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

第二章 聚类分析

2.1 聚类的基本概念

2.2 模式相似性测度

2.3 类的定义与类间距离

2.4 聚类算法

2.1 聚类的基本概念

（一）聚类分析的基本思想

（二）特征量的类型

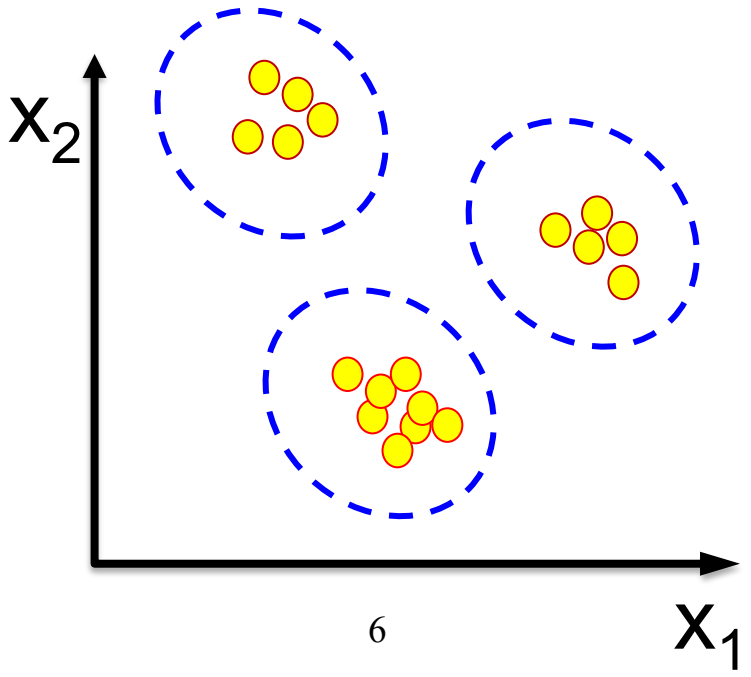
（三）方法的有效性

（一）聚类分析基本思想

假设：对象集客观存在着若干个自然类，每个自然类中个体的某些属性具有较强的相似性。

（一）聚类分析基本思想

方法：根据待分类的模式属性或特征相似程度进行分类，相似的模式归为一类，不相似的模式分划到不同类中。



（一）聚类分析基本思想

方法：根据待分类的模式属性或特征相似程度进行分类，相似的模式归为一类，不相似的模式分划到不同类中。

基本内容

- 特征提取
- 模式相似性度量
- 点与类间的距离
- 类与类间的距离
- 聚类准则及聚类算法
- 有效性分析

2.1 聚类的基本概念

（一）聚类分析的基本思想

（二）特征量的类型

（三）方法的有效性

（二）特征量的类型

- **物理量：直接反映特征的实际物理意义。**

  如:长度、重量、速度等。处理前需要离散化。

- **次序量：反映特征的次序关系或等级。**

  如:产品的等级、病症的级或期。已是离散量。

- **名义量：反映样本的状态特征，非数值的。**

  如男性与女性、事物的状态、种类等。需要数值化。这些特征的数值指标既无数量含义，也无次序关系，只是用数字代表各种状态。

2.1 聚类的基本概念

（一）聚类分析的基本思想

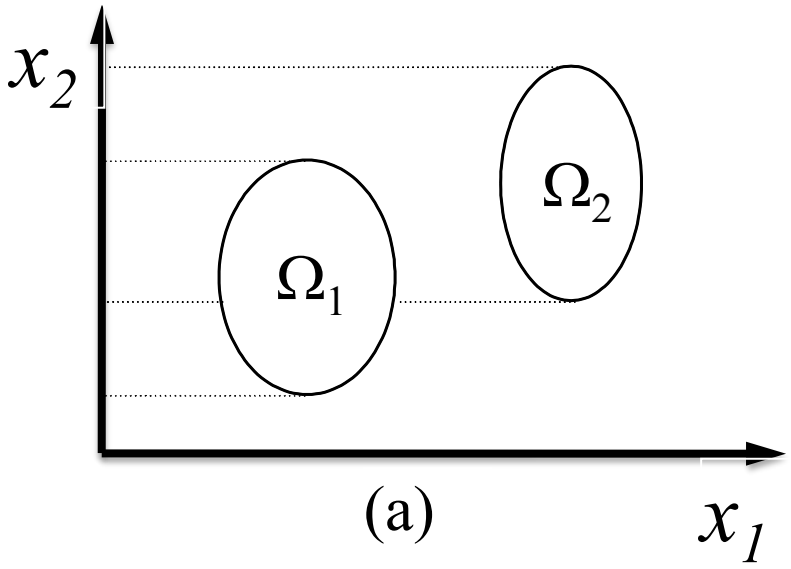
（二）特征量的类型

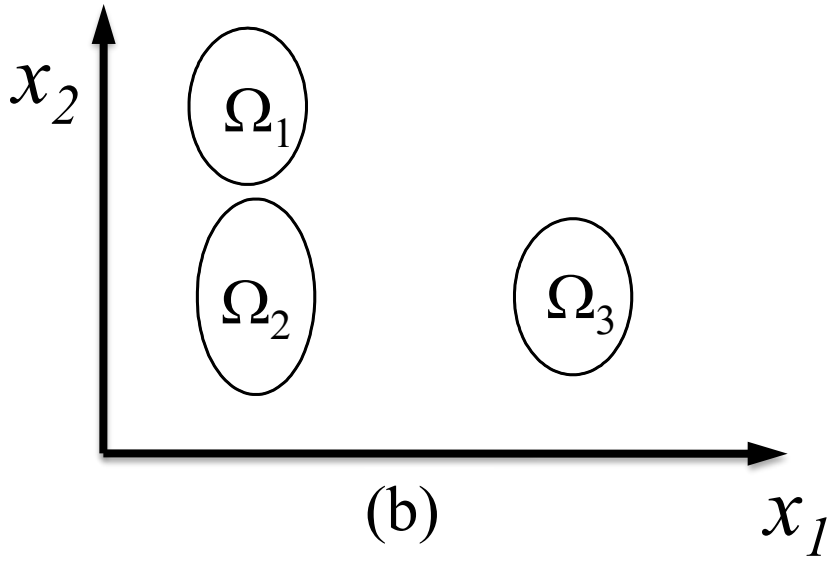
（三）方法的有效性

（三）方法的有效性

(1) 特征选取不当或不足使分类无效

(2) 特征选取过多可能有害无益，且增加分析负担

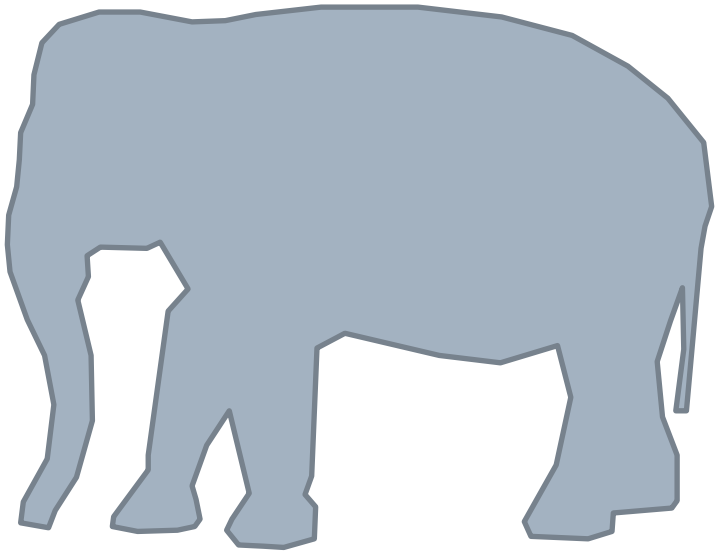




（三）方法的有效性

这是什么？

- A 墙纸
- B 动物
- C 天空
- C 织物



这是什么？

- A 大象
- B 长颈鹿
- C 斑马

班克斯(Banksy)

（三）方法的有效性

这是什么？

- A 墙纸
- B 动物
- C 天空
- C 织物

The Elephant in the Room by #Banksy

（三）方法的有效性

房间里的大象y #Banksy

Sep 16, 2006 - Los Angeles, CA, USA ...
alamy.com


Banksy | enwp.org ...
flickr.com


Banksy | Elephant art, Banksy art, Banksy
pinterest.at


Sep 16, 2006 - Los Angeles, CA, USA ...
alamy.com


Sep 16, 2006 - Los Angeles, CA, USA ...
alamy.com

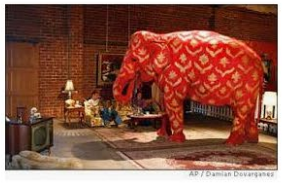

Sep 16, 2006 - Los Angeles, CA, USA ...
alamy.com


Banksy: quando la provocazione diventa ...
successo.com


Sep 16, 2006 - Los Angeles, CA, USA ...
alamy.com


Stephen Lemon questions fintech ...
businessinsider.com


Eléphant | A Lady in the City
aladyinthecity.wordpress.com


Street artist Banksy welcomes visitors ...
chron.com


Tai, a 38-year-old Asian elephant ...
stock.adobe.com


Artist Gallery | Widewalls
widewalls.ch


When Is It Okay to Use A...
artsy.net


bansky | Elephant painting ...
pinterest.nz


London Art News: Jolie buys Banksy art
coxsoft.blogspot.com


Banksy- Elephant in the Room...Thanks ...
pinterest.ca

Is the Banksy exhibition in Sydney ...
londonerinsydney.com


Artwork by Banksy Ne...
gettyimages.fi

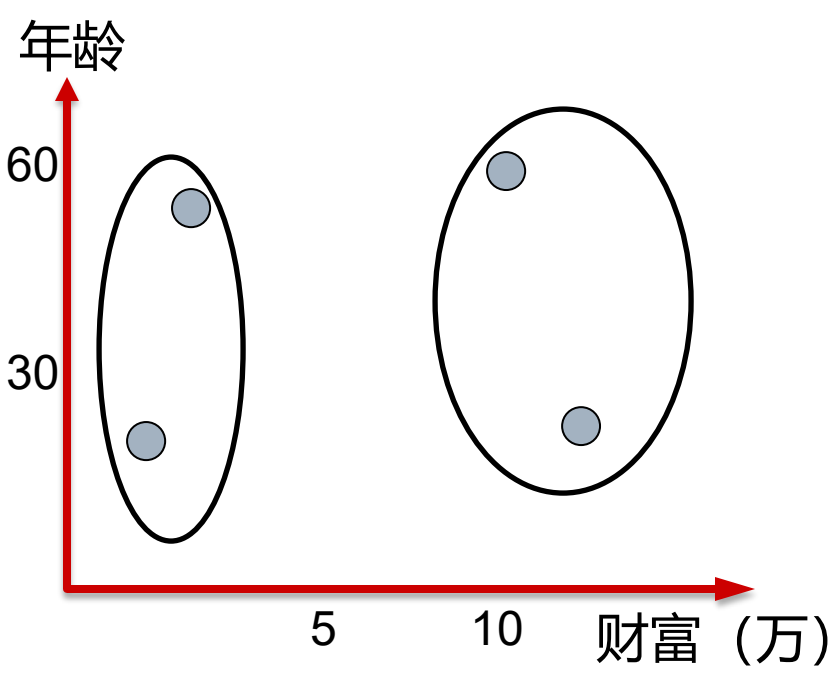

red dirt mule | Elephant art, Elephant ...
pinterest.com

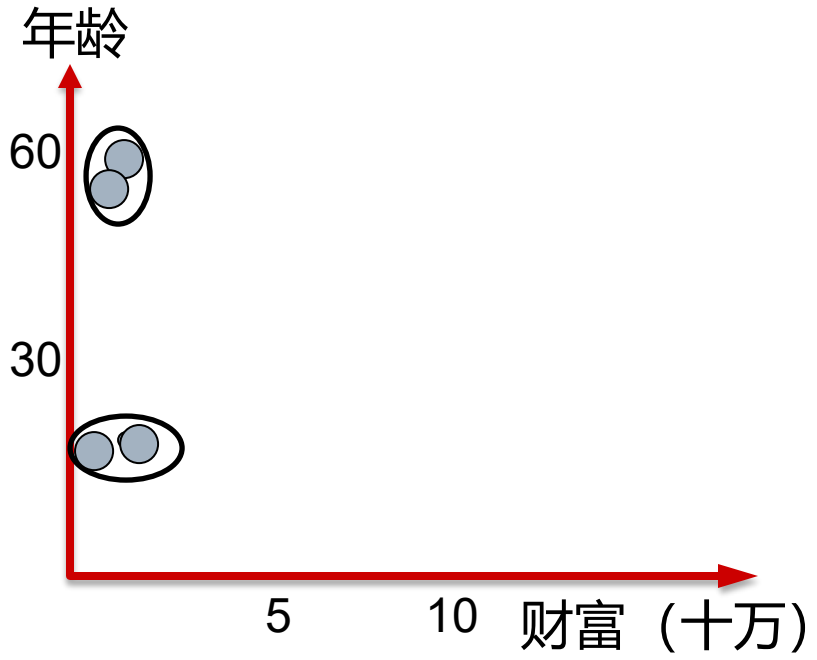

Street Artist" Thierry Guetta ...
stencilrevolution.com

The Elephant in the Room by #Banksy

（三）方法的有效性

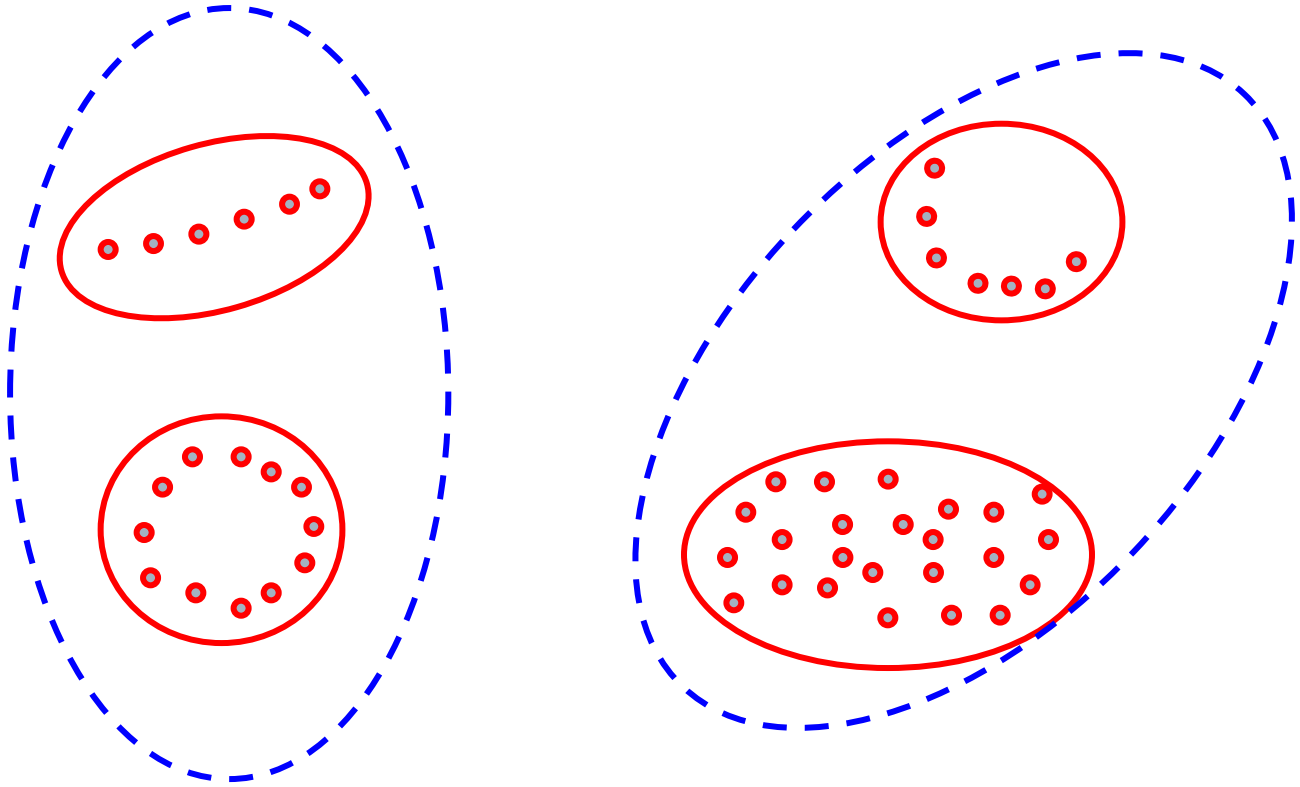
(3) 特征量纲对聚类结果的影响





（三）方法的有效性

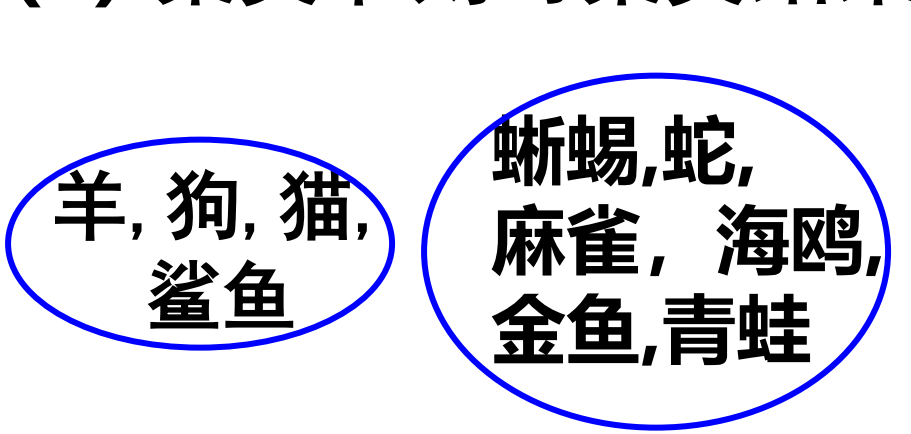
(4) 距离测度对聚类结果的影响



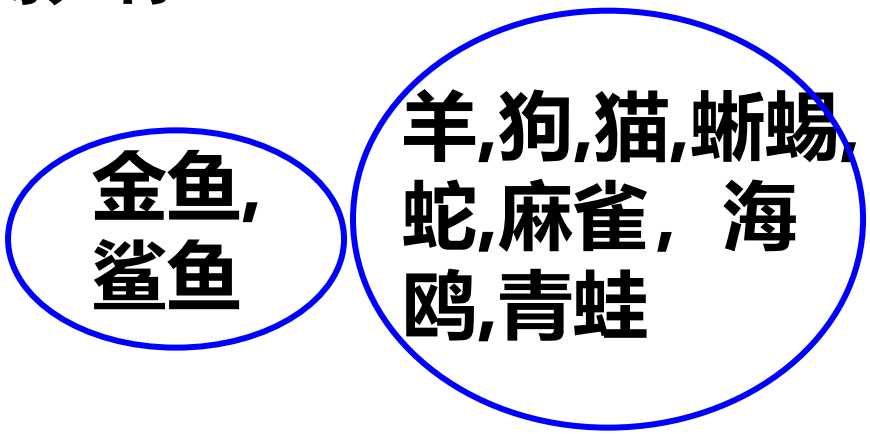
数据的粗聚类是两类,细聚类为4类

（三）方法的有效性

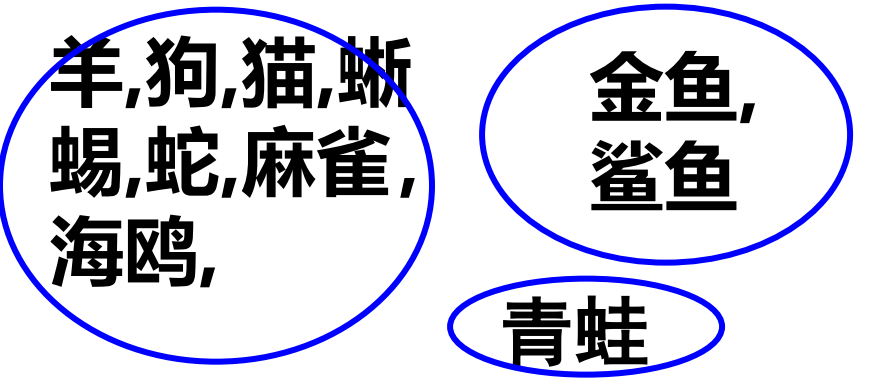
(5) 聚类准则对聚类结果的影响



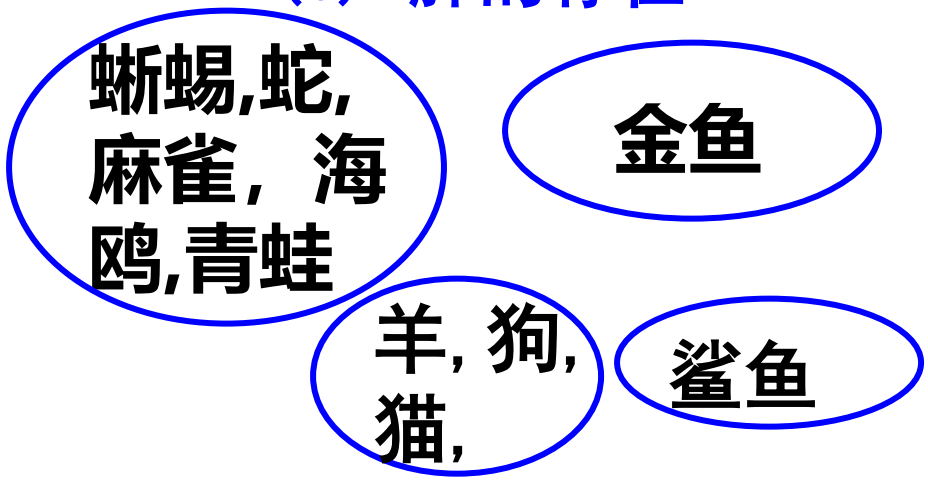
(a) 繁衍后代的方式



（b） 肺的存在



(c) 生存环境



(d) 繁衍后代的方式和是否存在肺

第二章 聚类分析

2.1 聚类的基本概念

2.2 模式相似性测度

2.3 类的定义与类间距离

2.4 聚类算法

2.2 模型相似性测度

（一） 距 离 测 度

（二） 相 似 测 度

（三） 匹 配 测 度

（一）距离测度

Distance (or Dissimilarity) Measure

设特征矢量 $\vec{x}$ 和 $\vec{y}$ 的距离为 $d(\vec{x},\vec{y})$

则 $d(\vec{x},\vec{y})$ 一般应满足如下公理

(1) $d(\vec{x},\vec{y}) \geq 0$, 当且仅当 $\vec{x}=\vec{y}$ 时等号成立，即 $d(\vec{x},\vec{y})=0 \Leftrightarrow \vec{x}=\vec{y}$

(2) $d(\vec{x},\vec{y})=d(\vec{y},\vec{x})$

(3) $d(\vec{x},\vec{y}) \leq d(\vec{x},\vec{z})+d(\vec{z},\vec{y})$ **(triangular inequality)**

（一）距离测度

- **概念**

  以两个矢量矢端的距离作为度量基础，
  距离测度值是两矢量各相应分量之差的函数 。

- **方法**
  - 欧氏(Euclidean)距离
  - 绝对值距离(街坊距离或Manhattan距离)
  - 切氏(Chebyshev)距离
  - 明氏(Minkowski)距离
  - Camberra距离(Lance距离、Willims距离)
  - 马氏(Mahalanobis)距离

（一）距离测度

设 $\vec{x}=(x_1,x_2,\cdots,x_n)'$ ， $\vec{y}=(y_1,y_2,\cdots,y_n)'$

(1) 欧氏（Euclidean）距离

$$d(\vec{x},\vec{y})=\|\vec{x}-\vec{y}\|=[\sum_{i=1}^{n}(x_i-y_i)^2]^{1/2}$$

(2) 绝对值距离（街坊距离或Manhattan距离）

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

(3) 切氏（Chebyshev）距离

$$d(\vec{x},\vec{y})=\max_i|x_i-y_i|$$

切氏距离 ≤ 欧氏距离 ≤ 绝对值距离 （作业）

（一）距离测度

(4) 明氏(Minkowski)距离

$$d(\vec{x},\vec{y})=[\sum_{i=1}^{n}|x_i-y_i|^m]^{1/m}$$

m=2:欧式距离，m=1:绝对值距离，m=∞:切氏距离

(5) Cambera距离(Lance距离、Willims距离)

$$d(\vec{x},\vec{y})=\sum_{i=1}^{n}\frac{|x_i-y_i|}{|x_i+y_i|}\quad(x_i,y_i\geq 0,x_i+y_i\neq 0)$$

该距离能克服量纲的影响，
但不能克服分量间的相关性。

（一）距离测度

(6)马氏(Mahalanobis)距离

设n维矢量$\vec{x}_i$和$\vec{x}_j$是矢量集$\{\vec{x}_1,\vec{x}_2,...,\vec{x}_m\}$中的两个矢量，它们的马氏距离定义为

$$d^2(\vec{x}_i,\vec{x}_j)=(\vec{x}_i-\vec{x}_j)^{\mathrm{T}}V^{-1}(\vec{x}_i-\vec{x}_j)$$

式中 $V=\dfrac{1}{m-1}\sum_{i=1}^{m}(\vec{x}_i-\bar{\vec{x}})(\vec{x}_i-\bar{\vec{x}})^{\mathrm{T}}$，$\bar{\vec{x}}=\dfrac{1}{m}\sum_{i=1}^{m}\vec{x}_i$

性质：对一切非奇异线性变换都是不变的。即，具有坐标系比例、旋转、平移不变性，并且从统计意义上尽量去掉了分量间的相关性。

马氏距离具有非奇异线性变换不变性

证明： 设，有非奇异线性变换：$\vec{y} = A\vec{x}$

则 $\vec{\bar{y}} = \frac{1}{m}\sum_{i=1}^{n}\vec{y}_i = \frac{1}{m}\sum_{i=1}^{n}A\vec{x}_i = A\frac{1}{m}\sum_{i=1}^{n}\vec{x}_i = A\vec{\bar{x}}$

$$V_y = \frac{1}{m-1}\sum_{i=1}^{m}(\vec{y}_i - \vec{\bar{y}})(\vec{y}_i - \vec{\bar{y}})'$$

$$= \frac{1}{m-1}\sum_{i=1}^{m}(A\vec{x}_i - A\vec{\bar{x}})(A\vec{x}_i - A\vec{\bar{x}})'$$

$$= \frac{1}{m-1}\sum_{i=1}^{m}A(\vec{x}_i - \vec{\bar{x}})(\vec{x}_i - \vec{\bar{x}})'A' \qquad \{\because (AB)' = B'A'\}$$

$$= A[\frac{1}{m-1}\sum_{i=1}^{m}(\vec{x}_i - \vec{\bar{x}})(\vec{x}_i - \vec{\bar{x}})']A' = AV_xA'$$

马氏距离具有非奇异线性变换不变性

故 $d_y^2(\vec{y}_i,\vec{y}_j)=(\vec{y}_i-\vec{y}_j)'V_y^{-1}(\vec{y}_i-\vec{y}_j)$

$=(A\vec{x}_i-A\vec{x}_j)'V_y^{-1}(A\vec{x}_i-A\vec{x}_j)$

$=(\vec{x}_i-\vec{x}_j)'A'V_y^{-1}A(\vec{x}_i-\vec{x}_j)$

$=(\vec{x}_i-\vec{x}_j)'A'(AV_xA')^{-1}A(\vec{x}_i-\vec{x}_j)$

$=(\vec{x}_i-\vec{x}_j)'A'A'^{-1}V_x^{-1}A^{-1}A(\vec{x}_i-\vec{x}_j)$ $\{\because (AB)^{-1}=B^{-1}A^{-1}\}$

$=(\vec{x}_i-\vec{x}_j)'V_x^{-1}(\vec{x}_i-\vec{x}_j)$

$=d_x^2(\vec{x}_i,\vec{x}_j)$

马氏距离的一般定义

设 $\vec{x}$ 、 $\vec{y}$ 是从期望矢量为 $\vec{\mu}$ 、协方差矩阵为Σ的母体G中抽取的两个样本，则它们间的马氏距离定义为

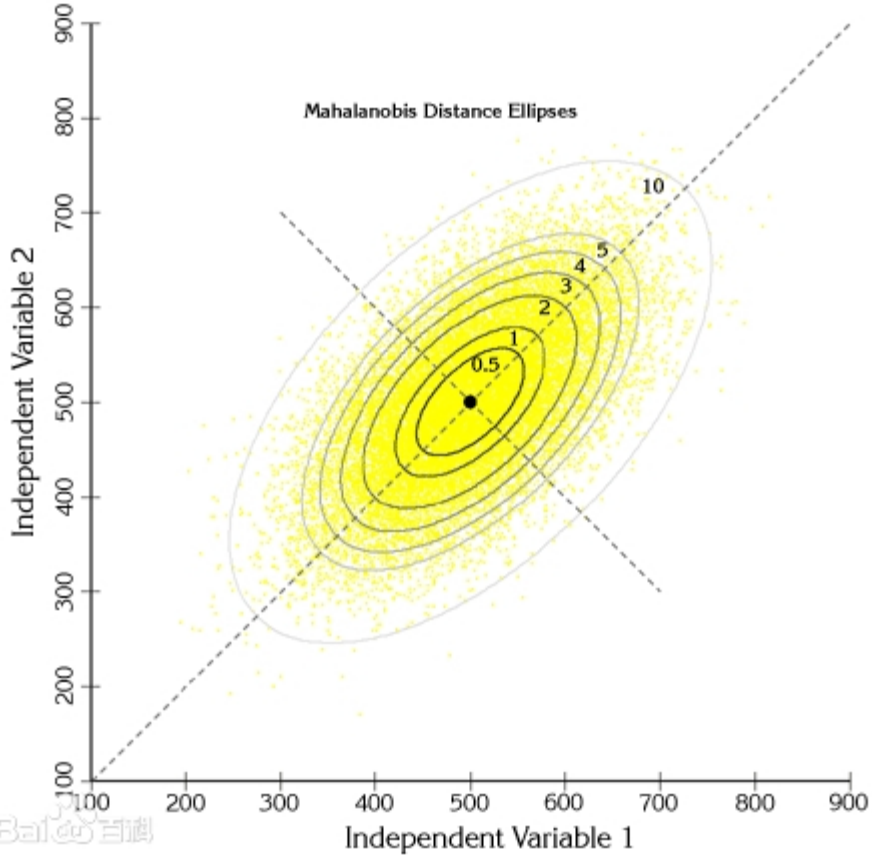
$$d^2(\vec{x},\vec{y}) = (\vec{x}-\vec{y})'\Sigma^{-1}(\vec{x}-\vec{y})$$

当 $\vec{x}$ 和 $\vec{y}$ 是分别来自两个数据集中的样本时，设C是它们的互协方差阵，则它们间的马氏距离定义为

$$d^2(\vec{x},\vec{y}) = (\vec{x}-\vec{y})'C^{-1}(\vec{x}-\vec{y})$$

- **当Σ、V、C为单位矩阵时，马氏距离⇔欧氏距离。**
- **对于正态分布，等概率密度点轨迹是到均值矢量的马氏距离为常数的点所构成的超椭球面。**

马氏距离的物理意义


Mahalanobis Distance Ellipses
10
5
4
3
2
1
0.5
Independent Variable 2
Independent Variable 1
100
200
300
400
500
600
700
800
900

例2.1

已知一个二维正态母体G的分布为 $N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}\right)$

求点 $A:\begin{pmatrix}1\\1\end{pmatrix}$ 和 $B:\begin{pmatrix}1\\-1\end{pmatrix}$ 至均值点 $M:\ \vec{\mu}=\begin{pmatrix}0\\0\end{pmatrix}$ 的距离。

解：由题设，可得 $\Sigma=\begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}$ $\Sigma^{-1}=\dfrac{1}{0.19}\begin{pmatrix}1 & -0.9\\-0.9 & 1\end{pmatrix}$

从而马氏距离

$$d_M{}^2(A,M)=\begin{pmatrix}1 & 1\end{pmatrix}\Sigma^{-1}\begin{pmatrix}1\\1\end{pmatrix}=0.2\big/0.19 \qquad d_M{}^2(B,M)=\begin{pmatrix}1 & -1\end{pmatrix}\Sigma^{-1}\begin{pmatrix}1\\-1\end{pmatrix}=3.8\big/0.19$$

它们之比达 $\sqrt{19}$ 倍。

若用欧氏距离，则算得的距离值相同：

$$d_E^2(A,M)=2 \qquad d_E^2(B,M)=2$$

由分布函数知，A、B两点的概率密度分别为

$$p(1,1)=0.2157 \qquad p(1,-1)=0.00001658$$

2.2 模型相似性测度

（一） 距 离 测 度

（二） 相 似 测 度

（二）相似测度

- **概念**

  重点考虑两矢量的方向是否相近，而忽略矢量长度。

- **方法**
  - 角度相似系数
  - 相关系数
  - 指数相似系数
  - 最值相似系数

（二）相似测度

- **重点考虑两矢量的方向是否相近，而忽略矢量长度。**

(1) 角度相似系数(夹角余弦)

矢量之间的相似性可用它们的夹角余弦来度量

$$\cos(\vec{x},\vec{y})=\frac{\vec{x}'\vec{y}}{\|\vec{x}\|\|\vec{y}\|}=\frac{\vec{x}'\vec{y}}{[(\vec{x}'\vec{x})(\vec{y}'\vec{y})]^{1/2}}$$

(2) 相关系数

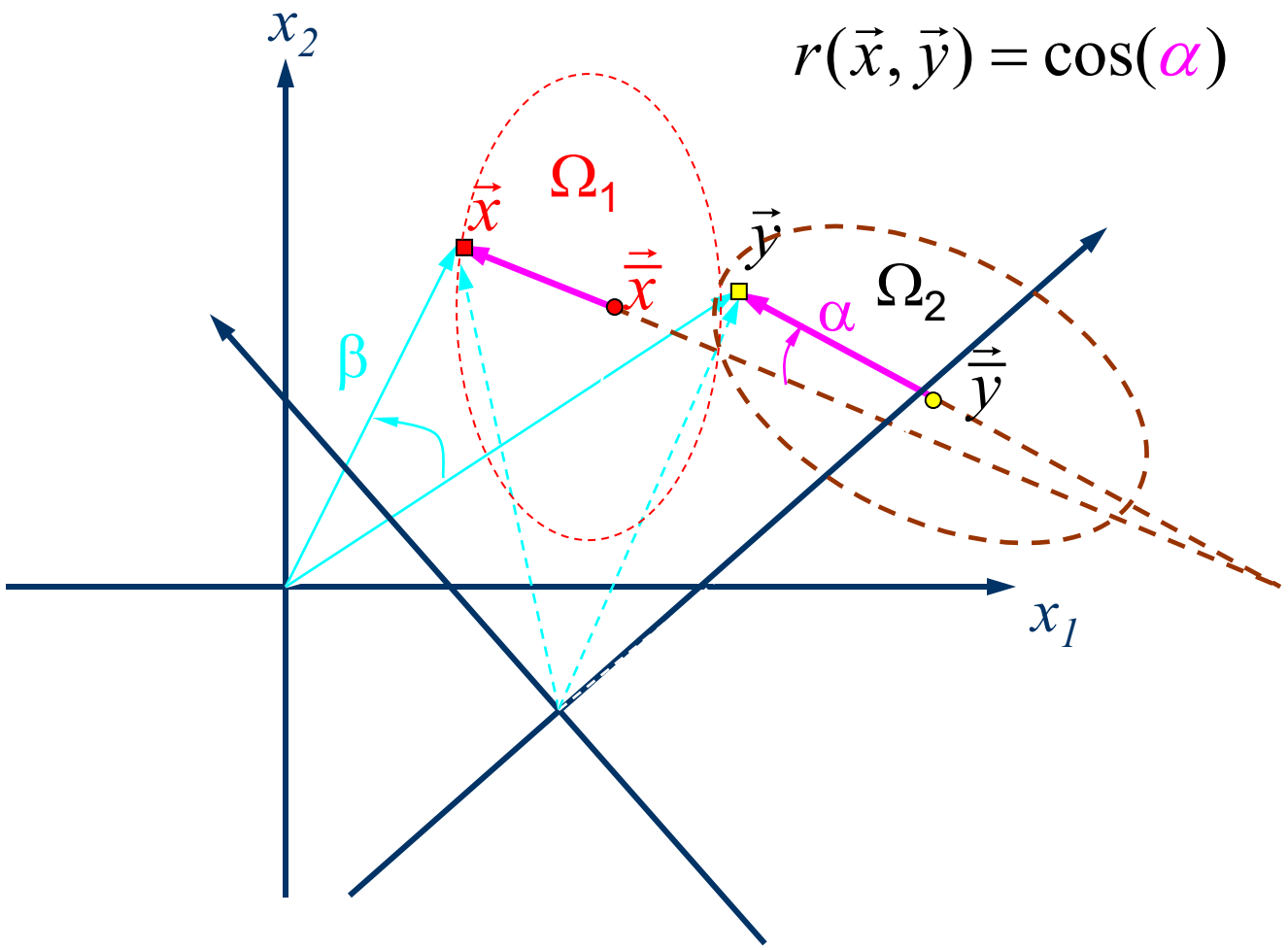
数据中心化后的矢量夹角余弦

$$r(\vec{x},\vec{y})=\frac{(\vec{x}-\bar{\vec{x}})'(\vec{y}-\bar{\vec{y}})}{[(\vec{x}-\bar{\vec{x}})'(\vec{x}-\bar{\vec{x}})(\vec{y}-\bar{\vec{y}})'(\vec{y}-\bar{\vec{y}})]^{1/2}}$$

性质：相关系数具有坐标系平移、旋转、比例不变性。

相关系数的几何意义

$$r(\vec{x},\vec{y})=\cos(\alpha)$$



（二）相似测度

(3) 指数相关系数

$$e(\vec{x},\vec{y})=\frac{1}{n}\sum_{i=1}^{n}\exp[-\frac{3}{4}\frac{(x_i-y_i)^2}{\sigma_i^2}]$$

这里假设 $\vec{x}$ 和 $\vec{y}$ 的维数n相同、概率分布相同。

σ_i^2 是第i个分量的方差。

性质：不受量纲变化的影响。

从函数构造看属于距离度量，

但由于差矢量长度对其的影响并不大，

所以将其归为相似性测度。

小 结

一、影响分类的因素

（1）分类准则；（2）特征量的选择；（3）量纲。

二、模式相似性测度

（一） 距 离 测 度

(1) 欧氏距离

(2) 马氏距离 $d^2(\vec{x}_i,\vec{x}_j)=(\vec{x}_i-\vec{x}_j)'V^{-1}(\vec{x}_i-\vec{x}_j)$

对坐标系平移、旋转、比例不变（非奇异线性变换不变）。

（二） 相 似 测 度

相关系数 (特征矢量的方向) $r(\vec{x},\vec{y})=\dfrac{(\vec{x}-\bar{\vec{x}})'(\vec{y}-\bar{\vec{y}})}{[(\vec{x}-\bar{\vec{x}})'(\vec{x}-\bar{\vec{x}})(\vec{y}-\bar{\vec{y}})'(\vec{y}-\bar{\vec{y}})]^{1/2}}$

对坐标系平移、旋转、比例不变。

（三） 匹 配 测 度

(0-1) 匹配系数, Tanimoto相似性测度

2.3 类间距离

2.3.1 类间距离测度方法

（一）最近距离

两个聚类ω_k和ω_l之间的最近距离定义为

$$D_{kl} = \min_{i,j}[d_{ij}]$$

式中，d_{ij}表示$x_i \in \omega_k$与$x_j \in \omega_l$间的距离。

如果ω_l由ω_p和ω_q两类合并而成，则有递推公式

$$D_{kl} = \min[D_{kp}, D_{kq}]$$

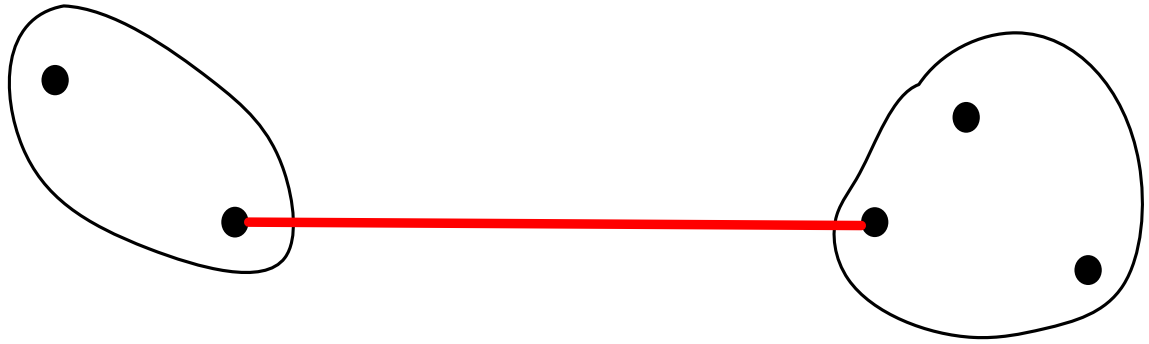
（二）最远距离

$$D_{kl} = \max_{i,j}[d_{ij}]$$

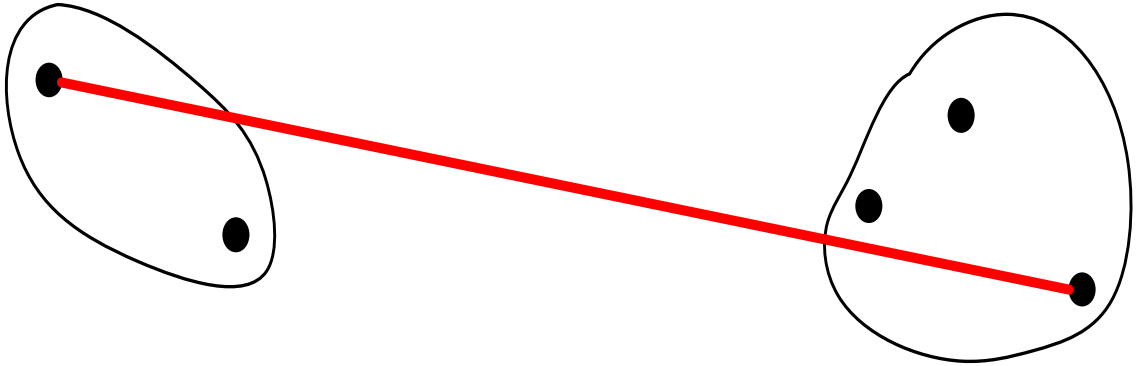
递推公式

$$D_{kl} = \max[D_{kp}, D_{kq}]$$

2.3.1 类间距离测度方法



最近距离法图示



最远距离法图示

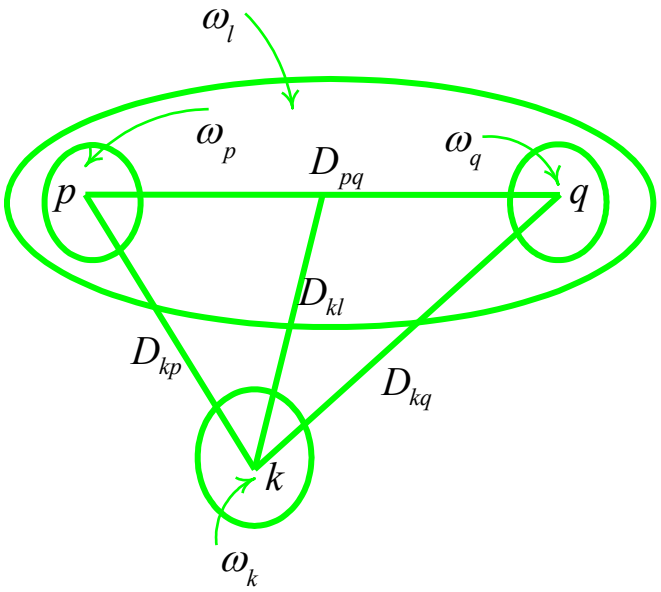
2.3.1 类间距离测度方法

设ω_i有n_i个样本，$i=k,l,p,q$；如果ω_l由ω_p和ω_q两类合并而成。

（三）中间距离

递推公式

$$D_{kl}^2 = \frac{1}{2}D_{kp}^2 + \frac{1}{2}D_{kq}^2 - \frac{1}{4}D_{pq}^2$$



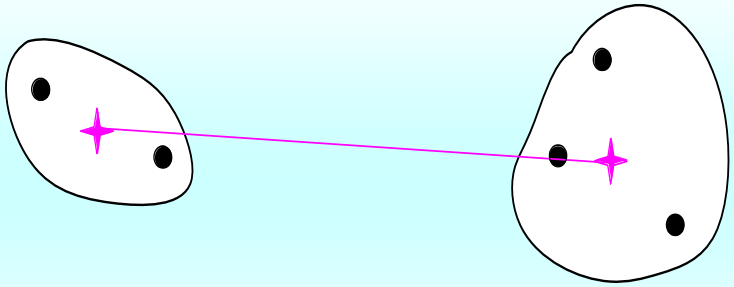
2.3.1 类间距离测度方法

设ω_i有n_i个样本，$i=k,l,p,q$；如果ω_l由ω_p和ω_q两类合并而成。

（四）重心距离

$$D_{pq}^2=(\vec{x}_p-\vec{x}_q)'(\vec{x}_p-\vec{x}_q)$$

$\vec{x}_p$ 和$\vec{x}_q$分别是ω_p和ω_q的重心



重心距离法图示

递推公式

$$D_{kl}^2=\frac{n_p}{n_p+n_q}D_{kp}^2+\frac{n_q}{n_p+n_q}D_{kq}^2-\frac{n_pn_q}{(n_p+n_q)^2}D_{pq}^2$$

式中 $D_{ij}^2=(\vec{x}_i-\vec{x}_j)'(\vec{x}_i-\vec{x}_j)$ ，$\vec{x}_i$ 和$\vec{x}_j$分别是ω_i和ω_j的重心，$i,j=k,l,p,q$ 。

2.3.1 类间距离测度方法

（五） 平均距离

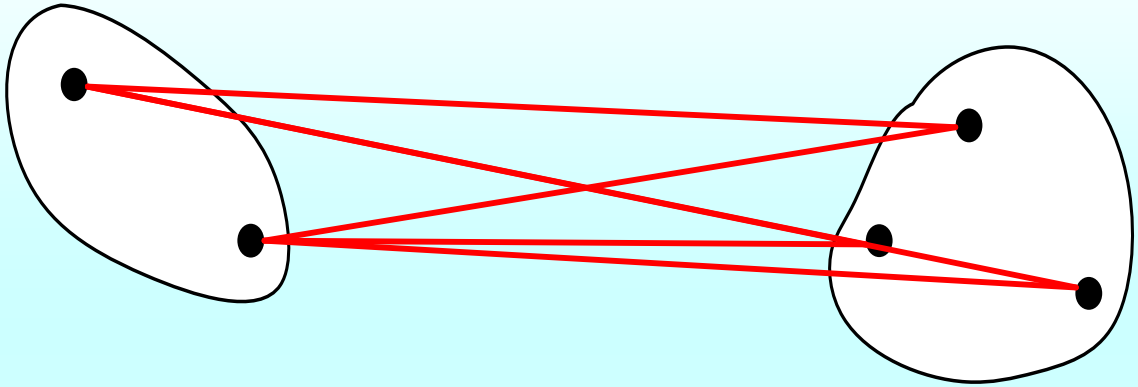
两类ω_p和ω_q间的距离平方定义为这两类元素两两之间的平均平方距离，即

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{\vec{x}_i \in \omega_p, \vec{x}_j \in \omega_p} d_{ij}^2$$

设$\omega_l = \omega_p \cup \omega_q$，类平均距离的递推公式为

$$D_{kl}^2 = \frac{n_p}{n_p + n_q} D_{kp}^2 + \frac{n_q}{n_p + n_q} D_{kq}^2$$

2.3.1 类间距离测度方法



平均距离法图示

2.3.1 类间距离测度方法

（六） 离差平方和法

设类ω_t的重心是$\vec{x}_t$，　ω_t的类内离差平方和定义为

$$s_t = \sum_{\vec{x}_i \in \omega_t} (\vec{x}_i - \vec{x}_t)'(\vec{x}_i - \vec{x}_t)$$

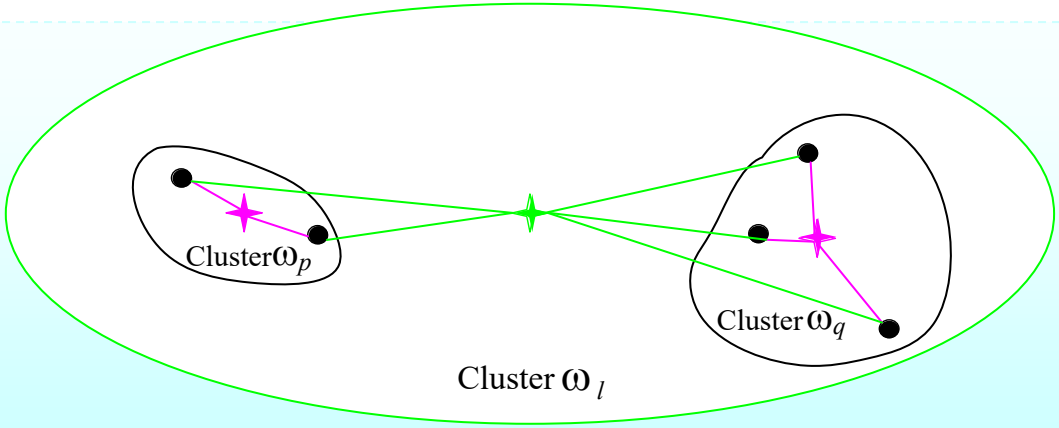
设$\omega_l = \omega_p \cup \omega_q$，则$s_l$要变大。把两类合并所增加的离差平方和定义为两类平方距离，即$D^2_{pq} = s_l - s_p - s_q$ 可以证明

$$D^2_{pq} = \frac{n_p n_q}{n_p + n_q}(\vec{x}_p - \vec{x}_q)'(\vec{x}_p - \vec{x}_q) \quad \vec{x}_p = \frac{1}{n_p}\sum_{\vec{x}_i \in \omega_p} \vec{x}_i, \quad \vec{x}_q = \frac{1}{n_q}\sum_{\vec{x}_j \in \omega_q} \vec{x}_j$$

ω_k与$\omega_l = \omega_p \cup \omega_q$的离差平方和的递推公式

$$D^2_{kl} = \frac{n_k + n_p}{n_k + n_l} D^2_{kp} + \frac{n_k + n_q}{n_k + n_l} D^2_{kq} - \frac{n_k}{n_k + n_l} D^2_{pq}$$

2.3.1 类间距离测度方法



离差平方和法图示

递推公式

$$D_{kl}^2 = \frac{n_k + n_p}{n_k + n_l} D_{kp}^2 + \frac{n_k + n_q}{n_k + n_l} D_{kq}^2 - \frac{n_k}{n_k + n_l} D_{pq}^2$$

谱系聚类算法中多采用平均距离法和离差平方和法计算类间距离。

类间距离递推公式

$$D_{kl}^2 = \alpha_p D_{kp}^2 + \alpha_q D_{kq}^2 + \beta D_{pq}^2 + \gamma \left| D_{kp}^2 - D_{kq}^2 \right| \text{（其中} \omega_l = \omega_p \cup \omega_q \text{）}$$

	α_p	α_q	β	γ
最近距离	1/2	1/2	0	-1/2
最远距离	1/2	1/2	0	1/2
中间距离	1/2	1/2	-1/4	0
重心距离	$n_p/(n_p+n_q)$	$n_q/(n_p+n_q)$	$-\alpha_p\alpha_q$	0
平均距离	$n_p/(n_p+n_q)$	$n_q/(n_p+n_q)$	0	0
可变平均距离	$(1-\beta)n_p/(n_p+n_q)$	$(1-\beta)n_q/(n_p+n_q)$	<1	0
可变距离	$(1-\beta)/2$	$(1-\beta)/2$	<1	0
离差平方和	$(n_k+n_p)/(n_k+n_l)$	$(n_k+n_q)/(n_k+n_l)$	$-n_k/(n_k+n_l)$	0

$$\min[a,b] = [\frac{a^2+b^2}{2} - \frac{|a^2-b^2|}{2}]^{1/2}, \quad \max[a,b] = [\frac{a^2+b^2}{2} + \frac{|a^2-b^2|}{2}]^{1/2}$$

2.3.1 类间距离测度方法

各种类间距离的性能

(1) **最远距离**：类域半径增长最慢，倾向成团聚类

(2) **最近距离**：可能产生细长分布的类

(3) **中间距离**：克服重心距离聚类合并后重心靠近较大聚类重心的问题

(4) **离差平方和**：倾向于将孤立点或较小类与较大类合并。

(5) **平均距离和重心距离**：对孤立点和噪声不敏感

2.4 准则函数

2. 4. 1 点与点的集合的距离

$$\text{if } d(\vec{x},\omega_i)=\min_{j=1,2,\ldots,c}[d(\vec{x},\omega_j)],\ \text{then } \vec{x}\in\omega_i$$

其中，$d(\vec{x},\omega_i)$为待分类模式$\vec{x}$到聚合类ω_i的距离

（一）第一类（在无集合中点的分布的先验知识时用）

（1）最小距离 $d(\vec{x},\omega_i)=\min_{\vec{y}\in\omega_i}[\vec{x},\vec{y}]$

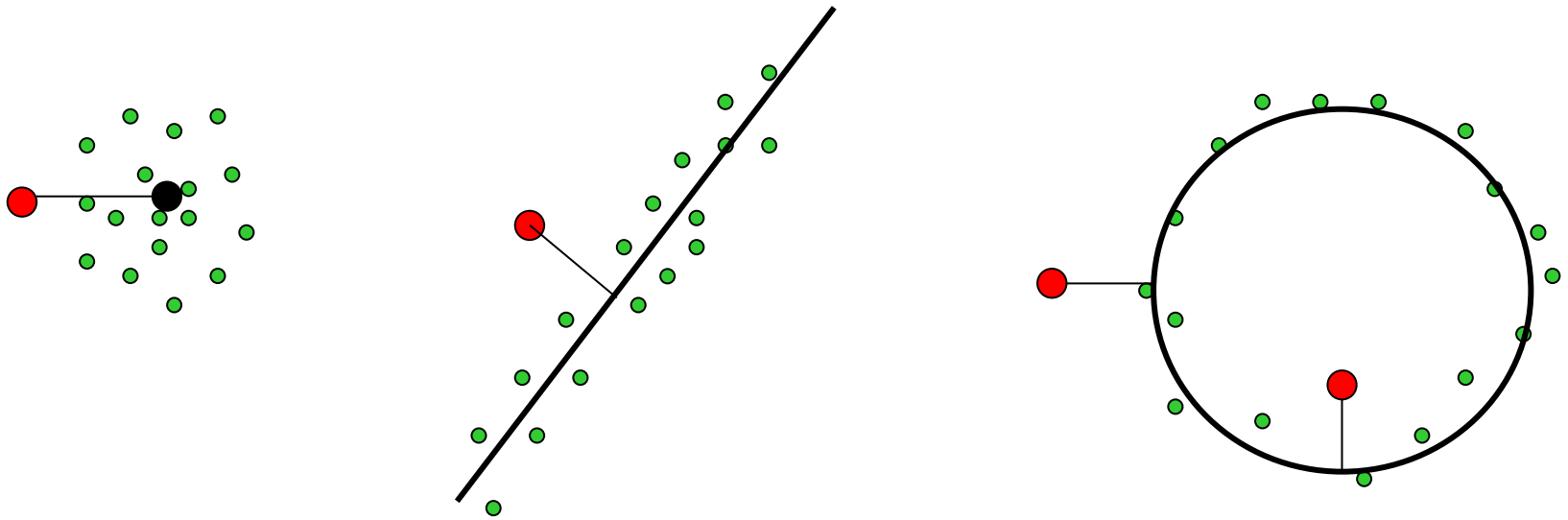
（2）最大距离 $d(\vec{x},\omega_i)=\max_{\vec{y}\in\omega_i}[\vec{x},\vec{y}]$

（3）平均距离 $d(\vec{x},\omega_i)=\frac{1}{N_i}\sum_{\vec{y}\in\omega_i}d(\vec{x},\vec{y})$

2.4.1 点与点的集合间的距离

(二) 第二类（有集合中点的分布的先验知识时用）

点到代表点集的**模型**的距离



(a) 团状分布
点模型

(b) 线状分布
超平面模型

(c) 环状分布
超球面模型

2.4.1 点与点的集合间的距离

（二）第二类（有集合中点的分布的先验知识时用）

点到代表点集的**模型**的距离

1）点模型 （若集合为团状分布）

（1）均矢 $\vec{m}_a = \frac{1}{N_i}\sum_{\vec{y}\in\omega_i}\vec{y}$ ，其中，N_i为ω_i类中的点数。

（2）最值中心：集合ω_i中的某一点$\vec{m}_c$若满足

$$\sum_{\vec{y}\in\omega_i} d(\vec{m}_c,\vec{y}) = \min_{\vec{x}_j\in\omega_i}[\sum_{\vec{y}\in\omega_i} d(\vec{x}_j,\vec{y})]$$

则称$\vec{m}_c$为集合ω_i中的最值中心。

2.4.1 点与点的集合间的距离

（二）第二类（有集合中点的分布的先验知识时用）

点到代表点集的**模型**的距离

1）点模型 （若集合为团状分布）

（3）中值中心：集合ω_i中的某一点$\vec{m}_m$若满足

$$\underset{\vec{y}\in\omega_i}{med}[d(\vec{m}_m,\vec{y})]=\underset{\substack{\vec{y}\in\omega_i\\ \vec{z}\in\omega_i}}{med}[d(\vec{y},\vec{z})]$$

则称$\vec{m}_m$为集合ω_i中的中值中心。

待分类模式$\vec{x}$到聚合类ω_i的距离定义为$d(\vec{x},\vec{m}_i),\ \ (i=a,c,m)$

$d(\vec{x},\vec{m}_i)$ 是$\vec{x}$到$\vec{m}_i$ 的欧氏距离、明氏距离或马氏距离。

2.4.1 点与点的集合间的距离

(二) 第二类（有集合中点的分布的先验知识时用）

点到代表点集的**模型**的距离

2）超平面模型 （若集合为平面状分布）

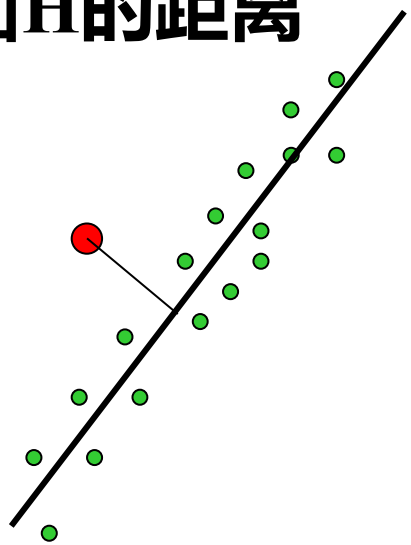
待分类模式 $\vec{x}$ 到代表聚合类 ω_i 的某超平面H的距离

$$H:\vec{w}'\vec{z}+w_0=0$$

其中 $\vec{w}=(w_1,w_2,\cdots,w_n)',\vec{z}=(z_1,z_2,\cdots,z_n)'$

$$d(\vec{x},H)=\min_{\vec{z}\in H}[d(\vec{x},\vec{z})]$$

采用欧氏距离时有 $$d(\vec{x},H)=\frac{|\vec{w}'\vec{x}+w_0|}{\|\vec{w}\|}$$



(b) 线状分布 超平面模型

2.4.1 点与点的集合间的距离

(二) 第二类（有集合中点的分布的先验知识时用）

点到代表点集的**模型**的距离

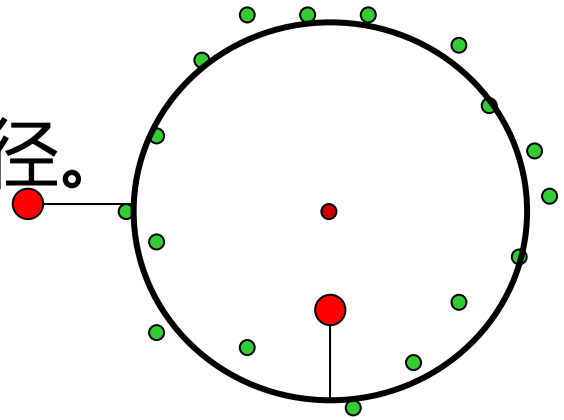
3）超球面模型 （若集合为超球面状分布）

待分类模式 $\vec{x}$ 到代表聚合类 ω_i 的某超球面S的距离

$$S:(\vec{z}-\vec{c})'(\vec{z}-\vec{c})=r^2$$

其中，$\vec{c}$ 和 r，为超球面的中心和半径。

$$d(\vec{x},S)=\min_{\vec{z}\in S}[d(\vec{x},\vec{z})]$$



(c) 环状分布
超球面模型

2.4.2 聚类准则函数

设待分类模式 $\{\vec{x}_i, i=1,2,...,N\}$，将它们分成c类，分类后各模式记为 $\{\vec{x}_i^{(j)}, j=1,2,\cdots,c, i=1,2,...,n_j\}$，$n_j$是$\omega_j$中的样本个数。

设计准则评价分类过程或分类结果的优劣

（一）类内距离准则（误差平方和准则）

$$J_W=\sum_{j=1}^{c}\sum_{i=1}^{n_j}\left\|\vec{x}_i^{(j)}-\vec{m}_j\right\|^2=\sum_{j=1}^{c}\sum_{i=1}^{n_j}(\vec{x}_i^{(j)}-\vec{m}_j)'(\vec{x}_i^{(j)}-\vec{m}_j)\to\min$$

$$\vec{m}_j=\frac{1}{n_j}\sum_{i=1}^{n_j}\vec{x}_i^{(j)}\quad(j=1,2,\cdots,c)$$

- **适用于各类模式呈团状分布的情况。**

2.4.2 聚类准则函数

（二）类间距离准则

$$J_B = \sum_{j=1}^{c} (\vec{m}_j - \vec{m})'(\vec{m}_j - \vec{m}) \to \max$$

式中，$\vec{m} = \frac{1}{N}\sum_{i=1}^{N} \vec{x}_i$ 是总的样本均值矢量，$\vec{m}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \vec{x}_i^{(j)} \quad (j = 1, 2, \cdots, c)$

对于两类问题，可以定义

$$J_{B2} = (\vec{m}_1 - \vec{m}_2)'(\vec{m}_1 - \vec{m}_2)$$

2.4.2 聚类准则函数

（三）基于类内、类间距离的准则函数

构造能同时使$J_w \to \min$和$J_B \to \max$的准则函数

- 类内离差矩阵（Scatter Matrix）

$$S_W^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j}(\vec{x}_i^{(j)} - \vec{m}_j)(\vec{x}_i^{(j)} - \vec{m}_j)', \quad \vec{x}_i^{(j)} \in \omega_j, \vec{m}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\vec{x}_i^{(j)}, (j = 1, 2, \cdots, c)$$

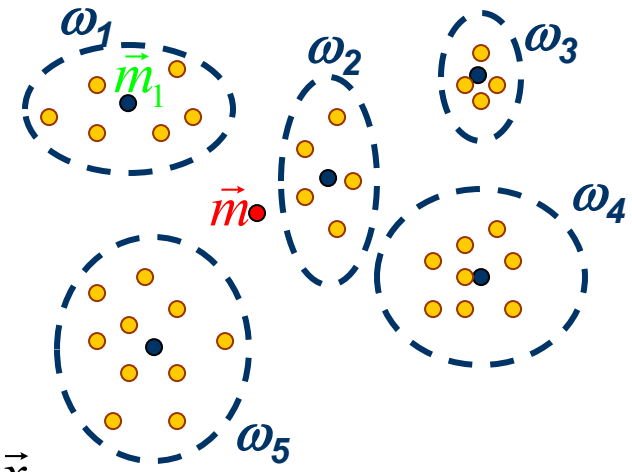
- 总的类内离差矩阵

$$S_W = \sum_{j=1}^{c}\frac{n_j}{N}S_W^{(j)}$$

- 类间离差矩阵

$$S_B = \sum_{j=1}^{c}\frac{n_j}{N}(\vec{m}_j - \vec{m})(\vec{m}_j - \vec{m})'$$

$$\vec{m} = \frac{1}{N}\sum_{i=1}^{N}\vec{x}_i$$



2.4.2 聚类准则函数

(三) 基于类内、类间距离的准则函数

- 总的离差矩阵

$$S_T = \frac{1}{N}\sum_{i=1}^{N}(\vec{x}_i - \vec{m})(\vec{x}_i - \vec{m})'$$

- 总的离差矩阵和类内离差和类间离差的关系：

$$S_T = S_W + S_B$$

S_W、S_B和S_T分别从不同方面反映了模式散布的结构信息，通常采用它们的数值特征，如迹、行列式、特征值等构造准则函数。

2.4.2 聚类准则函数

(三) 基于类内、类间距离的准则函数

聚类的基本目标是使 J_{WB}=Tr[S_B]→max 和 J_{WW} =Tr[S_W]→min

因此可定义如下聚类准则函数

$$J_1 = Tr\left[S_W^{-1} S_B\right]$$

$$J_2 = \left|S_W^{-1} S_B\right|$$

$$J_3 = Tr\left[S_W^{-1} S_T\right]$$

$$J_4 = \left|S_W^{-1} S_T\right|$$

$J_i \rightarrow \max$，$(i=1,2,3,4)$

即，**类内越“紧”，类间越“开”，聚类效果越好。**

小 结

- 聚类的基本概念
- 模式相似性测度
  - 欧氏(Euclidean)距离
  - 马氏(Mahalanobis)距离
- 类的定义与类间距离
- 聚类算法

End