

**注：决定将第8章提前到本次课来讲，将第7章放到后面讲**

# 第8章第1讲

## 数据聚类

## Data Clustering

向 世 明

smxiang@nlpr.ia.ac.cn

中科院自动化研究所 模式识别国家重点实验室

**助教：** 何文浩 (wenhao.he@nlpr.ia.ac.cn)  
杨红明 (hongming.yang@nlpr.ia.ac.cn)

# 8.1 引言

- 聚类

- 物以类聚，人以群分。
- 将数据分成多个类别，在同一个类内，对象（实体）之间具有较高的相似性，不同类对象间差异性较大。
- 对一批没有类别标签的样本集，按照样本之间的相似程度分类，相似的归为一类，不相似的归为其它类。这种分类称为**聚类分析**，也称为无监督分类。
- 聚类的质量(或结果)取决于对**度量标准**的选择。
- 聚类结果**因不同任务而不同**。



## 身份识别 vs 姿态估计

# 8.1 引言

- 聚类任务

- 给定一个样本集合  $X$ ，给定一种度量样本间相似度或者相异度（距离）的标准。聚类系统的输出是关于样本集  $X$  的一个划分，即  $D = \{D_1 \cup D_2 \cup \dots \cup D_k\}$ 。其中， $D_i (i=1,2,\dots,k)$  是  $X$  的一个子集，且满足：

- $D_1 \cup D_2 \cup \dots \cup D_k = X$

- $D_i \cap D_j = \emptyset, i \neq j$

- $D$  中成员  $D_1, D_2, \dots, D_k$  叫做类或者簇(cluster)，每个类均通过一些特征来描述：

- 通过类中心或者类的边界点来表示；
  - 使用聚类树采用图形化方式来表示。

# 8.1 引言

- 聚类方法分类

- 按照聚类标准

- **统计聚类方法**：基于全局数据的聚类，即从全体样本中通过距离比较，获得聚类中心。主要采用欧氏距离度量、马氏距离度量等。
    - **概念聚类方法**：将数据按按一定的方式和准则进行分组，得到的分组代表着不同的概念。

- 按聚类所处理的数据类型

- **数值型**数据聚类、**离散型**数据聚类、**混合型**数据聚类。

# 8.1 引言

- 聚类方法分类

- 按照度量准则

- 基于**距离**的聚类方法：基于各种不同的距离或者相似性来度量点对之间的关系，如**K-means**等。
    - 基于**密度**的聚类方法：采用密度函数对样本进行描述，并得到聚类结果。
    - 基于**连通性**的聚类方法：主要包含**基于图**的方法。高度连通的数据通常被聚为一簇，如**谱聚类**。

# 8.1 引言

- 聚类方法分类

- 按照不同的技术路线

- **划分法**：采用一定的规则对数据进行划分，如**K-means**等。
    - **层次法**：对给定样本进行层次划分，如**层级聚类**。
    - **密度法**：对数据的密度进行评价，如**混合高斯模型**、**Mean-Shift方法**。
    - **网格法**：将数据空间划分为有限个单元网络结构，然后基于网络结构进行聚类，如**矢量量化**。
    - **模型法**：为每一个簇引入一个模型，然后对数据进行划分，使其满足各自分派的模型。

# 8.1 引言

- **挑战性问题**

- **可伸缩性**

- 可伸缩性是指聚类算法无论对于小数据集还是大数据集，都应有效；无论对小类别数据还是大别类数据，都应有效。

- **具有不同类型的数据处理能力**

- 既可处理数值型数据，也可处理非数值型数据；既可处理离散数据，也可处理连续域内的数据。比如布尔型、时序型、枚举型、以及这些类型的混合。

- **能够发现任意形状的聚类**

- 能够发现任意形状的簇，球状的、位于同一流形上的数据。因此，选择合适的距离度量很关键。



# 8.11 一些挑战性问题

- 挑战性问题

- 能够处理高维数据

- 既可处理属性较少的数据，也可处理属性较多的数据。
- 在高维空间聚类更具挑战性，随着维数的增加，具有相同距离的两个样本其相似程度可以相差很远。对于高维稀疏数据，这一点更突出。

- 对噪声鲁棒

- 在实际中，绝大多数样本集都包含噪声、空缺、部分未知属性、孤立点、甚至错误数据。

# 8.11 一些挑战性问题

- 挑战性问题

- 具有约束的聚类

- 在实际应用中，通常需要在某种约束条件下进行聚类，既满足约束条件，以希望有高聚类精度，是一个挑战性问题。

- 对初始输入参数鲁棒

- 具有自适应的簇数判定能力（一直没有解决好）。
    - 对初始聚类中心鲁棒。

- 能够解决用户的问题

- 聚类结果能被用户所理解，并能带来经济效益，特别是在数据挖掘领域。

## 8.2 距离与相似性度量

- 距离

- 设有  $d$  维空间的三个样本  $\mathbf{x}$ ,  $\mathbf{y}$  和  $\mathbf{z}$ , 记  $d(\cdot, \cdot)$  为一个  $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  的映射, 如满足如下几个条件则称  $d(\cdot, \cdot)$  为一个距离:

- $d(\mathbf{x}, \mathbf{y}) \geq 0$

非负性

- $d(\mathbf{x}, \mathbf{x}) = 0$

自相似性

- $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$

对称性

- $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$

三角不等式

- 距离可以描述对点间的相异程度, 距离越大, 两个点越不相似; 距离越小, 两个点越相似。

## 8.2 距离与相似性度量

- 设  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , **Minkowski 距离度量** 定义如下:

$$d(\mathbf{x}, \mathbf{y}) = \left( \sum_{i=1}^d |x_i - y_i|^q \right)^{\frac{1}{q}}$$



$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^d |x_i - y_i|$$

城区距离  
曼哈顿距离

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^d |x_i - y_i|^2}$$

欧氏距离

$$d(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq d} |x_i - y_i|$$

切比雪夫距离

## 8.2 距离与相似性度量

- 设  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , **Mahalanobis (马氏)距离**定义如下:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{M} (\mathbf{x} - \mathbf{y})}$$

其中,  $\mathbf{M}$ 是半正定矩阵。

- $\mathbf{M}$ 为单位矩阵时, 退化为欧氏距离度量。
- $\mathbf{M}$ 为对角矩阵时, 退化为**特征加权**欧氏距离

## 8.2 距离与相似性度量

- 相似性

- 设  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , **余弦相似度**定义如下:

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^d x_i y_i}{\sqrt{\sum_{i=1}^d x_i^2} \sqrt{\sum_{i=1}^d y_i^2}}$$

(两个模为1的向量之内积)

## 8.2 距离与相似性度量

- 相似性

- 设  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ ，其每维特征只取  $\{0,1\}$  中的一个值。为了定义数据点之间的距离，通常先计算出如下几个值：

- $f_{00}$ ：样本  $\mathbf{x}$  和  $\mathbf{y}$  中满足  $x_i=y_i=0$  的属性的个数
- $f_{10}$ ：样本  $\mathbf{x}$  和  $\mathbf{y}$  中满足  $x_i=1 \& y_i=0$  的属性的个数
- $f_{01}$ ：样本  $\mathbf{x}$  和  $\mathbf{y}$  中满足  $x_i=0 \& y_i=1$  的属性的个数
- $f_{11}$ ：样本  $\mathbf{x}$  和  $\mathbf{y}$  中满足  $x_i=y_i=1$  的属性的个数

- 进一步，可定义如下几种类型的相似性度量：

## 8.2 距离与相似性度量

- 相似性

- **简单匹配系数**(simple matching coefficient, SMC):

$$s_{SMC}(\mathbf{x}, \mathbf{y}) = \frac{f_{00} + f_{11}}{f_{00} + f_{10} + f_{01} + f_{11}}$$

- **Jaccard 相似系数**:

$$s_J(\mathbf{x}, \mathbf{y}) = \frac{f_{11}}{f_{10} + f_{01} + f_{11}}$$

- **Tanimoto 系数**:

$$s_T(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\mathbf{x}^T \mathbf{x} + \mathbf{y}^T \mathbf{y} - \mathbf{x}^T \mathbf{y}} = \frac{f_{11}}{\mathbf{x}^T \mathbf{x} + \mathbf{y}^T \mathbf{y} - f_{11}}$$

$\mathbf{x}$ 取1的个数

$\mathbf{y}$ 取1的个数



## 8.2 距离与相似性度量

- 举例：计算如下**两位顾客x和y**的相似度：

商品	面包	啤酒	牛奶	咖啡	茶叶	鸡蛋	猪肉	牛肉	洋葱	土豆	大米	白糖
x	1	1	1	1	0	0	0	1	1	1	0	1
y	1	0	1	0	1	1	0	0	0	1	1	1

商品	莲藕	花生	可乐	豆腐	菠菜	黄瓜	面粉	酱油	辣椒	白酒	黄鱼	茄子
x	0	0	1	0	1	1	1	0	1	1	1	0
y	1	0	1	1	1	1	1	1	0	0	1	0

## 8.2 距离与相似性度量

- 类间距离:

- **最短距离法**: 定义两个类中最近的两个样本的距离为类间距离。

$$d(D_a, D_b) = \min\{d(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in D_a, \mathbf{y} \in D_b\}$$

- **最长距离法**: 定义两个类中最远的两个样本的距离为类间距离。

$$d(D_a, D_b) = \max\{d(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in D_a, \mathbf{y} \in D_b\}$$

- **类直径**: 类直径反映类中样本之间的差异, 可定义为类中各样本至**类中心点**的欧氏距离平方和:

$$r(D_a) = \sum_{\mathbf{x} \in D_a} (\mathbf{x} - \bar{\mathbf{x}})^T (\mathbf{x} - \bar{\mathbf{x}})$$

## 8.3 混合密度函数

- 目标—利用样本估计密度中的一些参数
  - 混合密度估计可为数据聚类提供方法论上的指导。
- 假定：
  - 样本来自于  $c$  个不同类别， $c$  是已知的。
  - 每类出现的先验概率  $P(\omega_j)$  是已知的， $j = 1, 2, \dots, c$ 。
  - 类条件概率密度函数  $p(\mathbf{x}|\omega_j, \theta_j)$  的形式是已知的。
  - $c$  个参数向量  $\theta_j$ ， $j = 1, 2, \dots, c$ ，是未知的。
  - 样本的类别标签也是未知的。
- 样本的生成过程：首先通过类先验概率  $P(\omega_j)$  随机选择一个类别，然后通过类条件概率密度函数  $p(\mathbf{x}|\omega_j, \theta_j)$  随机选择一个样本。

## 8.3 混合密度函数

- 设总体样本的概率密度函数为：

$$p(\mathbf{x} | \boldsymbol{\theta}) = \sum_{j=1}^c p(\mathbf{x} | \omega_j, \boldsymbol{\theta}_j) \boxed{P(\omega_j)}$$

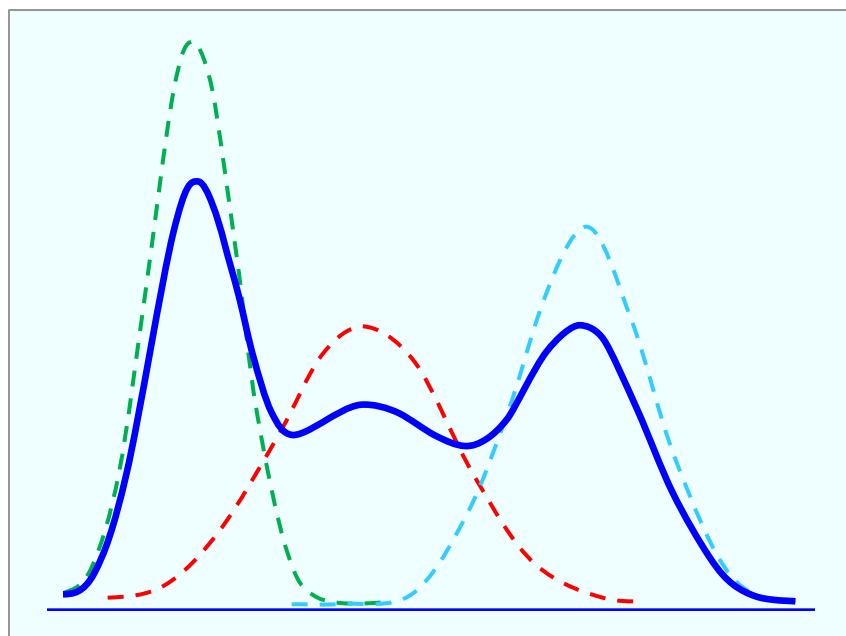
混合比例

其中， $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_c\}$ 。称上述密度函数为**混合密度**；称条件概率密度函数  $p(\mathbf{x} | \omega_j, \boldsymbol{\theta}_j)$  为**成分密度**；称先验概率为**混合参数**。此处主要考察参数  $\boldsymbol{\theta}$ 。

**基本任务：**估计  $\boldsymbol{\theta}$ 。一旦  $\boldsymbol{\theta}$  得到估计，可以将上述混合密度分解为多个已知的密度成分，并且可以采用最大化后验概率来确定样本的类别。

## 8.3 混合密度函数及参数可辨识性

- 举例：一维高斯混合模型：



三个高斯分布的混合

## 8.4 最大似然估计

- 任务：

- 给定一个包含  $n$  个无类别标签的数据集  $D=\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ ，假定这些样本**独立地**从如下混合型概率密度函数中采样得到：

$$p(\mathbf{x} | \boldsymbol{\theta}) = \sum_{j=1}^c p(\mathbf{x} | \omega_j, \boldsymbol{\theta}_j) P(\omega_j)$$

- 根据这些样本，采用最大似然估计方法对  $\boldsymbol{\theta}$  进行估计。
- $D$  中数据的联合密度 (假定独立采样)：

$$p(D | \boldsymbol{\theta}) = \prod_{k=1}^n p(\mathbf{x}_k | \boldsymbol{\theta})$$

## 8.4 最大似然估计

- 目标：估计一个  $\hat{\theta}$  使  $p(D | \theta)$  最大。
  - 考虑对数似然 (log-likelihood):

$$\begin{aligned} f_{lh}(\theta) &= \ln(p(D | \theta)) \\ &= \sum_{k=1}^n \ln(p(\mathbf{x}_k | \theta)) \\ &= \sum_{k=1}^n \ln \left( \sum_{j=1}^c p(\mathbf{x}_k | \omega_j, \theta_j) P(\omega_j) \right) \end{aligned}$$

$f_{lh}(\theta)$ 对参数 $\theta$ 的梯度(假定参数独立):

$$p(\mathbf{x}_k | \theta) = \sum_{j=1}^c p(\mathbf{x}_k | \omega_j, \theta_j) P(\omega_j)$$

仅考虑包含  $\theta_i$  的项

$$\nabla_{\theta_i} f_{lh}(\theta) = \sum_{k=1}^n \frac{1}{p(\mathbf{x}_k | \theta)} \nabla_{\theta_i} p(\mathbf{x}_k | \theta)$$

$$= \sum_{k=1}^n \frac{1}{p(\mathbf{x}_k | \theta)} \nabla_{\theta_i} (p(\mathbf{x}_k | \omega_i, \theta_i) P(\omega_i))$$

$$= \sum_{k=1}^n \frac{P(\omega_i)}{p(\mathbf{x}_k | \theta)} \nabla_{\theta_i} (p(\mathbf{x}_k | \omega_i, \theta_i))$$

相等

$$= \sum_{k=1}^n \frac{P(\omega_i)}{p(\mathbf{x}_k | \theta)} p(\mathbf{x}_k | \omega_i, \theta_i) \nabla_{\theta_i} \ln(p(\mathbf{x}_k | \omega_i, \theta_i))$$

$$= \sum_{k=1}^n \frac{P(\omega_i, \mathbf{x}_k | \theta_i)}{p(\mathbf{x}_k | \theta)} \nabla_{\theta_i} \ln(p(\mathbf{x}_k | \omega_i, \theta_i))$$

$$= \sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \theta) \nabla_{\theta_i} \ln(p(\mathbf{x}_k | \omega_i, \theta_i))$$



$f_{lh}(\theta)$ 对参数的梯度:

$$\nabla_{\theta_i} f_{lh}(\theta) = \sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \theta) \nabla_{\theta_i} \ln(p(\mathbf{x}_k | \omega_i, \theta_i))$$

单个样本对梯度的贡献:

第  $i$  个成分密度

$$\nabla_{\theta_i} f_{lh}(\theta | \mathbf{x}_k) = P(\omega_i | \mathbf{x}_k, \theta) \nabla_{\theta_i} \ln(p(\mathbf{x}_k | \omega_i, \theta_i))$$

$\mathbf{x}_k$  属于第  $i$  个成分的后验概率

单个样本  $\mathbf{x}_k$  对“似然函数关于  $\theta_i$  的梯度”之贡献 等于  
“ $\mathbf{x}_k$  属于第  $i$  个成分的后验概率” 乘以 “ $\mathbf{x}_k$  对第  $i$  个成分密度  $p(\mathbf{x} | \omega_i, \theta_i)$  的对数关于  $\theta_i$  的梯度”。

## 8.4 最大似然估计

- 令梯度等于零，可得如下  $c$  个方程：

$$\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\theta}_i} \ln \left( p(\mathbf{x}_k | \omega_i, \hat{\boldsymbol{\theta}}_i) \right) = 0, \quad i = 1, 2, \dots, c$$

- 求解上述方程可得待估计的  $\hat{\boldsymbol{\theta}}$ 。
- 进一步：当未知量中包含先验概率  $P(\omega_i)$ （即混合比例）时，应限制如下两个条件：

$$P(\omega_i) \geq 0, \quad i = 1, 2, \dots, c, \quad \text{且} \quad \sum_{i=1}^c P(\omega_i) = 1.$$

## 8.4 最大似然估计

- 实际上，如果似然函数可微，且  $P(\omega_i) \neq 0$ ，那么  $P(\omega_i)$  和  $\hat{\theta}_i$  必然同时满足以下条件：

条件1: 
$$\hat{P}(\omega_i) = \frac{1}{n} \sum_{k=1}^n \hat{P}(\omega_i | \mathbf{x}_k, \hat{\theta})$$

条件2: 
$$\sum_{k=1}^n \hat{P}(\omega_i | \mathbf{x}_k, \hat{\theta}) \nabla_{\theta_i} \ln \left( p(\mathbf{x}_k | \omega_i, \hat{\theta}_i) \right) = 0, \quad i = 1, 2, \dots, c$$

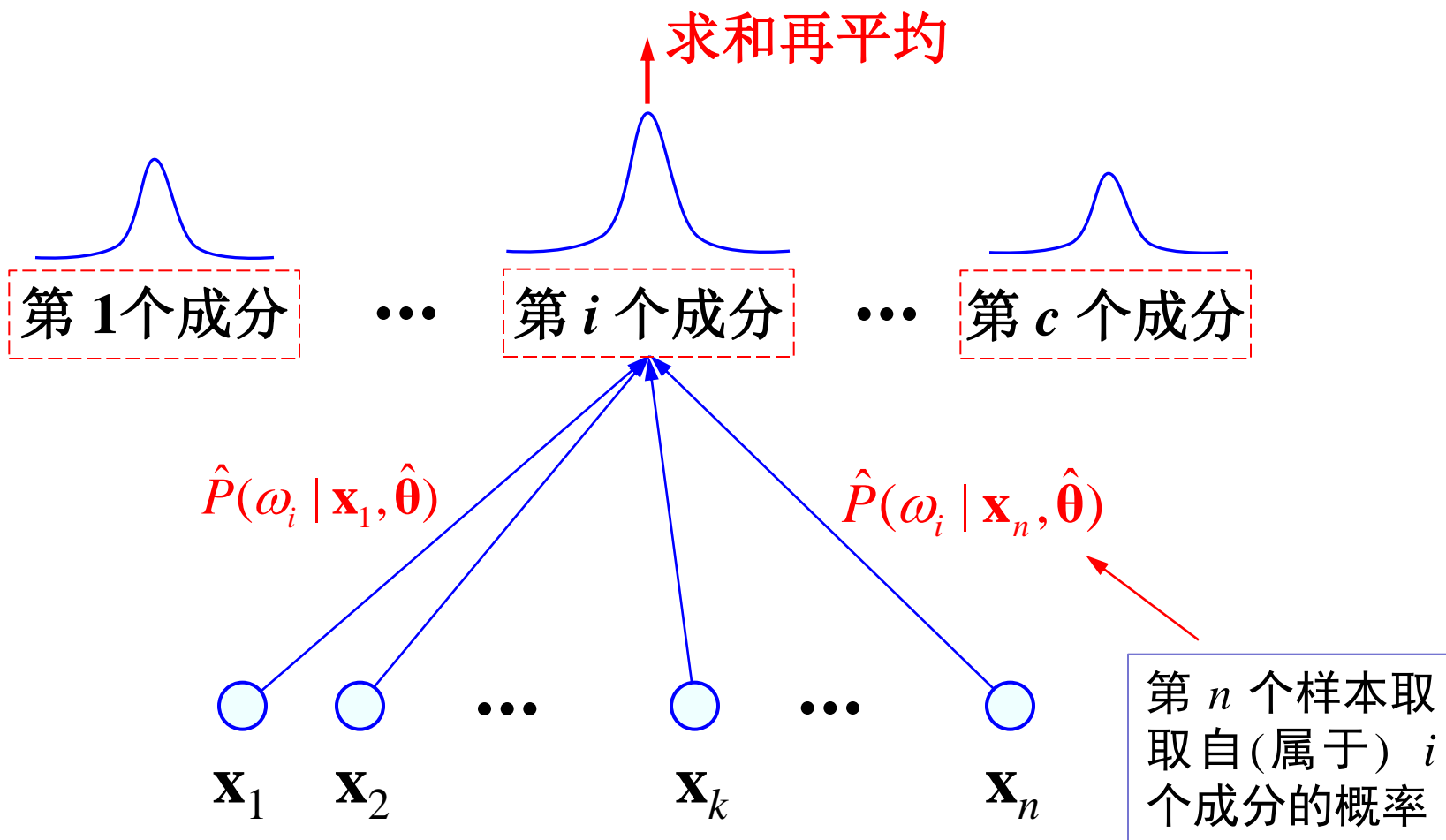
其中, 
$$\hat{P}(\omega_i | \mathbf{x}_k, \hat{\theta}) = \frac{p(\mathbf{x}_k | \omega_i, \hat{\theta}_i) \hat{P}(\omega_i)}{\sum_{j=1}^c p(\mathbf{x}_k | \omega_j, \hat{\theta}_j) \hat{P}(\omega_j)} = \frac{p(\mathbf{x}_k, \omega_i | \hat{\theta}_i)}{p(\mathbf{x}_k | \theta)}$$

全概率公式

# 对类先验的估计

(条件1的直观解释)

$$\hat{P}(\omega_i) = \frac{1}{n} \sum_{k=1}^n \hat{P}(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\theta}}), \quad i = 1, \dots, c$$



- 关于条件1的证明

- 首先，考虑所有变量时对数似然函数可以写成：

$$\begin{aligned} f_{lh}(\boldsymbol{\theta}, \boldsymbol{\alpha}) &= \sum_{k=1}^n \ln(p(\mathbf{x}_k | \boldsymbol{\theta})) = \sum_{k=1}^n \ln \left( \sum_{j=1}^c p(\mathbf{x}_k | \omega_j, \boldsymbol{\theta}_j) P(\omega_j) \right) \\ &= \sum_{k=1}^n \ln \left( \sum_{j=1}^c p(\mathbf{x}_k | \omega_j, \boldsymbol{\theta}_j) \alpha_j \right) \end{aligned}$$

其中引入新记号： $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_c]^T = [P(\omega_1), P(\omega_2), \dots, P(\omega_c)]^T$

- 然后，拉格朗日函数可以写成：

$$L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = f_{lh}(\boldsymbol{\theta}, \boldsymbol{\alpha}) + \underset{\substack{\uparrow \\ \text{拉格朗日乘子}}}{\lambda} \left( \sum_{j=1}^c \alpha_j - 1 \right) \quad \because \sum_{j=1}^c \alpha_j = 1$$

拉格朗日乘子

- 关于条件1的证明(续)

- 求目标函数关于变量的偏导数，并令其等于0：

$$\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{\alpha})}{\partial \alpha_i} = \sum_{k=1}^n \frac{p(\mathbf{x}_k | \omega_i, \boldsymbol{\theta}_i)}{p(\mathbf{x}_k | \boldsymbol{\theta})} + \lambda = 0, \quad i = 1, 2, \dots, c$$

- 在方程的两边乘以  $\alpha_i$ ，并将  $c$  个方程相加，可得

$$\begin{aligned} \sum_{i=1}^c \sum_{k=1}^n \frac{p(\mathbf{x}_k | \omega_i, \boldsymbol{\theta}_i) \alpha_i}{p(\mathbf{x}_k | \boldsymbol{\theta})} + \lambda \sum_{i=1}^c \alpha_i &= 0 \\ \Rightarrow \lambda &= - \sum_{i=1}^c \sum_{k=1}^n \frac{p(\mathbf{x}_k | \omega_i, \boldsymbol{\theta}_i) \alpha_i}{p(\mathbf{x}_k | \boldsymbol{\theta})} = - \sum_{k=1}^n \sum_{i=1}^c \frac{p(\mathbf{x}_k | \omega_i, \boldsymbol{\theta}_i) \alpha_i}{p(\mathbf{x}_k | \boldsymbol{\theta})} \\ &= - \sum_{k=1}^n \frac{p(\mathbf{x}_k | \boldsymbol{\theta})}{p(\mathbf{x}_k | \boldsymbol{\theta})} = -n \end{aligned}$$

- 关于条件1的证明(续)

- 最后由如下公式

$$\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{\alpha})}{\partial \alpha_i} = \sum_{k=1}^n \frac{p(\mathbf{x}_k | \omega_i, \boldsymbol{\theta}_i)}{p(\mathbf{x}_k | \boldsymbol{\theta})} + \lambda = 0, \quad i = 1, 2, \dots, c$$

- 可得

$$\sum_{k=1}^n \frac{p(\mathbf{x}_k | \omega_i, \boldsymbol{\theta}_i) \alpha_i}{p(\mathbf{x}_k | \boldsymbol{\theta})} = n \alpha_i, \quad i = 1, 2, \dots, c$$

$$\begin{aligned} \Rightarrow \alpha_i &= \frac{1}{n} \sum_{k=1}^n \frac{p(\mathbf{x}_k | \omega_i, \boldsymbol{\theta}_i) \alpha_i}{p(\mathbf{x}_k | \boldsymbol{\theta})} = \frac{1}{n} \sum_{k=1}^n \frac{p(\mathbf{x}_k | \omega_i, \boldsymbol{\theta}_i) p(\omega_i)}{p(\mathbf{x}_k | \boldsymbol{\theta})} \\ &= \frac{1}{n} \sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \boldsymbol{\theta}) \end{aligned}$$

因此，条件1得证。

## 8.5 正态分布情形下的非监督参数估计

- 本节讨论混合密度的各分量成分均为**多维正态分布**的情形：

$$p(\mathbf{x}|\omega_i, \theta_i) \sim N(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$N(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right)$$

- 考虑如下两种情形：

Case	$\boldsymbol{\mu}_i$	$\boldsymbol{\Sigma}_i$	$P(\omega_i)$	$c$
1	?	$\checkmark$	$\checkmark$	$\checkmark$
2	?	?	?	$\checkmark$



## 8.5 正态分布情形下的非监督参数估计

- 情形一：均值  $\mu_i$  未知

- 似然函数如下：

$$\ln p(\mathbf{x} | \omega_i, \mu_i) = -\ln \left( (2\pi)^{d/2} |\Sigma|^{1/2} \right) - \frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i)$$

- 梯度：

$$\nabla_{\mu_i} \ln p(\mathbf{x} | \omega_i, \mu_i) = \Sigma_i^{-1} (\mathbf{x} - \mu_i)$$

- 均值  $\mu_i$  需要满足的方程 (由前述条件2)：

$$\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\mu}) \Sigma_i^{-1} (\mathbf{x}_k - \hat{\mu}_i) = 0, \quad i = 1, 2, \dots, c$$

$$\hat{\mu} = [\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_c]^T$$

- 情形一：均值  $\mu_i$  未知

- 通过两边乘以  $\Sigma_i$ ，于是有：

$$\hat{\mu}_i = \frac{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\mu}) \mathbf{x}_k}{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\mu})}$$

- 进一步，令：

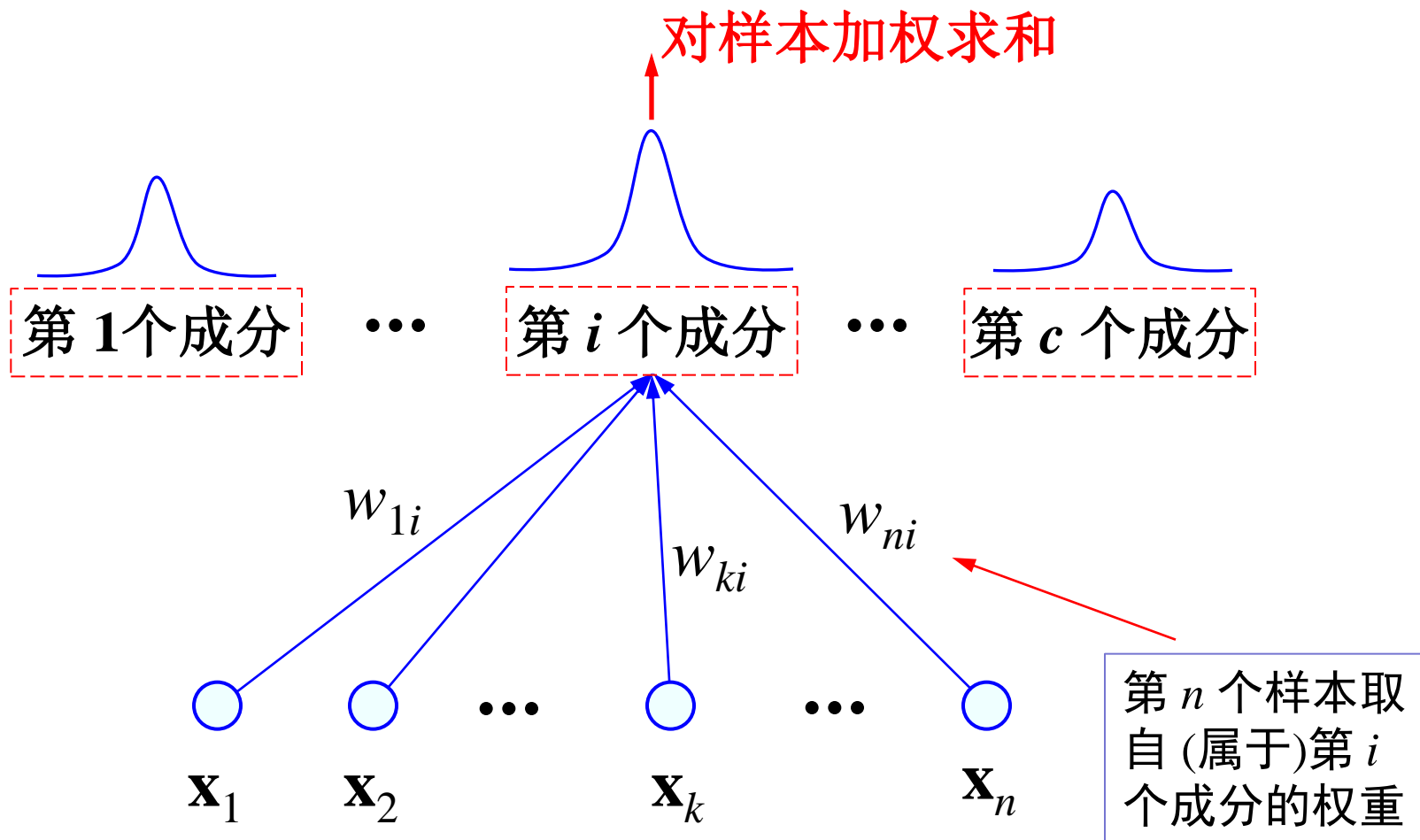
$$w_{ki} = \frac{P(\omega_i | \mathbf{x}_k, \hat{\mu})}{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\mu})}, \quad k = 1, \dots, n; \quad i = 1, \dots, c$$

$$\Rightarrow \hat{\mu}_i = \sum_{k=1}^n w_{ki} \mathbf{x}_k$$

上式表明，类均值的最大似然估计为样本的加权平均。权值表明样本  $\mathbf{x}_k$  属于第  $i$  类的可能性。

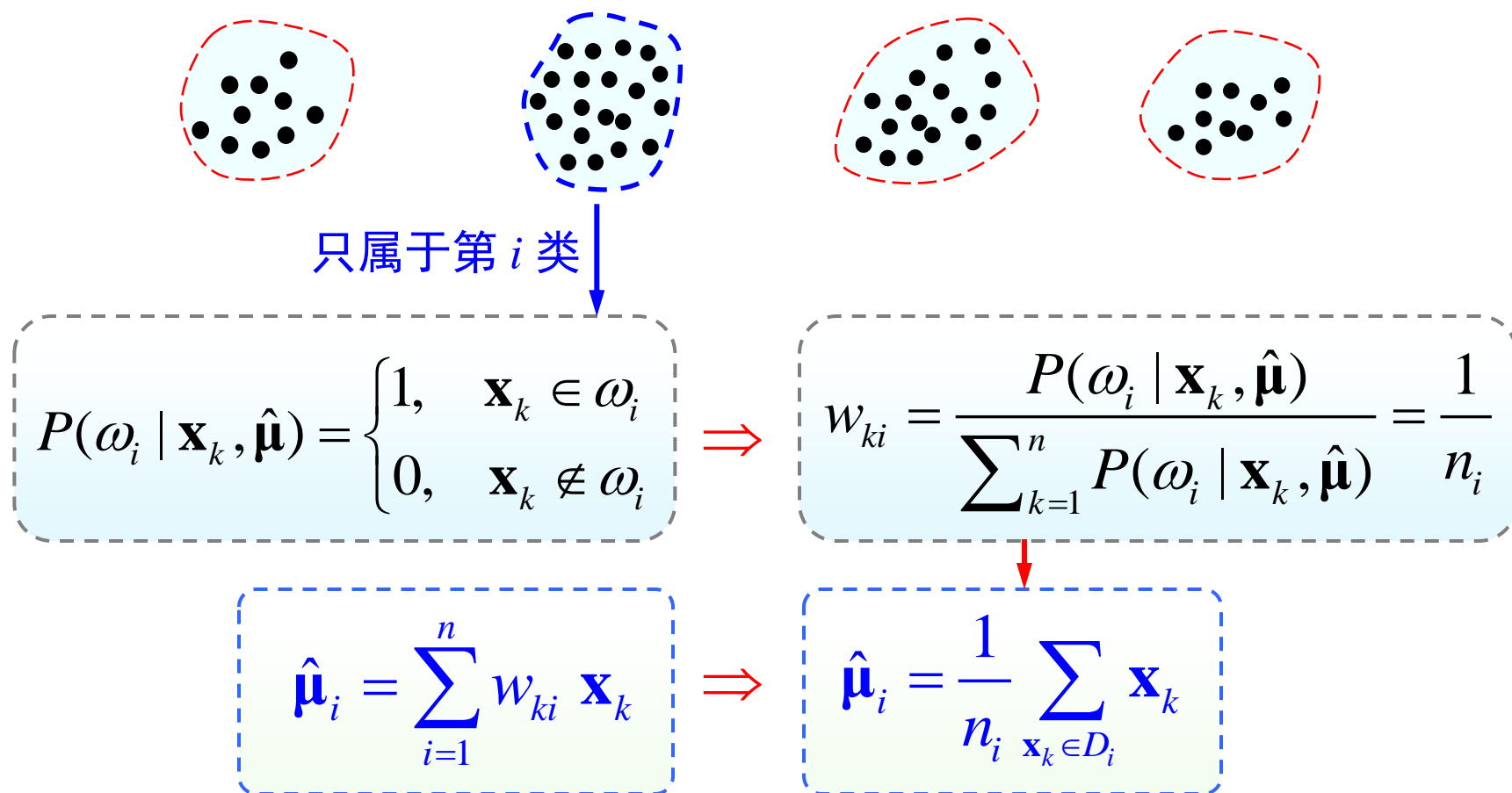
对  $\mu_i$  的估计(解释):

$$\hat{\mu}_i = \sum_{k=1}^n w_{ki} \mathbf{x}_k, \quad w_{ki} = \frac{P(\omega_i | \mathbf{x}_k, \hat{\mu})}{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\mu})}$$



- 情形一：均值  $\mu_i$  未知

- 如果样本满足：  $p(\omega_i, | \mu_i, \mathbf{x}_k)=1$ ，其它均为 0，  $\hat{\mu}_i$  将等于所有属于第  $i$  类的样本的均值。



## 8.5 正态分布情形下的非监督参数估计

- 情形一：均值  $\mu_i$  未知

- 如果  $\{\hat{\mu}_i\}$  充分接近其真值，则  $P(\omega_i | \mathbf{x}_k, \hat{\mu})$  将成为  $\mathbf{x}_k$  属于第  $i$  类的后验概率。
- 但  $\hat{\mu}_i$  的计算要通过类条件概率和类先验概率来计算：

$$\begin{aligned} P(\omega_i | \mathbf{x}_k, \hat{\mu}) &= \frac{p(\mathbf{x}_k, \omega_i | \hat{\mu})}{p(\mathbf{x}_k | \hat{\mu})} = \frac{p(\mathbf{x}_k, \omega_i | \hat{\mu}_i)}{p(\mathbf{x}_k | \hat{\mu})} \\ &= \frac{p(\mathbf{x}_k | \omega_i, \hat{\mu}_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x}_k | \omega_j, \hat{\mu}_j)P(\omega_j)} = \frac{N(\mathbf{x}_k | \hat{\mu}_i, \Sigma_i)P(\omega_i)}{\sum_{j=1}^c N(\mathbf{x}_k | \hat{\mu}_j, \Sigma_j)P(\omega_j)} \end{aligned}$$

第二个等式：由于“先选择类  $\omega_i$  再选择样本”这一同时发生的事件只与第  $i$  个成分相关。

- 情形一：均值  $\mu_i$  未知

- 但是，上述表示并不是关于  $\hat{\mu}_i$  的一个显示表达式，它与  $\hat{\mu}$  有关，因为后验概率包含待估参数(根据前一页，我们有)：

$$P(\omega_i | \mathbf{x}_k, \hat{\mu}) = \frac{N(\mathbf{x}_k | \hat{\mu}_i, \Sigma_i) P(\omega_i)}{\sum_{j=1}^c N(\mathbf{x}_k | \hat{\mu}_j, \Sigma_j) P(\omega_j)}$$

- 通常采用迭代求解 (给定各值初值)：

$$\hat{\mu}_i(t+1) = \frac{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\mu}(t)) \mathbf{x}_k}{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\mu}(t))}$$

算法本质：梯度下降法，也称爬山法 （最大似然）

- 一个例子：假定以下25个样本随机取自于如下分布：

$$p(x | \mu_1, \mu_2) = \frac{1}{3} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \mu_1)^2\right) + \frac{2}{3} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \mu_2)^2\right)$$

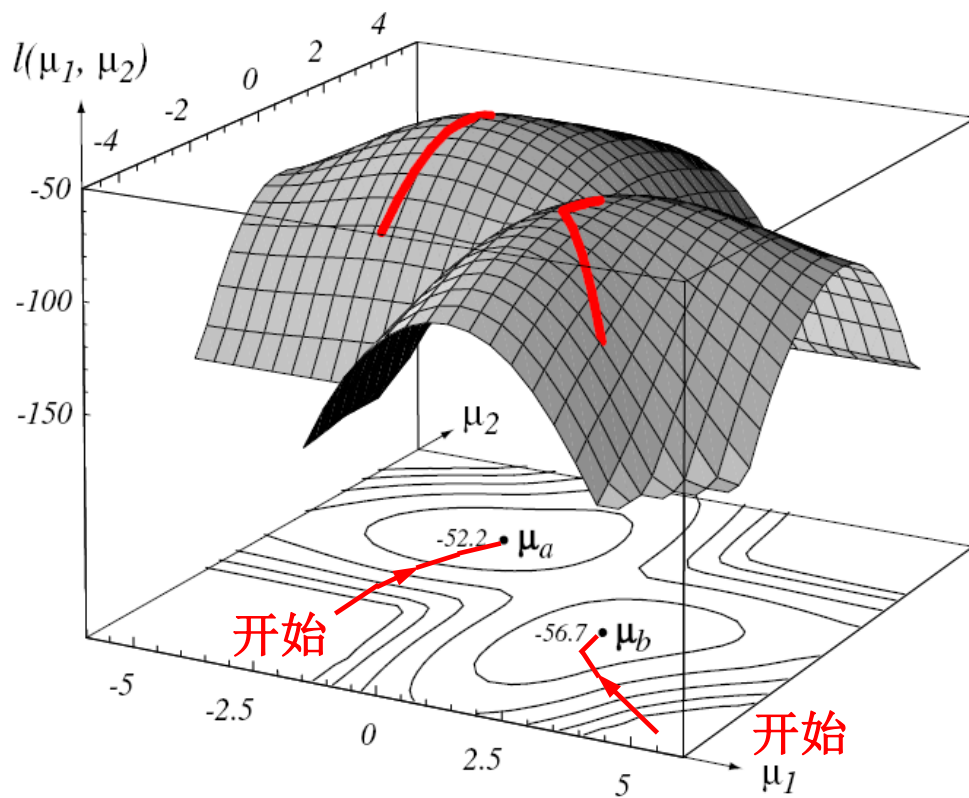
其中， $\mu_1 = -2$ ,  $\mu_2 = 2$ .

$k$	$x_k$	$\omega_1$	$\omega_2$
1	0.608		×
2	-1.590	×	
3	0.235		×
4	3.949		×
5	-2.249	×	
6	2.704		×
7	-2.473	×	
8	0.672		×

$k$	$x_k$	$\omega_1$	$\omega_2$
9	0.262		×
10	1.072		×
11	-1.773	×	
12	0.537		×
13	3.240		×
14	2.400		×
15	-2.499	×	
16	2.608		×

$k$	$x_k$	$\omega_1$	$\omega_2$
17	-3.458	×	
18	0.257		×
19	2.569		×
20	1.415		×
21	1.410		×
22	-2.653	×	
23	1.396		×
24	3.286		×
25	-0.712	×	

- 由25个样本生成的似然函数曲面 (25个对数似然相加得到):



目标函数有两个局部最大点在  $(\mu_1, \mu_2) = (-2, 2)$  和  $(2, -2)$  附近。每个最大点都是一个近似正确的解。因为类中心交换一下顺序也是可以的。

图中，两个不同的初始迭代点分别趋近于不同的局部最优点。



- 情形二：所有参数均未知 (但总类数已知)

- 对  $\mu_i, \Sigma_i$ ，样本  $\mathbf{x}_k$  的似然值有：

$$\begin{aligned}\ln p(\mathbf{x}_k | \omega_i, \mu_i, \Sigma_i) &= -\ln \left( (2\pi)^{d/2} |\Sigma_i|^{1/2} \right) - \frac{1}{2} (\mathbf{x}_k - \mu_i)^T \Sigma_i^{-1} (\mathbf{x}_k - \mu_i) \\ &= \ln \left( |\Sigma_i^{-1}|^{1/2} / (2\pi)^{d/2} \right) - \frac{1}{2} (\mathbf{x}_k - \mu_i)^T \Sigma_i^{-1} (\mathbf{x}_k - \mu_i)\end{aligned}$$

因为：

$$N(\mathbf{x} | \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp \left( -\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right)$$

- 情形二：所有参数均未知 (但总类数已知)
  - 对  $\ln p(\mathbf{x}_k | \omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$  其求梯度，并考虑所有样本，通过矩阵代数运算，我们有：

---

类先验 (混合比例) :

$$\hat{P}(\omega_i) = \frac{1}{n} \sum_{k=1}^n \hat{P}(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\theta}})$$

---

类均值:

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\mu}}) \mathbf{x}_k}{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\mu}})}$$

---

类协方差矩阵:

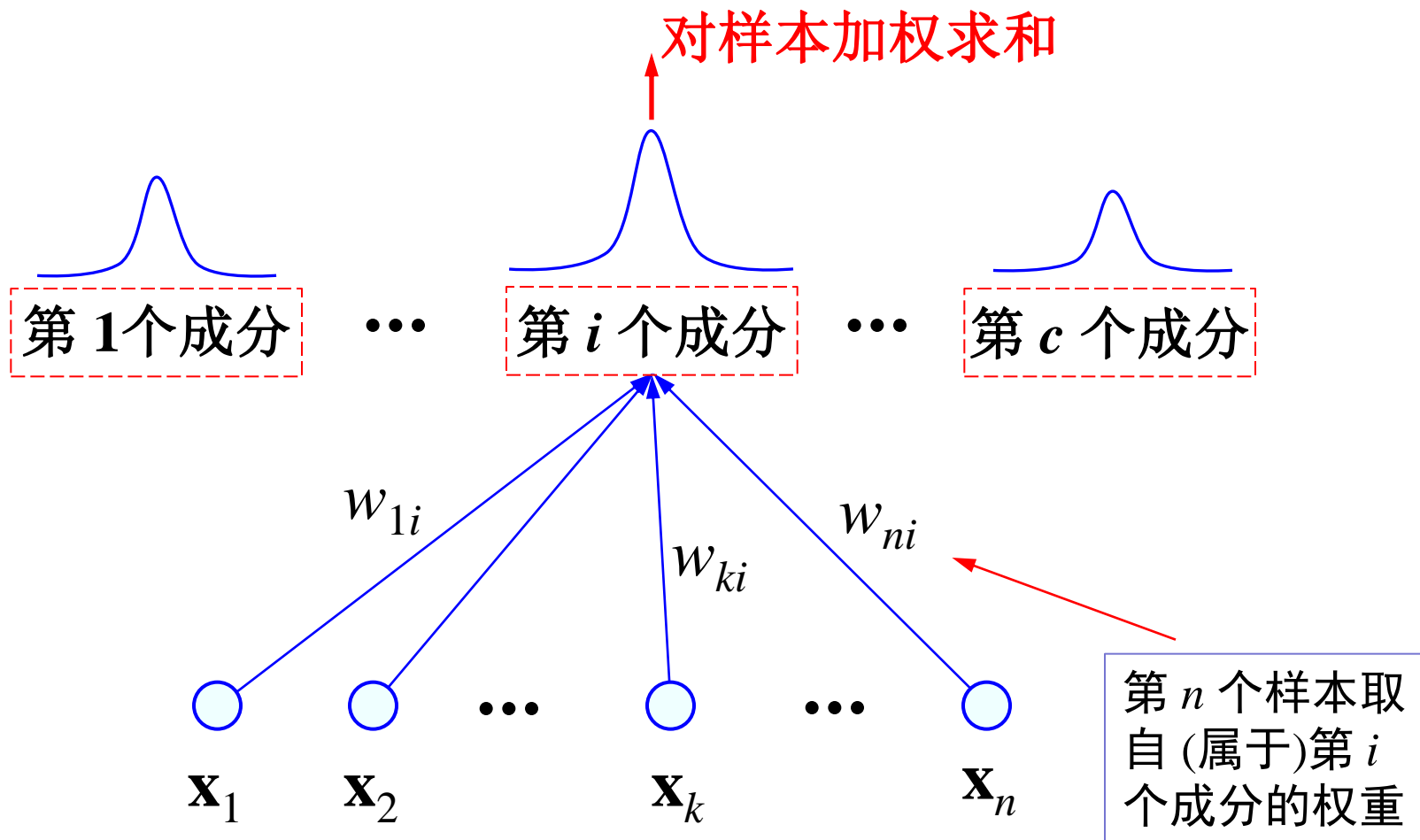
$$\hat{\boldsymbol{\Sigma}}_i = \frac{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\theta}}) (\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i) (\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)^T}{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\theta}})}$$

---

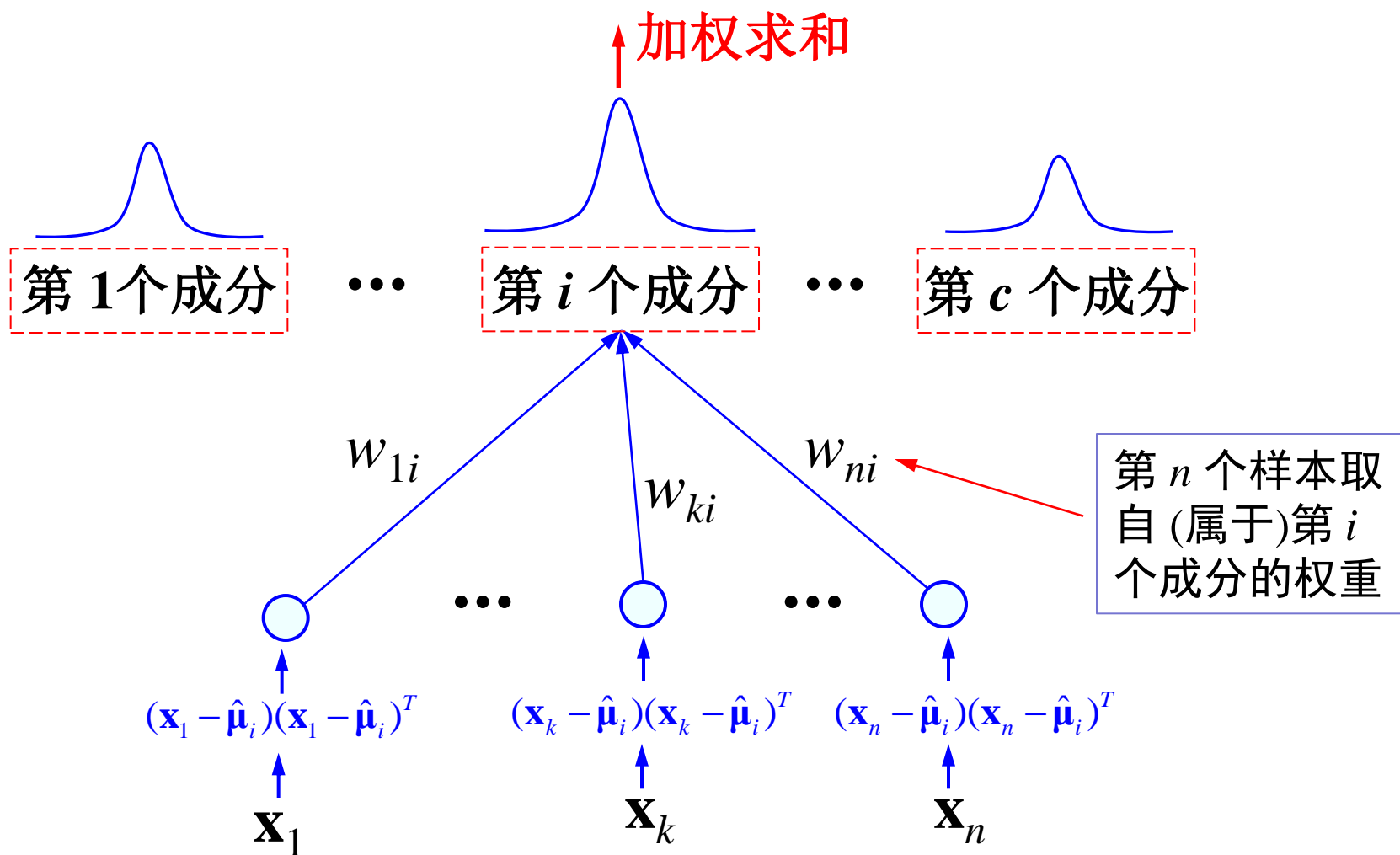
(  $\hat{\boldsymbol{\theta}}$  记录所有的未知参数 )

对  $\mu_i$  的估计(解释):

$$\hat{\mu}_i = \sum_{k=1}^n w_{ki} \mathbf{x}_k, \quad w_{ki} = \frac{P(\omega_i | \mathbf{x}_k, \hat{\mu})}{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\mu})}$$



对  $\Sigma_i$  的估计(解释): 
$$\hat{\Sigma}_i = \sum_{k=1}^n w_{ki} (\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)^T, \quad w_{ki} = \frac{P(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\mu}})}{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\mu}})}$$



- 其中，样本属于第  $i$  个成分的后验概率(此时可计算):

$$\begin{aligned}\hat{P}(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\theta}}) &= \frac{p(\mathbf{x}_k, \omega_i | \hat{\boldsymbol{\theta}}_i)}{p(\mathbf{x}_k | \hat{\boldsymbol{\theta}})} \\&= \frac{p(\mathbf{x}_k | \omega_i, \hat{\boldsymbol{\theta}}_i) \hat{P}(\omega_i)}{\sum_{j=1}^c p(\mathbf{x}_k | \omega_j, \hat{\boldsymbol{\theta}}_j) \hat{P}(\omega_j)} \\&= \frac{|\hat{\boldsymbol{\Sigma}}_i|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)\right) \hat{P}(\omega_i)}{\sum_{j=1}^c |\hat{\boldsymbol{\Sigma}}_j|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_j)\right) \hat{P}(\omega_j)}\end{aligned}$$

注意，分子只与  $i$  有关。

(  $\hat{\boldsymbol{\theta}}$  记录所有的未知参数 )

- 前面关于均值、类方差和混合比例的公式看起来很复杂。但实际上，**它们的含义确十分明显**。
  - 在极端情况下，即当样本  $\mathbf{x}_k$  来自于  $\omega_i$  类时，其后验概率  $\hat{P}(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\theta}})$  为 1，否则就为零，此时有：

只属于第  $i$  类：

$$P(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\mu}}) = \begin{cases} 1, & \mathbf{x}_k \in \omega_i \\ 0, & \mathbf{x}_k \notin \omega_i \end{cases} \rightarrow$$

$$\hat{P}(\omega_i) = \frac{n_i}{n},$$

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbf{x}_k^{(i)},$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} (\mathbf{x}_k^{(i)} - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}_k^{(i)} - \hat{\boldsymbol{\mu}}_i)^T$$

上标  $(i)$  表示属于  $\omega_i$  类的样本， $n_i$  表示属于  $\omega_i$  类样本的个数。

## 8.6 K-均值聚类 (K-means clustering)

- 在前一节中，有关参数估计的相关结论可从多方面简化，得到一些经典的算法。其中之一是著名的 **K-均值聚类** 算法。引入如下假设：
  - 各类出现的先验概率均相等；
  - 每个均本点以概率为1属于一个类（后验概率0-1近似）；
- 计算数据点到类中心的欧氏距离的平方，即计算  $\|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\|^2$ ，寻找与样本  $\mathbf{x}_k$  最近的类中心点，将  $\mathbf{x}_k$  分给最近的类（即假定协方差矩阵为单位阵）：

$$\hat{P}(\omega_i | \mathbf{x}_k, \hat{\boldsymbol{\theta}}) \approx \begin{cases} 1, & \text{if } \mathbf{x}_k \text{ is nearest to the center } \hat{\boldsymbol{\mu}}_i \\ 0, & \text{otherwise} \end{cases}$$

## 8.6 K-均值聚类

- 基于上述假定，对  $c$  个高斯成分的均值，我们有：

$$\hat{\mu}_i = \frac{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\mu}) \mathbf{x}_k}{\sum_{k=1}^n P(\omega_i | \mathbf{x}_k, \hat{\mu})} = \frac{1}{n_i} \sum_{\mathbf{x}_k \in \omega_i} \mathbf{x}_k, \quad i=1, 2, \dots, c$$

- 但是，样本  $\mathbf{x}_k$  属于哪一类需要通过计算  $\|\mathbf{x}_k - \hat{\mu}_i\|^2$  来判断，因此需要迭代进行。
- 通过迭代最终得到  $c$  个高斯成分的均值之后，以这些均值作为  $c$  个类（簇）的类中心，计算每个样本点到类中心的欧氏距离，将样本点归入到距离最近的类。从而完成 K-均值聚类的计算工作。



# 8.6 K-均值聚类

- 算法基本思想

---

## K-Means Clustering—Algorithm 1

---

- 1 begin initialization  $n, c, \mu_1, \mu_2, \dots, \mu_c$ .
  - 2 **do** classify  $n$  samples according to nearest  $\mu_i$
  - 3     re-compute  $\mu_i$
  - 4 **until** no change in  $\mu_i$
  - 5 return  $\mu_1, \mu_2, \dots, \mu_c$
-

- 一个例子：假定以下25个样本随机取自于如下分布：

$$p(x | \mu_1, \mu_2) = \frac{1}{3\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \mu_1)^2\right) + \frac{2}{3\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \mu_2)^2\right)$$

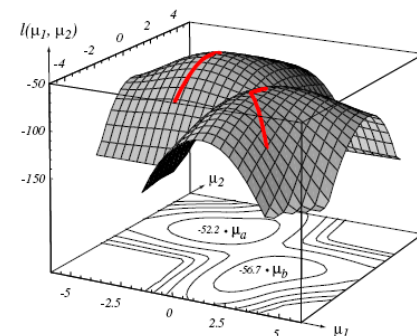
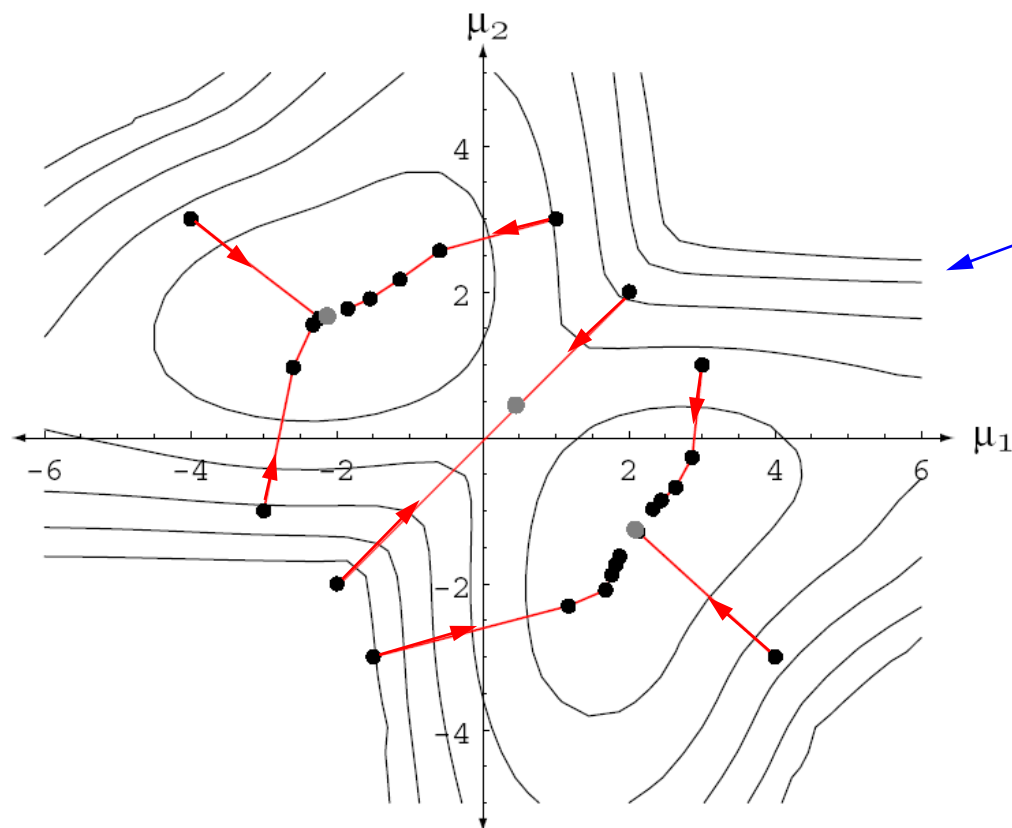
其中， $\mu_1 = -2, \mu_2 = 2$ .

$k$	$x_k$	$\omega_1$	$\omega_2$
1	0.608		×
2	-1.590	×	
3	0.235		×
4	3.949		×
5	-2.249	×	
6	2.704		×
7	-2.473	×	
8	0.672		×

$k$	$x_k$	$\omega_1$	$\omega_2$
9	0.262		×
10	1.072		×
11	-1.773	×	
12	0.537		×
13	3.240		×
14	2.400		×
15	-2.499	×	
16	2.608		×

$k$	$x_k$	$\omega_1$	$\omega_2$
17	-3.458	×	
18	0.257		×
19	2.569		×
20	1.415		×
21	1.410		×
22	-2.653	×	
23	1.396		×
24	3.286		×
25	-0.712	×	

- K-均值迭代过程:



对数似然  
函数等值线

8个初始点（二维向量）：3个迭代获得(-2,2)附近的点，3个迭代获得(2,-2)附近的点，两个得到(0,0)附近的点（错误）。

## 8.6 K-均值聚类

- 前面我们对K-均值算法从混合高斯密度函数估计的角度做了一个解释。
- 在估计混合密度均值时，我们考虑样本点至类中心的欧氏距离，以此为迭代准则来逐步地得到个类中心（即均值），从而完成K-均值聚类。
- 所以该算法的基础也可以解释为“最小误差平方和”准则。
- 下面从这个角度来进一步解释，并给出一个“最小误差平方和”准则下的K-均值聚类方法。

## 8.6 K-均值聚类

- 设  $n_i$  表示属于  $\omega_i$  类样本的个数， $\mathbf{m}_i$  是这些样本的均值  
(注：这里将  $\mu_i$  换成  $\mathbf{m}_i$ )：

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

- 考虑对所有样本的一个划分，计算划分后的样本与均值的误差平方和，得到如下“误差平方和”聚类准则：

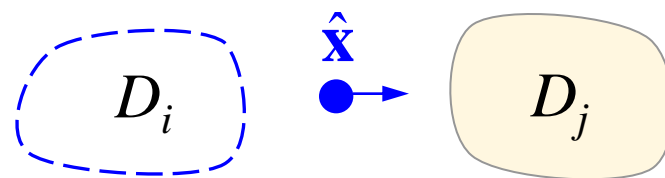
$$J_e = \sum_{i=1}^c J_i, \quad \text{其中,} \quad J_i = \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2$$

- 对于不同的划分（聚类），会得到不同的  $\mathbf{m}_i$ 。因此  $J_e$  的值也是不同的。使  $J_e$  最小的聚类就是误差平方和准则下的最优结果。因此，称这类聚类方法为**最小方差划分法**。

## 8.6 K-均值聚类

- 尽管我们的目标是对样本进行最优划分，但上述准则的**关键之处仍然在于对各均值的估计**。
- 从“正态分布情形下的非监督参数估计”的相关分析可知，难以得到解析解。从  $J_e$  准则的形式上看，也很难得到解析解。因此，需要采用迭代求解技术。
- 每一次迭代就是对样本的一个划分，通过划分的结果才能计算类中心。因此，要不断地调整属于各个类的样本，**有进有出**。
- 因此，下面的重点将介绍**在迭代的过程中如何对样本进行调整**。

## 8.6 K-均值聚类



- 迭代过程中的样本调整:

- 假设样本  $\hat{\mathbf{x}}$  从类  $D_i$  移动到  $D_j$ ，此时，两个类中心将同时进行变化:

$$\mathbf{m}_j^* = \mathbf{m}_j + \frac{\hat{\mathbf{x}} - \mathbf{m}_j}{n_j + 1}, \quad \mathbf{m}_i^* = \mathbf{m}_i - \frac{\hat{\mathbf{x}} - \mathbf{m}_i}{n_i - 1}$$

- 属于第  $j$  类的样本点引起的误差平方和将增加为:

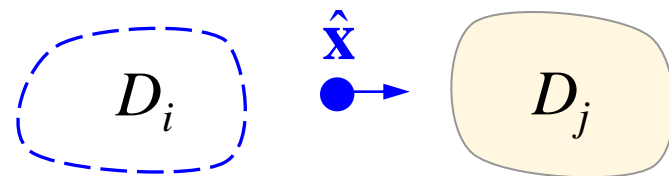
$$J_j^* = \sum_{\mathbf{x} \in D_j} \|\mathbf{x} - \mathbf{m}_j^*\|^2 + \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 = J_j + \frac{n_j \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2}{n_j + 1}$$

属于第  $j$  类的样本点引起的误差平方和将增加，推导如左。

$$\begin{aligned}
 J_j^* &= \sum_{\mathbf{x} \in D_j} \|\mathbf{x} - \mathbf{m}_j^*\|^2 + \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 \\
 &= \sum_{\mathbf{x} \in D_j} \left\| \mathbf{x} - \mathbf{m}_j - \frac{\hat{\mathbf{x}} - \mathbf{m}_j}{n_j + 1} \right\|^2 + \left\| \frac{n_j}{n_j + 1} (\hat{\mathbf{x}} - \mathbf{m}_j) \right\|^2 \\
 &= \sum_{\mathbf{x} \in D_j} \left( \|\mathbf{x} - \mathbf{m}_j\|^2 - \frac{2}{n_j + 1} (\mathbf{x} - \mathbf{m}_j)^T (\hat{\mathbf{x}} - \mathbf{m}_j) + \frac{\|\hat{\mathbf{x}} - \mathbf{m}_j\|^2}{(n_j + 1)^2} \right) + \left\| \frac{n_j}{n_j + 1} (\hat{\mathbf{x}} - \mathbf{m}_j) \right\|^2 \\
 &= J_j - \frac{2}{n_j + 1} (\hat{\mathbf{x}} - \mathbf{m}_j)^T \left( \sum_{\mathbf{x} \in D_j} \mathbf{x} - \sum_{\mathbf{x} \in D_j} \mathbf{m}_j \right) + \frac{n_j \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2}{(n_j + 1)^2} + \left\| \frac{n_j}{n_j + 1} (\hat{\mathbf{x}} - \mathbf{m}_j) \right\|^2 \\
 &= J_j - \frac{2}{n_j + 1} (\hat{\mathbf{x}} - \mathbf{m}_j)^T (n_j \mathbf{m}_j - n_j \mathbf{m}_j) + \frac{n_j \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2}{n_j + 1} \\
 &= J_j + \frac{n_j \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2}{n_j + 1}
 \end{aligned}$$







- 迭代过程中的样本调整：

- 属于第  $i$  类的样本点引起的误差平方和将减少为：

$$J_i^* = J_i - \frac{n_i \|\hat{\mathbf{x}} - \mathbf{m}_i\|^2}{n_i - 1}$$

- 如果**减少量大于增加量**，因此鼓励这种移动，即将样本  $\hat{\mathbf{x}}$  从类  $D_i$  移动到  $D_j$  会减少总体误差：

$$\frac{n_j \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2}{n_j + 1} < \frac{n_i \|\hat{\mathbf{x}} - \mathbf{m}_i\|^2}{n_i - 1} \quad \left( \text{如果减少量大于增加量，此时鼓励移动} \right)$$

从一个类引出样本会减少该类均方误差；但移入样本至一个类会增加该类均方误差。如果**减少量大于增加量**，对这样的样本进行移动是有利于总体误差减少的。

## K-Means Clustering—Algorithm2 (minimum squared error clustering)

- 1 begin initialization  $n, c, \mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_c$ .
- 2 do randomly select a sample  $\hat{\mathbf{x}}$
- 3  $i \leftarrow \arg \min_{i'} \|\mathbf{m}_{i'} - \hat{\mathbf{x}}\|$  // classify  $\hat{\mathbf{x}}$
- 4 if  $n_i \neq 0$ , then compute
$$\rho_j = \begin{cases} \frac{n_j}{n_j+1} \|\hat{\mathbf{x}} - \mathbf{m}_j\|, & j \neq i \\ \frac{n_j}{n_j-1} \|\hat{\mathbf{x}} - \mathbf{m}_j\|, & j = i \end{cases}$$
- 5 find the minimum  $\rho_k$  among all  $\rho_j, j=1,2,\dots,c$
- 6 if  $\rho_k \leq \rho_j$  for all  $j$ , then transfer  $\hat{\mathbf{x}}$  to  $D_k$
- 7 re-compute  $J_e, \mathbf{m}_i, \mathbf{m}_k$
- 8 until no change in  $J_e$  for all  $n$  samples
- 9 return  $\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_c$

## 8.6 K-均值聚类

- K-均值聚类算法是一种典型的**动态聚类方法**，具有如下三个要点：
  - (1) 选择欧氏距离度量作为样本间的相似性度量。
  - (2) 采用最大似然估计或最小均方误差作为评价聚类的准则函数。
  - (3) 给定某个初始分类，然后采用迭代算法寻找准则函数的极值。
- **优点：**
  - 是解决聚类问题的一种经典算法，简单、快速。
  - 对处理大数据集，该算法仍可保持其高效率。
  - 对于密集簇，聚类效果很好。
- **缺点：**
  - 必须事先给定簇的个数，且对初始值敏感。
  - 不适合于发现非凸曲面的簇以及大小相差很大的簇。
  - 对噪声、孤立数据点、野点很敏感。

## 8.6 K-均值聚类

- 关于初始点的选择建议：
  - 凭经验选择初始代表点，根据问题相关性。
  - 将数据随机地分成  $c$  类，计算每类中心，以此作为初始点。
  - 用密度法选择初始点，以每个样本为中心，在一个球形区域内估计样本密度，类似parzen窗方法，逐步地将数据划分至不同的密度区域。
  - 中心分解方法：先将所有数据看成一个聚类，计算聚类中心，然后寻找与该中心最远的点，划入一部分数据点至该最远点所在的区域；对剩下数据，以此类推。

# 8.7 模糊K-均值聚类

- 模糊集的基本知识

- 从集合论的角度，一个类可以看作是一个集合。聚类就是将一个集合划分为若干个子集的过程。
- 1965年，Zadeh 提出了著名的模糊集理论，由此形成了一门新的学科：模糊数学和模糊技术。
- 模糊集理论是对传统集合理论的一种推广。在传统集合理论中，一个元素或者属于一个集合，或者不属于一个集合。对于模糊集而言，一个元素是以一定的程度属于某个集合，也可以以不同的程度属于几个集合。**这一描述引伸出一个重要的概念——模糊集中元素的“隶属度”。**
- 隶属度函数是表示一个对象  $x$  属于集合  $A$  的程度，其自变量的取值范围为所有可能属于集合  $A$  的对象。

# 8.7 模糊K-均值聚类

- 模糊K-均值聚类准则

- 基本出发点：假定样本  $\mathbf{x}_j$  以一定的模糊程度属于某一类，比如第  $i$  类，记为： $\mu_i(\mathbf{x}_j)$ 。该假定也可以理解为  $\mathbf{x}_j$  属于第  $i$  类的概率，即令： $\mu_i(\mathbf{x}_j) = P(\omega_i | \mathbf{x}_j, \theta)$ 。
- 聚类准则修正如下：

$$J_{fuz} = \sum_{i=1}^c \sum_{j=1}^n [\mu_i(\mathbf{x}_j)]^b \|\mathbf{x}_j - \mathbf{m}_i\|^2$$

其中，上标  $b$  是一个自由参数，如果等于0，则退化为经典的 K-均值聚类算法。

- 定义不同的隶属度函数将得到不同的模糊聚类算法。

## 8.7 模糊K-均值聚类

- 一个经典的方法是假定  $\mathbf{x}_j$  属于各类的隶属度之和为1:

$$\sum_{i=1}^c \mu_i(\mathbf{x}_j) = 1, \quad j = 1, 2, \dots, n$$

- 在上述约束条件下, 对  $J_{fuz}$  目标函数求极值, 分别对  $\mathbf{m}_i$  和  $\mu_i(\mathbf{x}_j)$  求偏导数, 并令其等于 0, 则有:

$$\mathbf{m}_i = \frac{\sum_{j=1}^n [\mu_i(\mathbf{x}_j)]^b \mathbf{x}_j}{\sum_{j=1}^n [\mu_i(\mathbf{x}_j)]^b},$$

$$\mu_i(\mathbf{x}_j) = \frac{\left(1/\|\mathbf{x}_j - \mathbf{m}_i\|^2\right)^{1/(b-1)}}{\sum_{k=1}^c \left(1/\|\mathbf{x}_j - \mathbf{m}_k\|^2\right)^{1/(b-1)}}, \quad i = 1, \dots, c, j = 1, \dots, n$$

## 8.7 模糊K-均值聚类

- 算法步骤

---

### Fuzzy K-Means Clustering

---

- 1 begin initialization  $n, c, \mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_c$ ,
  - 2 given  $\mu_i(\mathbf{x}_j)$ ,  $i = 1, 2, \dots, c$ ,  $j = 1, 2, \dots, n$ ,
  - 3 let  $\sum_i \mu_i(\mathbf{x}_j) = 1$ ,  $j = 1, 2, \dots, n$
  - 4 **do** the following computations:
    - 5 update  $\mathbf{m}_j$
    - 6 update  $\mu_i(\mathbf{x}_j)$
  - 7 **until** small change in  $\mathbf{m}_j$  and  $\mu_i(\mathbf{x}_j)$
  - 8 return  $\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_c$
-



# 8.7 模糊K-均值聚类

- 优点：
  - 算法的鲁棒性会更好。
- 缺点：
  - 仍处理不好野点 (outlier)
  - 仍对初始值敏感
  - 仍需知道类别数

## 8.8 常用的聚类准则

- 均方误差准则

$$J_e = \sum_{i=1}^c \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2, \quad \text{其中, } \mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

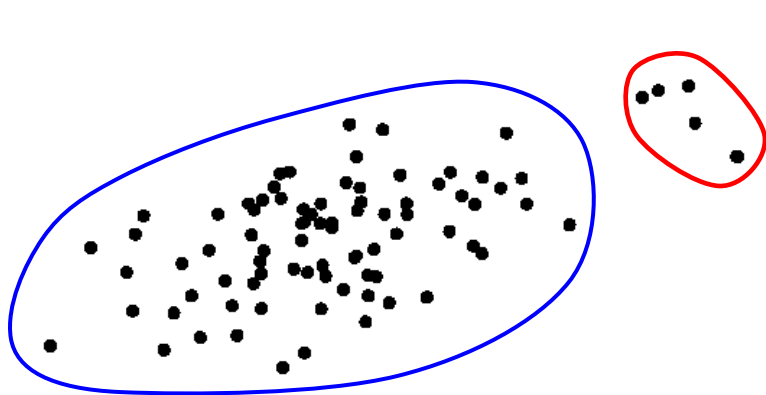
该准则的直观解释是：对于一个给定的簇  $D_i$ ，用其均值作为该簇所有样本的代表点，称为聚类中心。 $J_e$  度量了**总体平方误差**。

$J_e$  的值取决于样本如何被分成不同簇，同时  $J_e$  的值也取决于簇的多少。

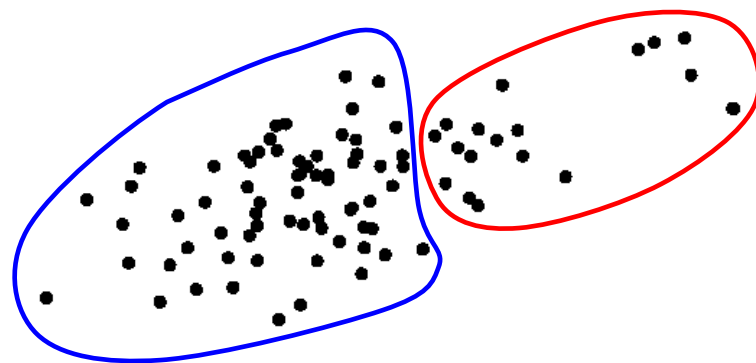
## 8.8 常用的聚类准则

- 均方误差准则的适用范围

- $J_e$  适合于度量簇内数据形成一个紧凑的“云团”的情形。也就是说， $J_e$  不太适用于分散的数据点。
- $J_e$  不太适用于各簇数据不平衡的情形。



$J_e$  更大



$J_e$  更小

## 8.8 常用的聚类准则

- 均方误差准则扩展

- 将簇内数据点的均值带入  $J_e$ ，可得：

$$J_e = \frac{1}{2} \sum_{i=1}^c n_i \bar{s}_i, \quad \text{其中, } \bar{s}_i = \frac{1}{n_i^2} \sum_{\mathbf{x} \in D_i} \sum_{\mathbf{x}' \in D_i} \|\mathbf{x} - \mathbf{x}'\|^2$$

可见,  $\bar{s}_i$  即为属于同一簇  $D_i$  的点对之间的平均距离。  
该表达式同时也表明，欧氏距离将作为相似性的度量方式。

上述形式更容易扩展。也就是说，可以引入其它度量方式来代替点对之间的欧氏距离：

$$\bar{s}_i = \frac{1}{n_i^2} \sum_{\mathbf{x} \in D_i} \sum_{\mathbf{x}' \in D_i} s(\mathbf{x}, \mathbf{x}')$$

## 8.8 常用的聚类准则

- 散度准则

- 散度准则：**类内散度最小，类间散度最大。**
- 散度度量数据点之间的分散程度，采用矩阵表示：

类均值	$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$	$\mathbf{m} = \frac{1}{n} \sum_{\mathbf{x} \in D} \mathbf{x} = \frac{1}{n_i} \sum_{i=1}^c n_i \mathbf{m}_i$	总均值
类散度	$\mathbf{S}_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$	$\mathbf{S}_W = \sum_{i=1}^c \mathbf{S}_i$	总类内散度
类间散度	$\mathbf{S}_B = \sum_{i=1}^c n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$	$\mathbf{S}_T = \sum_{\mathbf{x} \in D} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T$	总散度
$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$			

## 8.8 常用的聚类准则

- 为什么叫散度？
  - **类散度矩阵**实际上为协方差矩阵（相差一个系数），其主对角线元素代表方差，因此具有数据分布“分散程度”的含义。从宏观上刻画样本之间的离散程度。
  - **总类内散度矩阵**为类内散度矩阵之和，刻画：“从总体来看类内各个样本与其所在类之间的离散度”。
  - **类间散度矩阵**则描述类与类之间的总体离散程度。

## 8.8 常用的聚类准则

- 散度与距离之间的关系：

- 设一个簇的中心点为  $\mathbf{m}$ 。对于该簇的一个样本  $\mathbf{x}$ ，它对类散度矩阵的贡献为： $(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T \in \mathbb{R}^{d \times d}$ 。
- 该**矩阵的迹**等于  $(\mathbf{x}-\mathbf{m})^T(\mathbf{x}-\mathbf{m})$ ，即样本  $\mathbf{x}$  到类中心点  $\mathbf{m}$  的距离的平方。
- 因此，类散度矩阵的迹等于类内所有点到类中心点的距离平方和：

$$\sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2 = \text{tr}(\mathbf{S}_i), \quad i = 1, 2, \dots, c$$

- 该和反应了样本分布的聚集程度。该和越小，数据分布越紧凑。

## 8.8 常用的聚类准则

- 总类内散度迹最小准则

- 根据散度矩阵的迹与距离的关系，有如下关于总类内散度的迹的关系式：

$$J_e = \sum_{i=1}^c \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2 = \sum_{i=1}^c \text{tr}(\mathbf{S}_i) = \text{tr} \left( \sum_{i=1}^c \mathbf{S}_i \right) = \text{tr}(\mathbf{S}_W)$$

- 可见，**总类内散度迹最小准则与类均方误差最小准则是等价的。**

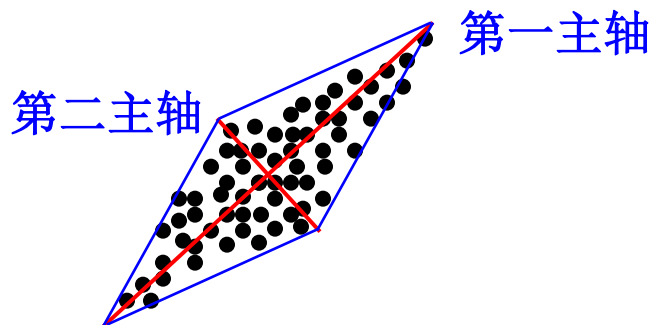
- 类间散度最大准则

$$\max \quad \text{tr}(\mathbf{S}_B) = \sum_{i=1}^c n_i \|\mathbf{m} - \mathbf{m}_i\|^2$$



## • 行列式准则

- 一个矩阵的行列式等于该矩阵的所有特征值之乘积。
- 对于数据的协方差矩阵而言，其第一个特征值反映数据沿第一主轴的分布。该值越大，数据沿此轴分布越长。



两个主轴大小之积可描述面积大小。

- 协方差矩阵的行列式正比于数据的分布所占的空间体积（平方）。
- 最小化总类内行列式准则： $\min J_d = |S_w|$
- 通常不采用这一准则的原因： $S_w$  可能非奇异！

## • 不变性准则

- 在  $\mathbf{S}_W$  为非奇异矩阵时，可以证明  $\text{tr}((\mathbf{S}_W)^{-1}\mathbf{S}_B)$  不会因为对数据施加一个任意的非奇线性变换而改变。
- 由于实对称矩阵的迹等于其所有特征值之和，于是有：

---

$$\text{tr}(\mathbf{S}_W^{-1}\mathbf{S}_B) = \sum_{i=1}^d \lambda_i, \quad |\mathbf{S}_W^{-1}\mathbf{S}_B| = \prod_{i=1}^d \lambda_i \quad (\text{最大化})$$

---

$$\text{tr}(\mathbf{S}_T^{-1}\mathbf{S}_W) = \sum_{i=1}^d \frac{1}{1 + \lambda_i}, \quad \frac{|\mathbf{S}_W|}{|\mathbf{S}_T|} = \prod_{i=1}^d \frac{1}{1 + \lambda_i} \quad (\text{最小化})$$

---

其中， $\lambda_i$  为矩阵  $(\mathbf{S}_W)^{-1}\mathbf{S}_B$  的特征值。

---

由于  $\mathbf{S}_T$  并不依赖于数据如何划分，所以最小化  $|\mathbf{S}_W|$  与最小化  $|\mathbf{S}_W|/|\mathbf{S}_T|$  是等价的。

Thank All of You!