

# 模式识别

## 第7讲 特征提取与选择

2018~2019学年



# 内容安排

---

一、绪论、数学基础（第1讲）

二、聚类分析（第2讲）

三、判别函数分类法（几何分类法）（第3、4讲）

四、统计决策分类法（概率分类法）（第5、6讲）

五、特征提取与选择（第7讲）

六、模糊模式识别（第8讲）

七、神经网络模式识别（第9讲）

期末考试（平时作业：40%，期末考试：60%）

## 五、特征提取与选择

---

5.1 基本概念

5.2 类别可分性测度

5.3 基于类内散布矩阵的单类模式特征提取

5.4 基于K-L变换的多类模式特征提取

## 5.1 基本概念

---

### 1、用于分类的模式特征的特点

在模式识别过程中，模式特征的确定比较复杂。研究领域不同，选择的特征也不同，但不论采用什么样的特征，都应该满足以下条件：

(1) **特征是可获取的**。因为模式识别系统的主要处理设备是计算机，所以作为观察对象的数字化表达，观察对象应该是可以通过数据采集设备输入到计算机的。如采集图像信息的图像卡和采集语音信息的声卡等。作为特征，既可以是数字化表达的结果，也可以是在数字化表达基础上形成的参数性质的值，如图像分割后的子目标特性表达。

(2) **类内聚集**。选择的特征对同一类应具有稳定性。由于模式类是由具有相似特性的若干个样本构成的，因此它们同属一类，首要前提是特性相似，反映在取值上，就应该有较好的稳定性。

(3) **类间离散**。选择的特征对不同的类应该有差异。若不同类的模式的特征值差异很小，则说明所选择的特征对于不同的类没有什么差异，作为分类的依据时，容易使不同的类产生混淆，使误识率增大。一般来讲。特征

的类间差异应该大于类内差异。

## 2、特征类别

### (1) 物理特征（观测型特征）

物理特征是比较直接、人们容易感知的特征，一般在设计模式识别系统时容易被选用。如为了描述指定班级中的某个学生，可以用以下物理特征：性别、身高、胖瘦、肤色等外在特征。物理特征虽然容易感知，却未必能非常有效地表征分类对象。

## (2) 结构特征（几何型特征）

结构特征的表达是先将观察对象分割成若干个基本构成要素，再确定基本要素间的相互连接关系。结构特征的表达能力一般要高于物理特征，在实际中已经有较多的成功应用，如指纹识别、汉字识别的成功都离不开结构特征的选择。

通过要素和相互连接关系表达对象，可以较好地表达复杂的图像图形信息。结构信息对对象的尺寸往往不太敏感。

结构特征比物理特征要抽象一些，但仍属比较容易感知的特征，如人的指纹特征、人脸的五官结构信息等，是认定人的身份的重要参数。

### (3) 数字特征（标记型特征）

一般来说，数字特征是为了表征观察对象而设立的特征，如给每个学生设立一个学号，作为标志每个学生的特征。由于学号是人为设定的，可以保证唯一性，但这种特征是抽象的，不容易被人感知。数字特征有时和观察对象的固有特性没有任何联系，有时则是物理或结构特征的计算结果。



### 3、特征的形成

在设计一个具体的模式识别系统时，往往是先接触一些训练样本，由领域专家和系统工程师联合研究模式类所包含的特征信息，并给出相应的表述方法。这一阶段的主要目标是获取尽可能多的表述特征。在这些特征中，有些可能满足类内聚集、类间离散的要求，有的则可能不满足，不能作为分类的依据。根据样例分析得到一组表述观察对象的特征值，而不论特征是否实用，称这一步为**特征形成**，得到的特征称为**原始特征**。

在这些原始特征中，有的特征对分类有效，有的则不起什么作用。若在得到一组原始特征后，不加筛选，全部用于分类函数确定，则有可能存在无效特征、相关特征，这既增加了分类决策的复杂度，还可能恶化分类器的性能。为此，需要对原始特征集进行处理，去除对分类作用不大的特征，从而可以在保证性能的条件下，通过降低特征空间的维数来减少分类方法的复杂度。

实现上述目的的方法有两种：特征提取和特征选择。

特征提取和特征选择都不考虑针对具体应用需求的原始特征形成过程，而是假设原始特征形成工作已经完成。然而在实际工作中，原始特征的获得并不容易，因为人具有非常直观的识别能力，因此反而很难为计算机“明确描述”用于分类的特性依据。如人脸的判定，人识别脸部特征非常容易，若用计算机来识别人脸，则需要得到多达上千个特征，难度很大。可以说，特征形成是模式识别过程中的重点和难点之一。

## 4、特征提取和特征选择

**特征选择**是指从一组特征中挑选出对分类最有利的特征，同时达到降低特征空间维数的目的。

**特征提取**是指通过变换的方法获取最有效并有利于分类的特征，实现特征维数的降低。变换后的特征称为**二次特征**，它们是原始特征的某种组合，最常用的是线性组合。

**主要目的：**在不降低或尽可能少地降低分类器性能的情况下，降低特征空间的维数。

## 5、类别可分性判据

- ◆ 衡量不同特征及其组合对分类是否有效的定量准则
- ◆ **理想准则**：某组特征使分类器错误概率最小
- ◆ 实际的类别可分性判据应满足的条件：
  - 度量特性： $J_{ij} > 0$ , if  $i \neq j$ ;  $J_{ij} = 0$ , if  $i = j$ ;  $J_{ij} = J_{ji}$
  - 与错误率有单调关系（可分性判据越大，误分率越小）
  - 当特征独立时有可加性： $J_{ij}(x_1, x_2, \dots, x_d) = \sum_{k=1}^d J_{ij}(x_k)$
  - 单调性： $J_{ij}(x_1, x_2, \dots, x_d) \leq J_{ij}(x_1, x_2, \dots, x_d, x_{d+1})$
- ◆ 常见类别可分性判据：
  - 基于距离度量
  - 基于概率分布
  - 基于熵函数

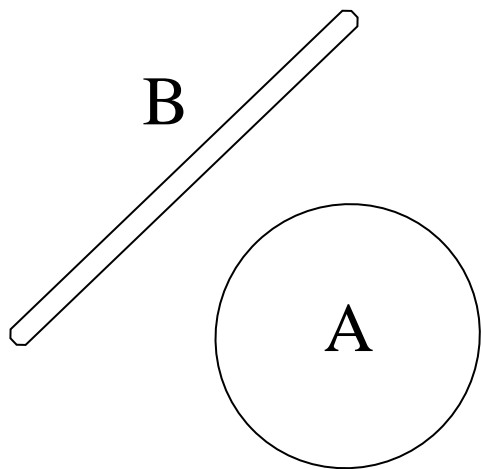
## 6、对分类特征的要求

- (1) 具有很大的识别信息量。即应具有**很好的类别可分性**。
- (2) 具有**可靠性**。模棱两可、似是而非、时是时非等不易判别的特征应丢掉。
- (3) 尽可能强的**特征间独立性**。相关性强的特征只选一个。
- (4) **数量尽量少**，同时损失的分类信息尽量小。

当模式在空间中发生移动、旋转、缩放时，特征值应保持不变，保证仍可得到同样的识别效果。

例：特征选择与特征提取的区别：对一个条形和圆进行识别。

(Toy example)



解：[法1]

① 特征抽取：测量三个结构特征

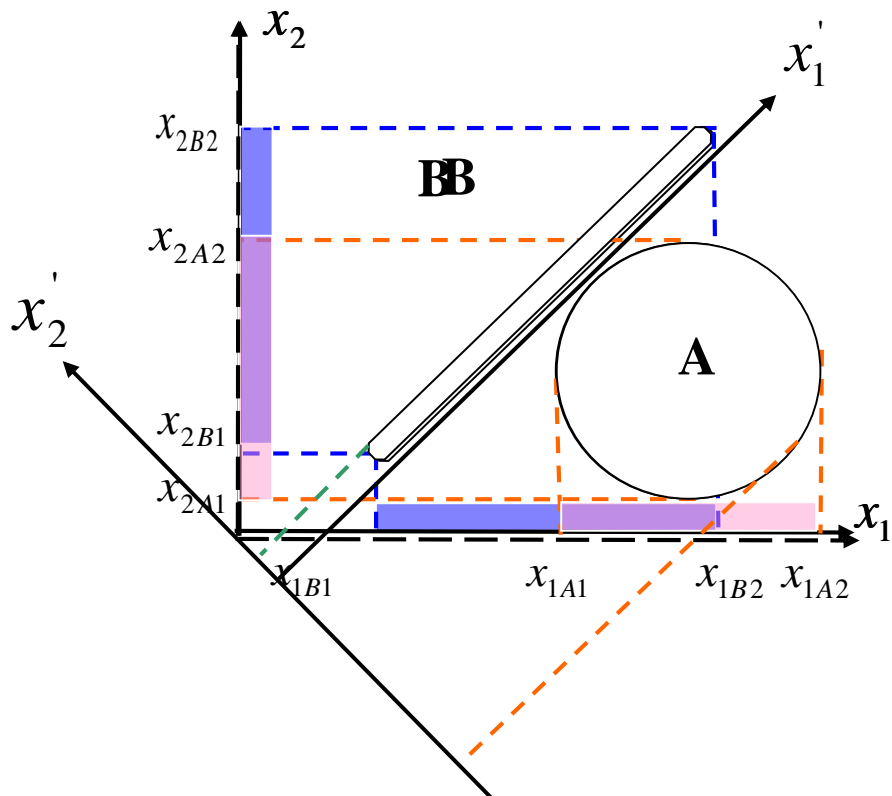
(a) 周长

(b) 面积

(c) 两个互相垂直的内径比

② 分析：(c)是具有分类能力的特征，故选(c)，  
扔掉(a)、(b)。

——特征选择：一般根据物理特征或结构特征进行压缩。



[法2]: ① 特征抽取: 测量物体向两个坐标轴的投影值, 则A、B各有2个值域区间。可以看出, 两个物体的投影有重叠, 直接使用投影值无法将两者区分开。

② 分析: 将坐标系按逆时针方向做一旋转变换, 或物体按顺时针方向变, 并适当平移等。根据物体在  $x'_2$  轴上投影的坐标值的正负可区分两个物体。

——特征提取, 一般用数学的方法进行压缩。



## 5.2 类别可分性测度

类别可分性测度：衡量类别间可分性的尺度。

类别可分性测度 { 几何分布：类内距离和类间距离  
概率分布：类概率密度间的距离  
误判概率：与错误率有关的距离

相似性测度：衡量模式之间相似性的一种尺度

### 5.2.1 基于距离的可分性测度

#### 1. 类内散布距离和类内散布矩阵

1)  $\omega_i$  类的类内距离一：类内（样本间）均方两两距离。

$$\overline{D_i^2} \triangleq E[\| \mathbf{X}_k - \mathbf{X}_l \|^2 | \omega_i] = E[(\mathbf{X}_k - \mathbf{X}_l)^T (\mathbf{X}_k - \mathbf{X}_l) | \omega_i]$$

其中 $\mathbf{X}_k$ 和 $\mathbf{X}_l$ 均为同一类模式样本。

若 $\{\mathbf{X}_i\}$ 中的样本相互独立，有

$$\begin{aligned}\overline{D_i^2} &= E[(\mathbf{X}_k^T \mathbf{X}_k - \mathbf{X}_k^T \mathbf{X}_l - \mathbf{X}_l^T \mathbf{X}_k + \mathbf{X}_l^T \mathbf{X}_l) | \omega_i] \\ &= 2E[\mathbf{X}_k^T \mathbf{X}_k] - 2E[\mathbf{X}_k^T]E[\mathbf{X}_l] = 2\text{Tr}[\mathbf{R}_i - \mathbf{M}_i \mathbf{M}_i^T] \\ \overline{D_i^2} &= E[(\mathbf{X}_k - \mathbf{M}_i + \mathbf{M}_i - \mathbf{X}_l)^T (\mathbf{X}_k - \mathbf{M}_i + \mathbf{M}_i - \mathbf{X}_l) | \omega_i] \\ &= E[(\mathbf{X}_k - \mathbf{M}_i)^T (\mathbf{X}_k - \mathbf{M}_i) + (\mathbf{X}_l - \mathbf{M}_i)^T (\mathbf{X}_l - \mathbf{M}_i) | \omega_i] \\ &= 2\text{Tr}[\mathbf{C}_i] = 2 \sum_{k=1}^n \sigma_k^2\end{aligned}$$

式中， $\mathbf{R}_i$ ：第*i*类的自相关矩阵；

$\mathbf{M}_i$ ：第*i*类的均值向量；  $\mathbf{C}_i$ ：第*i*类的协方差矩阵；

$\sigma_k^2$ ： $\mathbf{C}_i$ 主对角线上的元素，表示模式向量第*k*个分量的方差；

Tr：矩阵的迹（方阵主对角线上各元素之和）。

2)  $\omega_i$  类的类内距离二：类内（样本间）均方散布距离。

$$\overline{D_{w,i}^2} \triangleq E[\|\mathbf{X}_k - \mathbf{M}_i\|^2 | \mathbf{X}_k \in \omega_i]$$

可证： $\overline{D_i^2} = 2\overline{D_{w,i}^2}$  ——有2倍关系，因为前者有重复计算。

3) 两个类内距离的计算式。

$$\overline{D_{w,i}^2} \triangleq E[\| \mathbf{X}_k - \mathbf{M}_i \|^2 | \omega_i] = \frac{1}{n_i} \sum_{k=1}^{n_i} (\mathbf{X}_k^i - \mathbf{M}_i)^T (\mathbf{X}_k^i - \mathbf{M}_i)$$

$$\overline{D_w^2} \triangleq \sum_{i=1}^c P(\omega_i) \overline{D_{w,i}^2} = \sum_{i=1}^c P(\omega_i) \left[ \frac{1}{n_i} \sum_{k=1}^{n_i} (\mathbf{X}_k^i - \mathbf{M}_i)^T (\mathbf{X}_k^i - \mathbf{M}_i) \right]$$

$$\overline{D_i^2} \triangleq E[(\mathbf{X}_k - \mathbf{X}_l)^T (\mathbf{X}_k - \mathbf{X}_l) | \omega_i] = \frac{1}{n_i(n_i - 1)} \sum_{k=1}^{n_i} \sum_{\substack{l=1 \\ l \neq k}}^{n_i} (\mathbf{X}_k^i - \mathbf{X}_l^i)^T (\mathbf{X}_k^i - \mathbf{X}_l^i)$$

4) 类内散布矩阵：表示各样本点围绕类均值的散布情况，以及各特征间互相关情况——即第*i*类分布的协方差矩阵。

$$\mathbf{S}_{w,i} \triangleq E[(\mathbf{X} - \mathbf{M}_i)(\mathbf{X} - \mathbf{M}_i)^T | \omega_i]$$

$$= \frac{1}{n_i} \sum_{k=1}^{n_i} (\mathbf{X}_k^i - \mathbf{M}_i)(\mathbf{X}_k^i - \mathbf{M}_i)^T \quad \text{显然: } \overline{D_{w,i}^2} = \text{Tr}(\mathbf{S}_{w,i})$$

特征选择和提取的结果应使类内散布矩阵的迹愈小愈好。

## 2. 类间散布距离和类间散布矩阵

- 1) **类间距离一**：各模式类均值向量之间**距离平方的加权和——类间均方两两距离**，记为  $\overline{D^2}$ 。

$$\overline{D^2} \triangleq \frac{1}{2} \sum_{i=1}^c P(\omega_i) \sum_{j=1}^c P(\omega_j) \| \mathbf{M}_i - \mathbf{M}_j \|^2 \quad \text{可证:}$$

$$\overline{D^2} = \sum_{i=1}^c P(\omega_i) \| \mathbf{M}_i - \mathbf{M}_0 \|^2 = \sum_{i=1}^c P(\omega_i) (\mathbf{M}_i - \mathbf{M}_0)^T (\mathbf{M}_i - \mathbf{M}_0) \triangleq \overline{D_b^2}$$

- 2) **类间距离二**：各模式类均值向量与总体均值向量之间**距离平方的加权和——类间均方散布距离**，记为  $D_b^2$ （见上式）。

式中，  $P(\omega_i)$ ：  $\omega_i$  类的先验概率；  $\mathbf{M}_i$ ：  $\omega_i$  类的均值向量；

$\mathbf{M}_0$ ：所有  $c$  类模式的总体均值向量。

$$\mathbf{M}_0 \triangleq E[\mathbf{X}] = E_{\omega} E[\mathbf{X} / \omega = \omega_i] = \sum_{i=1}^c P(\omega_i) \mathbf{M}_i$$

- 3) 类间散布矩阵：表示  $c$  类模式在空间的散布情况，记为  $\mathbf{S}_b$ 。

$$\mathbf{S}_b \triangleq \sum_{i=1}^c P(\omega_i) (\mathbf{M}_i - \mathbf{M}_0) (\mathbf{M}_i - \mathbf{M}_0)^T$$

注意：与类间距离的转置位置不同。

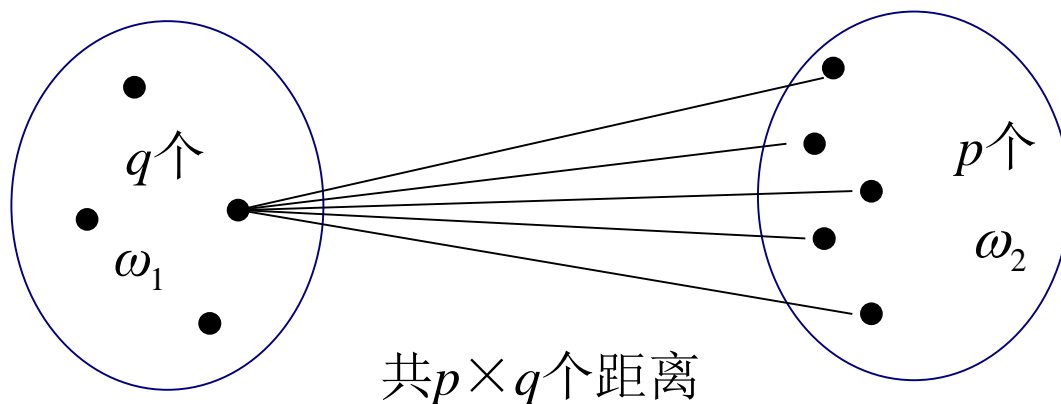
- 4) 类间散布距离与类间散布矩阵的关系： $\overline{D_b^2} = \text{Tr}(\mathbf{S}_b)$

类间散布矩阵的迹愈大愈有利于分类。

### 3. 多类模式向量间的总体距离和总体散布矩阵

#### 1) 两类情况的距离

设  $\omega_1$  类中有  $q$  个样本， $\omega_2$  类中有  $p$  个样本。



两个类之间的距离 =  $p \times q$  个点间距离的平均值

类似地  $\downarrow$  多类情况

多类间任意 **两点间两两距离的平均值**

$\downarrow$   
多类间任意 **两点间两两距离平方的平均值**

## 2) 多类情况的总体距离

任意类的组合

特定两类间  
任意样本的组合

(1) 多类模式向量间的总体平均距离  $J_d$

$$J_d \triangleq \frac{1}{2} \sum_{i=1}^c P(\omega_i) \sum_{j=1}^c P(\omega_j) \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} D^2(\mathbf{X}_k^i, \mathbf{X}_l^j) \quad (5-8)$$

式中,  $P(\omega_i)$  和  $P(\omega_j)$ :  $\omega_i$  和  $\omega_j$  类先验概率;  $c$ : 类别数;

$\mathbf{X}_k^i$ :  $\omega_i$  类的第  $k$  个样本;  $\mathbf{X}_l^j$ :  $\omega_j$  类的第  $l$  个样本;

$n_i$  和  $n_j$ :  $\omega_i$  和  $\omega_j$  类的样本数;

$D^2(\mathbf{X}_k^i, \mathbf{X}_l^j)$ :  $\mathbf{X}_k^i$  和  $\mathbf{X}_l^j$  间欧氏距离的平方。

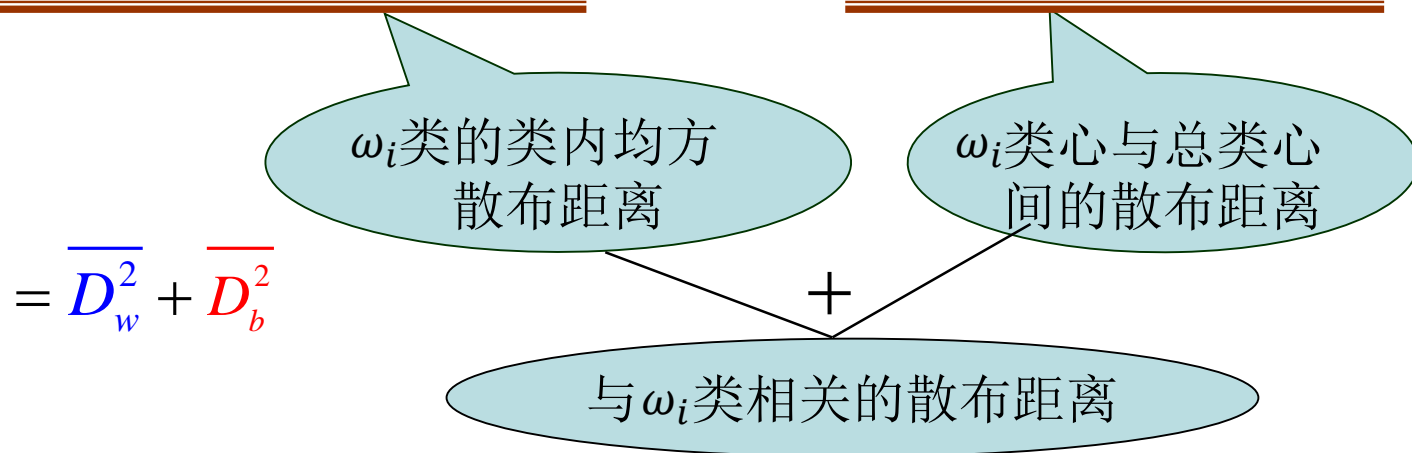
(2)  $J_d$  的另一种形式: 将以下三式代入(5-8)式

距离取欧式距离:  $D^2(\mathbf{X}_k^i, \mathbf{X}_l^j) = (\mathbf{X}_k^i - \mathbf{X}_l^j)^T (\mathbf{X}_k^i - \mathbf{X}_l^j) \quad (5-9)$

$\omega_i$  类的均值向量:  $\mathbf{M}_i \triangleq \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbf{X}_k^i \quad (5-10)$

所有类的总体均值向量:  $\mathbf{M}_0 \triangleq \sum_{i=1}^c P(\omega_i) \mathbf{M}_i \quad (5-11)$

$$J_d = \sum_{i=1}^c P(\omega_i) \left[ \frac{1}{n_i} \sum_{k=1}^{n_i} (\mathbf{X}_k^i - \mathbf{M}_i)^T (\mathbf{X}_k^i - \mathbf{M}_i) \right] + \sum_{i=1}^c P(\omega_i) \left[ (\mathbf{M}_i - \mathbf{M}_0)^T (\mathbf{M}_i - \mathbf{M}_0) \right]$$



多类模式向量之间的总均方距离=各类均方距离的先验概率加权和

多类模式向量之间的总体均方距离= { 模式类内散布距离加权和  
+  
模式类间散布距离加权和

### 3) 多类情况的散布矩阵

多类类间散布矩阵:

$$S_b \triangleq \sum_{i=1}^c P(\omega_i) (\mathbf{M}_i - \mathbf{M}_0) (\mathbf{M}_i - \mathbf{M}_0)^T$$

$\omega_i$ 类内散布矩阵:  $S_{w,i} \triangleq E[(X - M_i)(X - M_i)^T | \omega_i]$

$$= \frac{1}{n_i} \sum_{k=1}^{n_i} (X_k^i - M_i)(X_k^i - M_i)^T \quad \text{【计算式】}$$

多类类内散布矩阵:

$$S_w \triangleq \sum_{i=1}^c P(\omega_i) S_{w,i} = \sum_{i=1}^c P(\omega_i) E[(X - M_i)(X - M_i)^T | \omega_i]$$

$$= \sum_{i=1}^c P(\omega_i) \frac{1}{n_i} \sum_{k=1}^{n_i} (X_k^i - M_i)(X_k^i - M_i)^T \quad \text{【计算式】}$$

—— 各类模式协方差矩阵的先验加权平均值。

多类总体散布矩阵:

$$S_t \triangleq E[(X - M_0)(X - M_0)^T] = S_b + S_w$$

4) 多类模式的总体均方距离 $J_d$ 与总体散布矩阵 $S_t$ 的关系

$$J_d = \text{tr}(S_t) = \text{tr}(S_b + S_w)$$



## 均方距离与散布矩阵作为可分性测度的特点:

- \* 计算方便, 概念直观 (反映模式的空间分布情况) ;
- \* 与分类错误率没有直接的联系。



$$\begin{aligned} P_1(e) &= \int_{R_2} p(\mathbf{X} | \omega_1) d\mathbf{X} & P_2(e) &= \int_{R_1} p(\mathbf{X} | \omega_2) d\mathbf{X} \\ P(e) &= P(\omega_1)P_1(e) + P(\omega_2)P_2(e) \end{aligned}$$

## 5.2.2 基于概率分布的可分性测度

### 1. 散度

#### 1) 散度的定义

出发点: 对数似然比 含有 类别的可分性信息。

设  $\omega_i, \omega_j$  类的概率密度函数分别为  $p(\mathbf{X} | \omega_i)$  和  $p(\mathbf{X} | \omega_j)$

$\omega_i$  类对  $\omega_j$  类的对数似然比:  $l_{ij} \triangleq \ln \frac{p(\mathbf{X} | \omega_i)}{p(\mathbf{X} | \omega_j)}$

$$\omega_j \text{ 类对 } \omega_i \text{ 类的对数似然比: } l_{ji} = \ln \frac{p(\mathbf{X}|\omega_j)}{p(\mathbf{X}|\omega_i)}$$

对不同的 $\mathbf{X}$ ，似然函数不同，对数似然比体现的可分性不同，通常采用平均可分性信息——对数似然比的期望值。

$\omega_i$  类对数似然比的期望值:

$$E\{x\} = \int_{-\infty}^{\infty} xp(x)d(x)$$

$$I_{ij} = E[l_{ij}] = \int_{\mathbf{X}} p(\mathbf{X}|\omega_i) \ln \frac{p(\mathbf{X}|\omega_i)}{p(\mathbf{X}|\omega_j)} d\mathbf{X}$$

$\omega_j$  类对数似然比的期望值:

$$I_{ji} = E[l_{ji}] = \int_{\mathbf{X}} p(\mathbf{X}|\omega_j) \ln \frac{p(\mathbf{X}|\omega_j)}{p(\mathbf{X}|\omega_i)} d\mathbf{X}$$

散度等于两类间对数似然比期望值之和。

$\omega_i$  类对  $\omega_j$  类的散度定义为  $J_{ij}$ :

$$J_{ij} = I_{ij} + I_{ji} = \int_{\mathbf{X}} [p(\mathbf{X}|\omega_i) - p(\mathbf{X}|\omega_j)] \ln \frac{p(\mathbf{X}|\omega_i)}{p(\mathbf{X}|\omega_j)} d\mathbf{X}$$

散度表示了区分 $\omega_i$ 类和 $\omega_j$ 类的总的平均信息。

——特征选择和特征提取应使散度尽可能的太。

## 2) 散度的性质

(1)  $J_{ij} = J_{ji}$

$$J_{ij} = I_{ij} + I_{ji} = \int_X [p(\mathbf{X}|\omega_i) - p(\mathbf{X}|\omega_j)] \ln \frac{p(\mathbf{X}|\omega_i)}{p(\mathbf{X}|\omega_j)} d\mathbf{X}$$

$$J_{ji} = I_{ji} + I_{ij} = \int_X [p(\mathbf{X}|\omega_j) - p(\mathbf{X}|\omega_i)] \ln \frac{p(\mathbf{X}|\omega_j)}{p(\mathbf{X}|\omega_i)} d\mathbf{X}$$

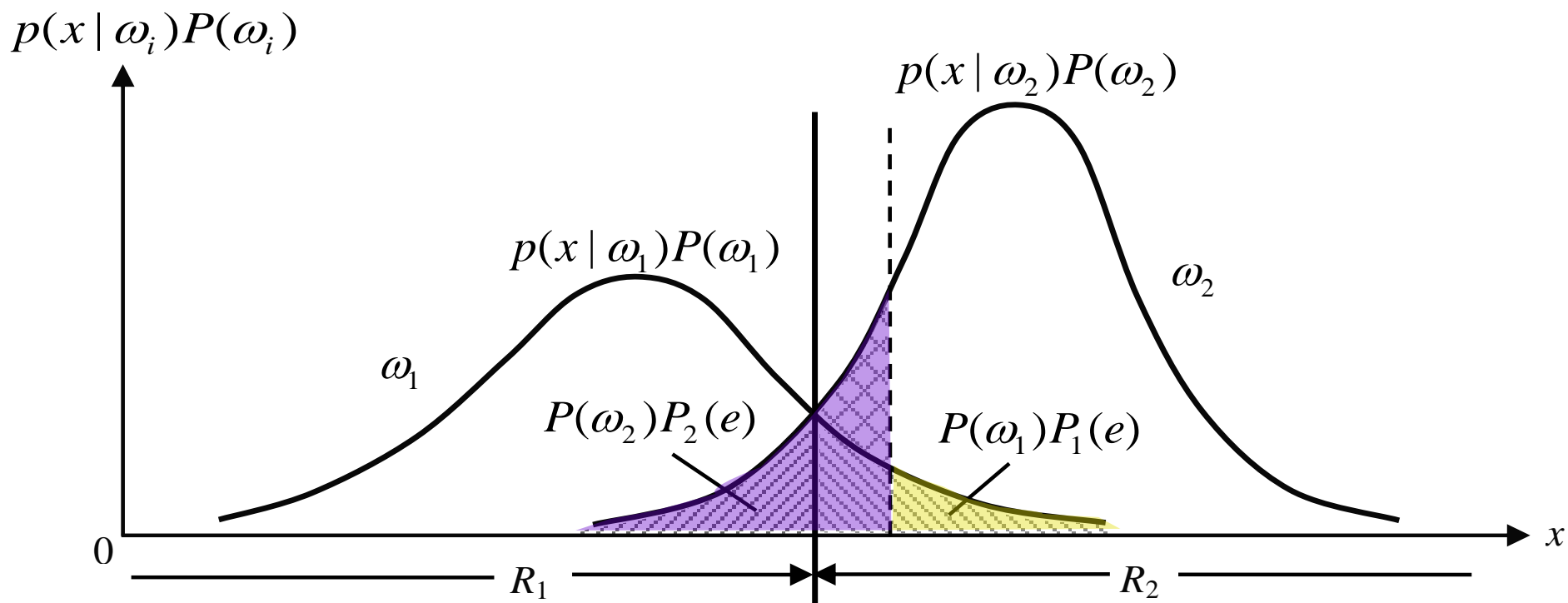
(2)  $J_{ij}$  为非负, 即  $J_{ij} \geq 0$ 。

当  $p(\mathbf{X} | \omega_i) \neq p(\mathbf{X} | \omega_j)$  时,  $J_{ij} > 0$ ,

$p(\mathbf{X} | \omega_i)$  与  $p(\mathbf{X} | \omega_j)$  相差愈大,  $J_{ij}$  越大。

当  $p(\mathbf{X} | \omega_i) = p(\mathbf{X} | \omega_j)$ , 两类分布密度相同,  $J_{ij} = 0$ 。

(3) 错误率分析中，两类概率密度曲线交叠越少，错误率越小。



由散度的定义式  $J_{ij} = I_{ij} + I_{ji} = \int_X [p(\mathbf{X}|\omega_i) - p(\mathbf{X}|\omega_j)] \ln \frac{p(\mathbf{X}|\omega_i)}{p(\mathbf{X}|\omega_j)} d\mathbf{X}$

可知，散度愈大，两类概率密度函数曲线相差愈大，交叠愈少，分类错误率愈小。

(4) 散度具有独立可加性：对于模式向量  $\mathbf{X} = [x_1, x_2, \dots, x_n]^T$ ，若各分量相互独立，则有【大家可以试证之】

$$J_{ij}(\mathbf{X}) = J_{ij}(x_1, x_2, \dots, x_n) = \sum_{k=1}^n J_{ij}(x_k)$$

据此可估计每一个特征在分类中的重要性：

散度较大的特征含有较大的可分信息——保留。

(5) 独立可加性：增加独立新特征使散度减小（意味什么？）

$$J_{ij}(x_1, x_2, \dots, x_n) \leq J_{ij}(x_1, x_2, \dots, x_n, x_{n+1})$$

### 3) 两个等方差正态分布模式类的散度

设 $\omega_i$ 类和 $\omega_j$ 类的概率密度函数分别为

$$p(\mathbf{X}|\omega_i) \sim N(\mathbf{M}_i, \mathbf{C})$$

$$p(\mathbf{X}|\omega_j) \sim N(\mathbf{M}_j, \mathbf{C})$$

可得到  $\omega_i$  类对  $\omega_j$  类的散度为

$$J_{ij} = \mathbf{Tr}[(\mathbf{C}^{-1}(\mathbf{M}_i - \mathbf{M}_j)(\mathbf{M}_i - \mathbf{M}_j)^{\mathbf{T}})] = (\mathbf{M}_i - \mathbf{M}_j)^{\mathbf{T}} \mathbf{C}^{-1}(\mathbf{M}_i - \mathbf{M}_j)$$

——两模式类均值向量间的**马氏距离的平方**

一维正态分布时：

$$J_{ij} = \frac{(m_i - m_j)^2}{\sigma^2}$$

两类均值向量间  
距离越远，散度愈大

每类的类内向量  
分布愈集中（方差越  
小），散度愈大

## 5.3 基于类内散布矩阵的单类模式特征提取

### 特征提取的目的:

对某类模式: 压缩模式向量的维数。

对多类分类: 压缩模式向量的维数;

保留类别间的鉴别信息, 突出可分性。

### 特征提取方法:

若  $\{X \in \omega_i\}$  是  $\omega_i$  类的一个  $n$  维样本集, 将  $X$  压缩成  $m$  维

向量  $X^*$  —— 寻找一个  $m \times n$  矩阵  $A$ , 并作变换:

$$\begin{array}{ccccc} & & X^* = AX & & \\ & \swarrow & | & \searrow & \\ m \times 1 & & m \times n & & n \times 1 \end{array} \quad (m < n)$$

注意: 维数降低后, 在新的  $m$  维空间里各模式类之间的分布规律应至少保持不变或更优化。

讨论:

- \* 根据类内散布矩阵如何确定变换矩阵 $\mathbf{A}$ ;
- \* 通过 $\mathbf{A}$ 如何进行特征提取。

## 1. 根据类内散布矩阵确定变换矩阵

设 $\omega_i$ 类模式的均值向量为 $\mathbf{M}$ , 类内散布矩阵(协方差矩阵)为 $\mathbf{C}$ :

$$\mathbf{M} = E\{\mathbf{X}\}$$

$$\mathbf{C} = E\{(\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})^T\}$$

式中, $\mathbf{X}$ 为 $n$ 维向量, $\mathbf{C}$ 为 $n \times n$ 的实对称矩阵。

设矩阵 $\mathbf{C}$ 的 $n$ 个特征值分别为 $\lambda_1, \lambda_2, \dots, \lambda_n$ 。

任一特征值是满足

$$|\lambda \mathbf{I} - \mathbf{C}| = 0$$

的一个解。



假定  $n$  个特征值对应的  $n$  个特征向量为  $\mathbf{u}_k$ ,  $k = 1, 2, \dots, n$ 。

则  $\mathbf{u}_k$  是满足

$$\mathbf{C}\mathbf{u}_k = \lambda_k \mathbf{u}_k$$

的一个非零解。

$\mathbf{u}_k$  是  $n$  维向量, 可表示为  $\mathbf{u}_k = [u_{k1}, u_{k2}, \dots, u_{kn}]^T$ 。

若  $\mathbf{u}_k$  为归一化特征向量, 根据实对称矩阵的性质, 有

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}$$

——  $n$  个特征向量相互正交。

若选  $n$  个归一化特征向量作为  $(\mathbf{A})_{n \times n}$  的行, 则  $\mathbf{A}$  为归一化正交矩阵:

$$\mathbf{A} = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \dots \\ \mathbf{u}_n^T \end{bmatrix}$$

$$\mathbf{A}^T = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_n]$$

$$\mathbf{A}\mathbf{A}^T = \mathbf{I}$$

利用  $\mathbf{A}$  对  $\omega_i$  类的样本  $\mathbf{X}$  进行变换, 得  $\mathbf{X}^* = \mathbf{A}\mathbf{X}$ 。

式中,  $\mathbf{X}$  和  $\mathbf{X}^*$  都是  $n$  维向量。

$\mathbf{A}_{n \times n}$

考察变换前后的分布规律:

均值向量  $\mathbf{M}^*$ 、协方差矩阵  $\mathbf{C}^*$  和类内距离  $\overline{D^2}$  的变化。

$$(1) \mathbf{M}^* = E\{\mathbf{X}^*\} = E\{\mathbf{A}\mathbf{X}\} = \mathbf{A}E\{\mathbf{X}\} = \mathbf{A}\mathbf{M}$$

$$(2) \mathbf{C}^* = E\{(\mathbf{X}^* - \mathbf{M}^*)(\mathbf{X}^* - \mathbf{M}^*)^T\} = E\{(\mathbf{A}\mathbf{X} - \mathbf{A}\mathbf{M})(\mathbf{A}\mathbf{X} - \mathbf{A}\mathbf{M})^T\} \\ = \mathbf{A}E\{(\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})^T\}\mathbf{A}^T = \mathbf{A}\mathbf{C}\mathbf{A}^T$$

$$= \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \dots \\ \mathbf{u}_n^T \end{bmatrix} \mathbf{C} [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_n] = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \dots \\ \mathbf{u}_n^T \end{bmatrix} [\lambda_1 \mathbf{u}_1 \ \lambda_2 \mathbf{u}_2 \ \dots \ \lambda_n \mathbf{u}_n] = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}$$

$$\mathbf{C}\mathbf{u}_k = \lambda_k \mathbf{u}_k$$

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}$$

变换后：

协方差矩阵为对角阵，说明  $\mathbf{X}^*$  的各分量不相关的！

——便于特征的取舍；

$\mathbf{X}^*$  的第  $k$  个分量的方差等于未变换时  $\mathbf{C}$  的特征值  $\lambda_k$ 。

(3) 变换后的类内均方距离

$$\begin{aligned}\overline{D^2} &= E\{\|\mathbf{X}_i^* - \mathbf{X}_j^*\|^2\} \\&= E\{(\mathbf{X}_i^* - \mathbf{X}_j^*)^T (\mathbf{X}_i^* - \mathbf{X}_j^*)\} \\&= E\{(\mathbf{A}\mathbf{X}_i - \mathbf{A}\mathbf{X}_j)^T (\mathbf{A}\mathbf{X}_i - \mathbf{A}\mathbf{X}_j)\} \\&= E\{(\mathbf{X}_i - \mathbf{X}_j)^T \mathbf{A}^T \mathbf{A} (\mathbf{X}_i - \mathbf{X}_j)\} \\&= E\{(\mathbf{X}_i - \mathbf{X}_j)^T (\mathbf{X}_i - \mathbf{X}_j)\} \\&= E\{\|\mathbf{X}_i - \mathbf{X}_j\|^2\}\end{aligned}$$

$$\mathbf{C}^* = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}$$

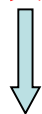
变换后的类内距离保持不变。注意 $\mathbf{A}$ 是 $n \times n$ ，不是前面的 $m \times n$ 。

根据以上特点得到构造变换矩阵的方法：

目标：构造一变换矩阵，可以将 $n$ 维向量 $\mathbf{X}$ 变换成 $m$ 维（ $m < n$ ）。

**思路：**

将变换前的 $\mathbf{C}$ 的 $n$ 个特征值从小到大排队



选择前 $m$ 个小的特征值对应的特征向量  
作为矩阵 $\mathbf{A}$ 的行（ $m \times n$ ）

后↓续

对 $\mathbf{X}$ 进行 $\mathbf{A}$ 变换

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} \Rightarrow \begin{bmatrix} x_1^* \\ \mathbf{M} \\ x_m^* \end{bmatrix}$$

优点：压缩了维数；

**类内距离减小，样本更密集——适合分类，不适合逼近！**  
——相当去掉了方差大的特征分量。

## 2. 特征提取的方法

设 $\{\mathbf{X}\}$ 为 $\omega_i$ 类的样本集， $\mathbf{X}$ 为 $n$ 维向量。

第一步：根据样本集求 $\omega_i$ 类的协方差矩阵（类内散布矩阵）。

$$\mathbf{C} = \frac{1}{N} \sum_{i=1}^N (\mathbf{X}_i - \mathbf{M})(\mathbf{X}_i - \mathbf{M})^T$$

其中，

$$\mathbf{M} = \frac{1}{N} \sum_{i=1}^N \mathbf{X}_i$$

第二步：计算 $\mathbf{C}$ 的特征值，对特征值从小到大进行排队，选择前 $m$ 个。

第三步：计算前 $m$ 个特征值对应的特征向量

$\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m$ ，并归一化处理。将归一化后的

特征向量的转置作为矩阵 $\mathbf{A}$ 的行。

$$\mathbf{A} = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_m^T \end{bmatrix}$$

第四步：利用 $\mathbf{A}$ 对样本集 $\{\mathbf{X}\}$ 进行变换。

$$\mathbf{X}^* = \mathbf{A}\mathbf{X}$$

则 $m$ 维（ $m < n$ ）模式向量 $\mathbf{X}^*$ 就是作为分类用的模式向量。

**例 5.2** 假定 $\omega_i$ 类的样本集为 $\{\mathbf{X}\} = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ ，三个样本分别为

$$\mathbf{X}_1 = [1, 1]^T, \quad \mathbf{X}_2 = [2, 2]^T, \quad \mathbf{X}_3 = [3, 1]^T$$

用类内散布矩阵进行特征提取，将二维样本变换成一维样本。

解：1) 求样本均值向量和协方差矩阵。

$$\mathbf{M} = \frac{1}{3} \sum_{i=1}^3 \mathbf{X}_i = [2, 1.3]^T$$

$$\mathbf{C} = \frac{1}{3} \sum_{i=1}^3 \mathbf{X}_i \mathbf{X}_i^T - \mathbf{M} \mathbf{M}^T = \begin{bmatrix} 0.7 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}$$

第二步：根据 $|\lambda \mathbf{I} - \mathbf{C}| = 0$ 求 $\mathbf{C}$ 的特征值，并进行选择。

由 
$$\begin{vmatrix} \lambda - 0.7 & -0.1 \\ -0.1 & \lambda - 0.3 \end{vmatrix} = 0$$

$$\mathbf{C} = \begin{bmatrix} 0.7 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}$$

得  $\lambda_1 = 0.2765 \quad \lambda_2 = 0.7236$   
 $\because \lambda_1 < \lambda_2 \quad \therefore \text{选 } \lambda_1$

第三步：计算 $\lambda_1$ 对应的特征向量 $\mathbf{u}_1$ 。由方程 $\mathbf{C}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$ 得

$$\mathbf{u}_1 = [0.5, -2.1]^T$$

归一化处理有

$$\mathbf{u}_1 = \frac{1}{\sqrt{0.5^2 + 2.1^2}} [0.5, -2.1]^T = \frac{1}{\sqrt{4.66}} [0.5, -2.1]^T$$

由归一化特征向量 $\mathbf{u}_1$ 构成变换矩阵 $\mathbf{A}$ ： $\mathbf{A} = \frac{1}{\sqrt{4.66}} [0.5, -2.1]$

第四步：利用  $A$  对  $X_1, X_2, X_3$  进行变换。

$$X_1^* = AX_1 = -0.74$$

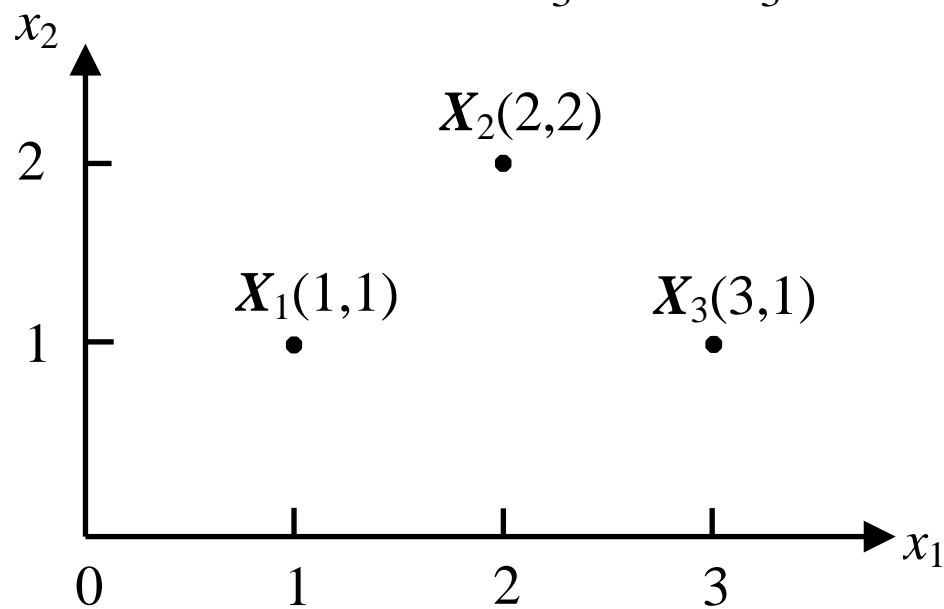
$$X_2^* = AX_2 = -1.48$$

$$X_3^* = AX_3 = -0.28$$

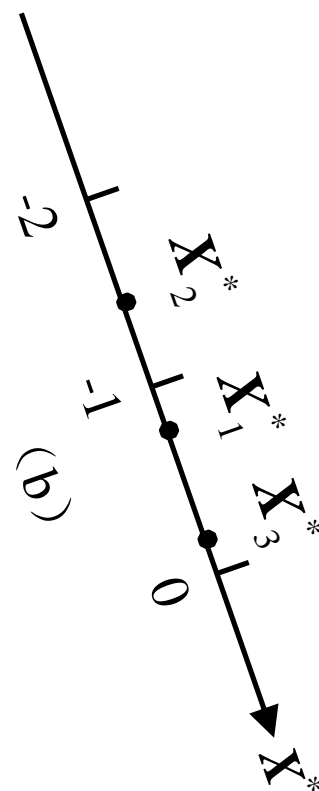
$$X_1 = [1, 1]^T$$

$$X_2 = [2, 2]^T$$

$$X_3 = [3, 1]^T$$



(a)  
变换前



变换后



## 5.4 基于K-L变换的多类模式特征提取

---

### 特征提取的目的：

对一类模式：维数压缩。

对多类模式：维数压缩，突出类别的可分性。

### 卡洛南-洛伊（Karhunen-Loeve）变换（K-L变换）：

- \* 一种常用的特征提取方法；
- \* 最小均方误差逼近意义下的最优正交变换；
- \* 适用于任意的概率密度函数；
- \* 在消除模式特征之间的相关性、突出差异性方面有最优的效果。

分为：连续K-L变换    离散K-L变换

## 1. K-L展开式

设 $\{\mathbf{X}\}$ 是  $n$  维随机模式向量  $\mathbf{X}$  的集合, 对每一个  $\mathbf{X}$  可以用确定的完备归一化正交向量系  $\{\mathbf{u}_j\}$  中的正交向量**展开式**:

$$\mathbf{X} = \sum_{j=1}^n a_j \mathbf{u}_j \quad a_j: \text{随机系数};$$

用有限项 ( $d < n$ ) 估计  $\mathbf{X}$  时:  $\hat{\mathbf{X}} = \sum_{j=1}^d a_j \mathbf{u}_j$

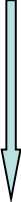
引起的均方误差:  $\xi = E[(\mathbf{X} - \hat{\mathbf{X}})^T (\mathbf{X} - \hat{\mathbf{X}})]$

代入  $\mathbf{X}$ 、 $\hat{\mathbf{X}}$ , 利用  $\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}$



$$\xi = E\left[\sum_{j=d+1}^n a_j^2\right]$$

$$\xi = E\left[\sum_{j=d+1}^n a_j^2\right]$$

由  $\mathbf{X} = \sum_{j=1}^n a_j \mathbf{u}_j$  两边  左乘  $\mathbf{u}_j^T$  得  $a_j = \mathbf{u}_j^T \mathbf{X}$ 。

$$\xi = E\left[\sum_{j=d+1}^n \mathbf{u}_j^T \mathbf{X} \mathbf{X}^T \mathbf{u}_j\right]$$

$$= \sum_{j=d+1}^n \mathbf{u}_j^T E[\mathbf{X} \mathbf{X}^T] \mathbf{u}_j$$

$\mathbf{u}_j$  为确定性向量

$$= \sum_{j=d+1}^n \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j \quad \mathbf{R}: \text{自相关矩阵。}$$

不同的  $\{\mathbf{u}_j\}$  对应不同的均方误差， $\mathbf{u}_j$  的选择应使  $\xi$  最小。

利用拉格朗日乘数法求使  $\xi$  最小的正交系  $\{\mathbf{u}_j\}$ ，令

$$g(\mathbf{u}_j) = \sum_{j=d+1}^n \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j - \sum_{j=d+1}^n \lambda_j (\mathbf{u}_j^T \mathbf{u}_j - 1) \quad \lambda_j: \text{拉格朗日乘数}$$

$$g(\mathbf{u}_j) = \sum_{j=d+1}^n \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j - \sum_{j=d+1}^n \lambda_j (\mathbf{u}_j^T \mathbf{u}_j - 1)$$

用函数  $g(\mathbf{u}_j)$  对  $\mathbf{u}_j$  求导，并令导数为零，得

$$(\mathbf{R} - \lambda_j \mathbf{I}) \mathbf{u}_j = 0 \quad j = d+1, \dots, n$$

——正是矩阵  $\mathbf{R}$  与其特征值和对应特征向量的关系式。

说明：当用 $\mathbf{X}$ 的自相关矩阵 $\mathbf{R}$ 的特征值对应的特征向量展开 $\mathbf{X}$ 时，截断误差最小。

选前 $d$ 项估计 $\mathbf{X}$ 时引起的均方误差为

$$\xi = \sum_{j=d+1}^n \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j = \sum_{j=d+1}^n \text{Tr}[\mathbf{u}_j \mathbf{R} \mathbf{u}_j^T] = \sum_{j=d+1}^n \lambda_j$$

$\lambda_j$  决定截断的均方误差， $\lambda_j$  的值小，那么  $\xi$  也小。

因此，当用 $\mathbf{X}$ 的正交展开式中的前 $d$ 项估计 $\mathbf{X}$ 时，展开式中的 $\mathbf{u}_j$ 应当是前 $d$ 个较大的特征值对应的特征向量。

## K-L变换具体方法:

对 $R$ 的特征值由大到小进行排队:  $\lambda_1 \geq \lambda_2 \geq \Lambda \geq \lambda_d \geq \lambda_{d+1} \geq \Lambda$

均方误差最小的 $\mathbf{X}$ 的近似式:  $\mathbf{X} = \sum_{j=1}^d a_j \mathbf{u}_j$  —— K-L展开

矩阵形式:  $\mathbf{X} = \mathbf{U}\mathbf{a}$  (5-49)

式中,  $\mathbf{a} = [a_1, a_2, \dots, a_d]^T$ ,  $\mathbf{U}_{n \times d} = [\mathbf{u}_1, \dots, \mathbf{u}_j, \dots, \mathbf{u}_d]$ 。

其中:  $\mathbf{u}_j = [u_{j1}, u_{j2}, \dots, u_{jn}]^T$

$$\mathbf{U}^T \mathbf{U} = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \dots \\ \mathbf{u}_d^T \end{bmatrix} [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_d] = \mathbf{I}$$

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}$$

对式(5-49)两边左乘 $\mathbf{U}^T$ :  $\mathbf{a} = \mathbf{U}^T \mathbf{X}$  —— K-L变换

系数向量 $\mathbf{a}$ 就是变换后的模式向量。

## 2. 利用自相关矩阵的K-L变换进行特征提取

设  $\mathbf{X}$  是  $n$  维模式向量， $\{\mathbf{X}\}$  是来自  $M$  个模式类的样本集，总样本数目为  $N$ 。将  $\mathbf{X}$  变换为  $d$  维 ( $d < n$ ) 向量的方法：

第一步：求样本集 $\{\mathbf{X}\}$ 的总体自相关矩阵 $\mathbf{R}$ 。

$$\mathbf{R} = E[\mathbf{X}\mathbf{X}^T] \approx \frac{1}{N} \sum_{j=1}^N \mathbf{X}_j \mathbf{X}_j^T$$

决定压缩  
后的维数 $d$

第二步：求  $\mathbf{R}$  的特征值  $\lambda_j$ ， $j = 1, 2, \dots, n$ 。对特征值由大到小进行排队，选择前  $d$  个较大的特征值。

第三步：计算  $d$  个特征值对应的特征向量  $\mathbf{u}_j$ ， $j = 1, 2, \dots, d$ ，归一化后构成变换矩阵  $\mathbf{U}$ 。

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d]$$

第四步：对 $\{\mathbf{X}\}$ 中的每个  $\mathbf{X}$  进行 K-L 变换，得变换后向量  $\mathbf{X}^*$ ：

$$\mathbf{X}^* = \mathbf{U}^T \mathbf{X}$$

$d$  维向量  $\mathbf{X}^*$  就是代替  $n$  维向量  $\mathbf{X}$  进行分类的模式向量。

### 3. 使用不同散布矩阵进行K-L变换

根据不同的散布矩阵进行K-L变换，对保留分类鉴别信息的效果不同。

#### 1) 采用多类类内散布矩阵 $S_w$ 作 K-L 变换

多类类内散布矩阵：

$$S_w = \sum_{i=1}^c P(\omega_i) E[(X - M_i)(X - M_i)^T \mid X \in \omega_i]$$

若要突出各类模式的主要特征分量的分类作用：

选用对应于大特征值的特征向量组成变换矩阵；

若要使同一类模式聚集于最小的特征空间范围：

选用对应于小特征值的特征向量组成变换矩阵。

#### 2) 采用类间散布矩阵 $S_b$ 作 K-L 变换

类间散布矩阵：

$$S_b = \sum_{i=1}^c P(\omega_i) (M_i - M_0)(M_i - M_0)^T$$

适用于类间距离比类内距离大得多的多类问题，选择与大特征值对应的特征向量组成变换矩阵。

### 3) 采用总体散布矩阵 $S_t$ 作 K-L 变换

把多类模式合并起来看成一个总体分布。

总体散布矩阵:  $S_t = E[(X - M_0)(X - M_0)^T] = S_b + S_w$

适合于多类模式在总体分布上具有良好的可分性的情况。

采用大特征值对应的特征向量组成变换矩阵，能够保留模式原有分布的主要结构。

**利用K-L变换进行特征提取的优点:**

- 1) 在均方逼近误差最小的意义下使新样本集  $\{X^*\}$  逼近原样本集  $\{X\}$  的分布，既压缩了维数、又保留了数据集的分布信息和类别鉴别信息。



- 2) 变换后的新模式向量各分量相对总体均值的方差等于原样本集总体自相关矩阵的大特征值，表明变换加强了模式类之间的差异性。

$$\mathbf{C}^* = E\{(\mathbf{X}^* - \mathbf{M}^*)(\mathbf{X}^* - \mathbf{M}^*)^T\} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_d \end{bmatrix}$$

- 3)  $\mathbf{C}^*$ 为对角矩阵说明了变换后样本各分量特征互不相关，即消除了原分量特征间的相关性，便于进一步进行特征的选择。

### K-L变换的不足之处:

- 1) 对两类问题容易得到较满意的结果。类别愈多，效果愈差。
- 2) 需要通过足够多的样本估计样本集的协方差矩阵或其它类型的散布矩阵。当样本数不足时，矩阵的估计会变得十分粗略，变换的优越性也就不能充分地显示出来。

3) 矩阵的本征值和本征向量缺乏统一的快速算法，计算较困难。

**例5.3** 两个二维模式类的样本分别为 **【注意：第三步的计算!】**

$$\omega_1: \mathbf{X}_1 = [2, 2]^T, \mathbf{X}_2 = [2, 3]^T, \mathbf{X}_3 = [3, 3]^T$$

$$\omega_2: \mathbf{X}_4 = [-2, -2]^T, \mathbf{X}_5 = [-2, -3]^T, \mathbf{X}_6 = [-3, -3]^T$$

利用**总体自相关矩阵** $\mathbf{R}$ 作K-L变换，把原样本压缩成一维样本。

解：第一步：计算总体自相关矩阵 $\mathbf{R}$ 。

$$\mathbf{R} = E\{\mathbf{X}\mathbf{X}^T\} = \frac{1}{6} \sum_{j=1}^6 \mathbf{X}_j \mathbf{X}_j^T = \begin{bmatrix} 5.7 & 6.3 \\ 6.3 & 7.3 \end{bmatrix}$$

第二步：计算 $\mathbf{R}$ 的本征值，并选择较大者。由 $|\mathbf{R} - \lambda \mathbf{I}| = 0$ 得

$$\lambda_1 = 12.85, \lambda_2 = 0.15, \text{ 选择 } \lambda_1。$$

第三步：根据 $\mathbf{R}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$ 计算 $\lambda_1$ 对应的特征向量 $\mathbf{u}_1$ ，令第1分量为1，归一化后为：

$$\mathbf{u}_1 = \frac{1}{\sqrt{2.3}} [1, 1.14]^T = [0.66, 0.75]^T$$

$$\mathbf{u}_1 = [0.66, 0.75]^T$$

变换矩阵为  $\mathbf{U} = [\mathbf{u}_1] = \begin{bmatrix} 0.66 \\ 0.75 \end{bmatrix}$

第四步：利用  $\mathbf{U}$  对样本集中每个样本进行 K-L 变换。

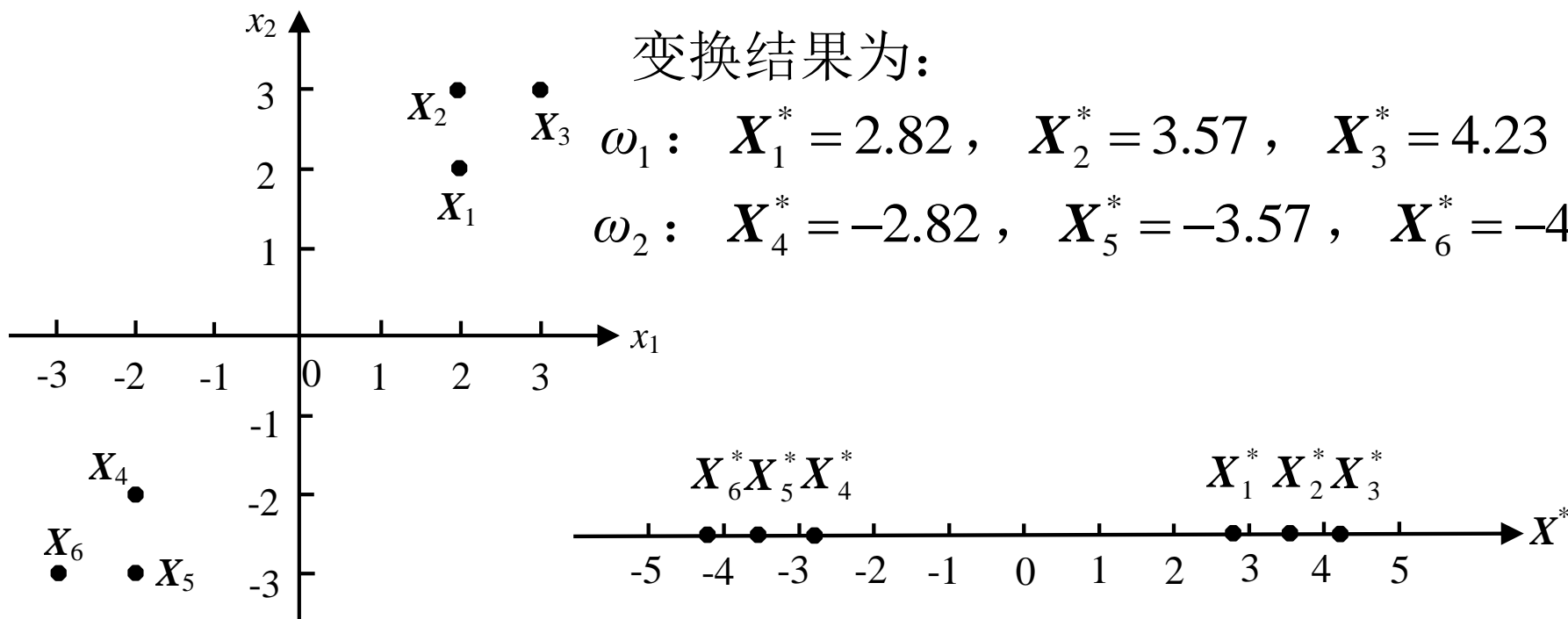
$$\mathbf{X}_1^* = \mathbf{U}^T \mathbf{X}_1 = [0.66 \ 0.75] \begin{bmatrix} 2 \\ 2 \end{bmatrix} = 2.82$$

.....

变换结果为：

$$\omega_1 : \mathbf{X}_1^* = 2.82, \mathbf{X}_2^* = 3.57, \mathbf{X}_3^* = 4.23$$

$$\omega_2 : \mathbf{X}_4^* = -2.82, \mathbf{X}_5^* = -3.57, \mathbf{X}_6^* = -4.23$$



## 5.5 特征选择

从 $n$ 个特征中选择 $d$ 个( $d < n$ )最优特征构成分类用特征向量。

### 5.5.1 特征选取择的准则

#### 1. 散布矩阵准则

类别可分性测度	特征选择准则
类间散布矩阵 $\mathbf{S}_b$ 多类类内散布矩阵 $\mathbf{S}_w$ 多类总体散布矩阵 $\mathbf{S}_t$	使 $\text{Tr}(\mathbf{S}_b)$ 最大 使 $\text{Tr}(\mathbf{S}_w)$ 最小
$J_1 = \text{Tr}(\mathbf{S}_w^{-1} \mathbf{S}_b)$ $J_2 = \frac{\text{Tr}(\mathbf{S}_b)}{\text{Tr}(\mathbf{S}_w)}$	使 $J_1 \sim J_4$ 最大
$J_3 = \ln \frac{ \mathbf{S}_b }{ \mathbf{S}_w }$ $J_4 = \frac{ \mathbf{S}_w + \mathbf{S}_b }{ \mathbf{S}_w }$	

## 2. 散度准则

用于正态分布的模式类。

两类的散度表达式

$$J_{ij} = \frac{1}{2} \text{Tr}[(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})(\mathbf{C}_i - \mathbf{C}_j)] + \frac{1}{2} \text{Tr}[(\mathbf{C}_i^{-1} + \mathbf{C}_j^{-1})(\mathbf{M}_i - \mathbf{M}_j)(\mathbf{M}_i - \mathbf{M}_j)^T]$$

书中符号有误

\* 平均散度

$$J = \sum_{i=1}^{c-1} \sum_{j=i+1}^c p(\omega_i)p(\omega_j)J_{ij} \quad \text{选择使} J \text{最大的特征子集}$$

\* 变换散度

$$J_{ij}^T = 100\% \times [1 - \exp(-J_{ij}/8)]$$

\* 平均变换散度

$$J^T = \sum_{i=1}^{c-1} \sum_{j=i+1}^c p(\omega_i)p(\omega_j)J_{ij}^T$$

## 5.5.2 特征选择的方法

从 $n$ 个特征中挑选 $d$ 个特征，所有可能的特征子集数为

$$C_n^d = \frac{n!}{(n-d)!d!}$$

组合数很大

穷举法：

计算出各种可能特征组合的某个测度值，加以比较，选择最优特征组。

特点：

可以得到最优特征组；

计算量大，难实现。采取搜索技术可降低计算量。

### 1. 最优搜索算法

分支定界算法：唯一能获得最优结果的搜索方法。

自上而下、具有回溯功能。

## 使用条件：

可分性测度  $J$  对维数单调。

## 方法：

- \* 将可能的特征组构成树结构。

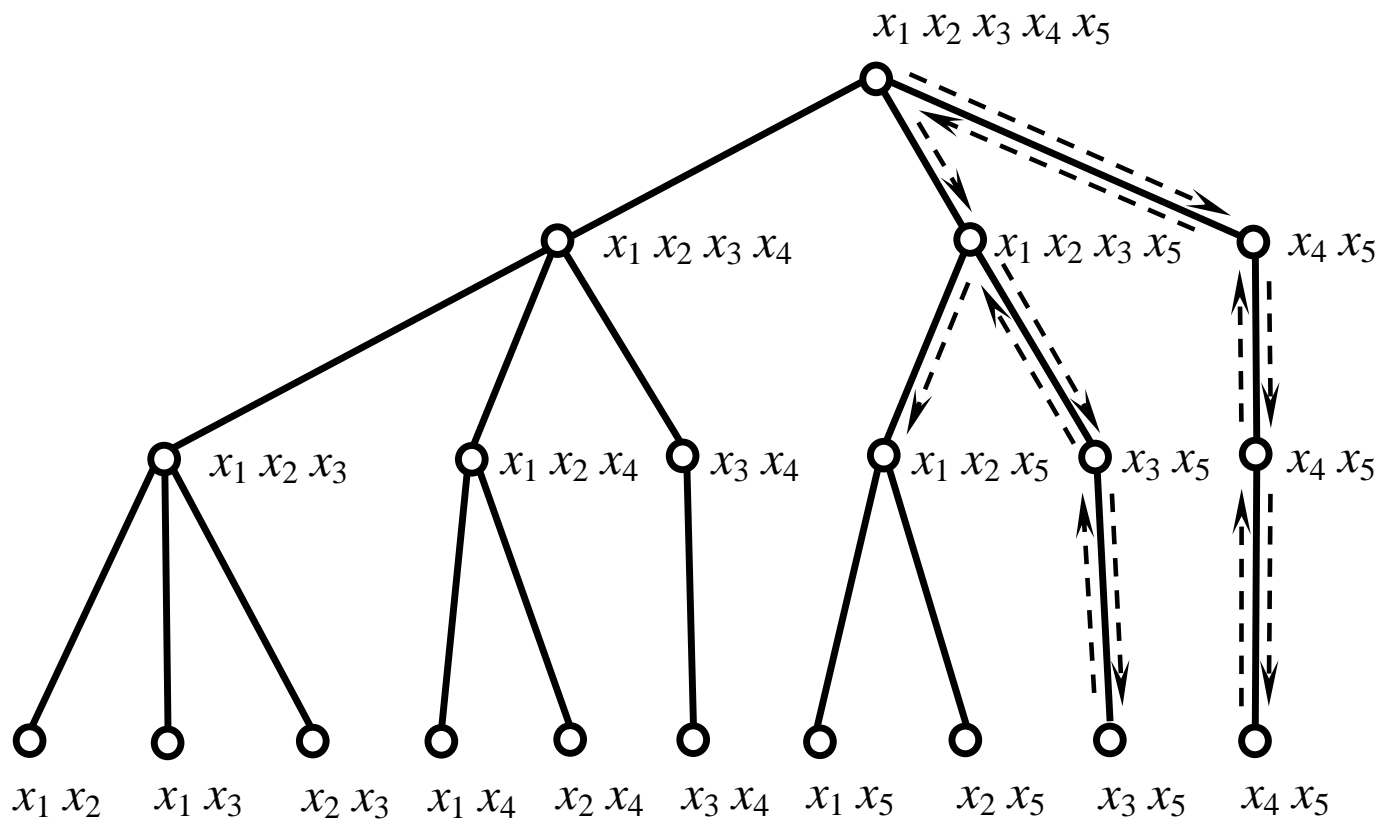
待选择的 $n$ 个原特征为根；

子结点的特征组元素个数逐级下降；

叶结点按照规定的特征数构成特征组合。

- \* 从最右边的叶结点开始，根据选择的测度回溯搜索。
- \* 找到最优特征组，结束。

例：从5个特征中选出2个特征作为模式向量。





## 2. 次优搜索算法

虽然不能得到最优解，但可减少计算量。

1) 单独最优特征组合

2) 顺序前进法 (Sequential Forward Selection, SFS)

广义顺序前进法 (Generalized SFS, GSFS)

3) 顺序后退法 (Sequential Backward Selection, SBS)

广义顺序后退法 (Generalized SBS, GSBS)

4) 增 $l$ 减 $r$ 法 ( $l-r$ 法) : SFS和SBS的组合。

广义的 $l-r$ 法

其他：模拟退火 (Simulated Annealing) 算法

Tabu搜索 (Tabu Search) 算法

遗传算法 (Genetic Algorithm)

## ■ **特征的选择与提取是模式识别中重要而困难的一步**

- 模式识别的第一步：分析各种特征的有效性并选出最有代表性的特征
- 降低特征维数在很多情况下是有效设计分类器的重要课题

## ■ **三大类特征：物理特征、结构特征、数学特征**

- 物理特征、结构特征：易于为人的直觉感知，但难于定量描述，因而不易用机器判别
- 数学特征：易于用机器定量描述和判别

# 证明题

【题1】定义：

$$\overline{D^2} @ \frac{1}{2} \sum_{i=1}^c P(\omega_i) \sum_{j=1}^c P(\omega_j) \| \mathbf{M}_i - \mathbf{M}_j \|^2$$

$$\overline{D_b^2} @ \sum_{i=1}^c P(\omega_i) (\mathbf{M}_i - \mathbf{M}_0)^T (\mathbf{M}_i - \mathbf{M}_0)$$

试证：  $\overline{D^2} = \overline{D_b^2}$

【题2】定义：

$$J_d @ \frac{1}{2} \sum_{i=1}^c P(\omega_i) \sum_{j=1}^c P(\omega_j) \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} (\mathbf{X}_k^i - \mathbf{X}_l^j)^T (\mathbf{X}_k^i - \mathbf{X}_l^j)$$

试证：

$$\begin{aligned} J_d &= \sum_{i=1}^c P(\omega_i) \left[ \frac{1}{n_i} \sum_{k=1}^{n_i} (\mathbf{X}_k^i - \mathbf{M}_i)^T (\mathbf{X}_k^i - \mathbf{M}_i) \right] \\ &\quad + \sum_{i=1}^c P(\omega_i) \left[ (\mathbf{M}_i - \mathbf{M}_0)^T (\mathbf{M}_i - \mathbf{M}_0) \right] \end{aligned}$$

# 课后作业

---

- 见另文。
- 下次上课前提交。
- 最好使用电子档。

**End of This Part**