根据知识判断（人为判断）

Kmeans动态演示



K=3



K=7

C均值聚类习题

现有样本集如下：

$$x_1=\begin{bmatrix}0 & 0\end{bmatrix}^T, x_2=\begin{bmatrix}1 & 0\end{bmatrix}^T, x_3=\begin{bmatrix}0 & 1\end{bmatrix}^T, x_4=\begin{bmatrix}1 & 1\end{bmatrix}^T$$

$$x_5=\begin{bmatrix}2 & 1\end{bmatrix}^T, x_6=\begin{bmatrix}1 & 2\end{bmatrix}^T, x_7=\begin{bmatrix}2 & 2\end{bmatrix}^T, x_8=\begin{bmatrix}3 & 2\end{bmatrix}^T$$

$$x_9=\begin{bmatrix}6 & 6\end{bmatrix}^T, x_{10}=\begin{bmatrix}7 & 6\end{bmatrix}^T, x_{11}=\begin{bmatrix}8 & 6\end{bmatrix}^T, x_{12}=\begin{bmatrix}6 & 7\end{bmatrix}^T$$

$$x_{13}=\begin{bmatrix}7 & 7\end{bmatrix}^T, x_{14}=\begin{bmatrix}8 & 7\end{bmatrix}^T, x_{15}=\begin{bmatrix}9 & 7\end{bmatrix}^T, x_{16}=\begin{bmatrix}7 & 8\end{bmatrix}^T$$

$$x_{17}=\begin{bmatrix}8 & 8\end{bmatrix}^T, x_{18}=\begin{bmatrix}9 & 8\end{bmatrix}^T, x_{19}=\begin{bmatrix}8 & 9\end{bmatrix}^T, x_{20}=\begin{bmatrix}9 & 9\end{bmatrix}^T$$

试用C均值算法将以上20个样本分成两类，并求出两类聚类中心。设初始聚类中心为

$$z_1(1)=x_1=\begin{bmatrix}0 & 0\end{bmatrix}^T \qquad z_2(1)=x_2=\begin{bmatrix}1 & 0\end{bmatrix}^T$$