



第五章

特征提取与选择

任课教师：柳欣老师

email: starxliu@163.com



5.1 特征提取与选择

模式识别的三大核心问题：

- 特征数据采集
- 分类识别
- 特征提取与选择

分类识别的正确率取决于对象的表示、训练学习和分类识别算法，我们在前面各章的介绍中详细讨论了后两方面的内容。本章介绍的特征提取与选择问题则是对象表示的一个关键问题。



5.1 特征提取与选择

通常在得到实际对象的若干具体特征之后，再由这些原始特征产生出对分类识别最有效、数目最少的特征，这就是特征提取与选择的任务。从本质上讲，我们的目的是使在最小维数特征空间中异类模式点相距较远（类间距离较大），而同类模式点相距较近（类内距离较小）。

一般来讲，不同类的模式可以被区分是由于它们所属类别在特征空间中的类域是不同的区域。显然，区域重叠的部分越小或完全没有重叠，类别的可分性就越好。因此可以用距离或离差测度（散度）来构造类别的可分性判据。



5.2 特征提取的必要性

- 特征选择和提取是构造模式识别系统时的一个重要课题
 - 在很多实际问题中，往往不容易找到那些最重要的特征，或受客观条件的限制，不能对它们进行有效的测量；
 - 因此在测量时，由于人们心理上的作用，只要条件许可总希望把特征取得多一些；
 - 另外，由于客观上的需要，为了突出某些有用信息，抑制无用信息，有意加上一些比值、指数或对数等组合计算特征；
 - 如果将数目很多的测量值不做分析，全部直接用作分类特征，不但耗时，而且会影响到分类的效果，产生“特征维数灾难”问题。

5.2 特征提取的必要性

- 为了设计出效果好的分类器，通常需要对原始测量值集合进行分析，经过选择或变换处理，组成有效的识别特征；
- 在保证一定分类精度的前提下，减少特征维数，即进行“降维”处理，使分类器实现快速、准确和高效的分类。
- 为达到上述目的，关键是所提供的识别特征应具有很好的可分性，使分类器容易判别。为此，需对特征进行选择。
 - 应去掉模棱两可、不易判别的特征；
 - 所提供的特征不要重复，即去掉那些相关性强且没有增加更多分类信息的特征。

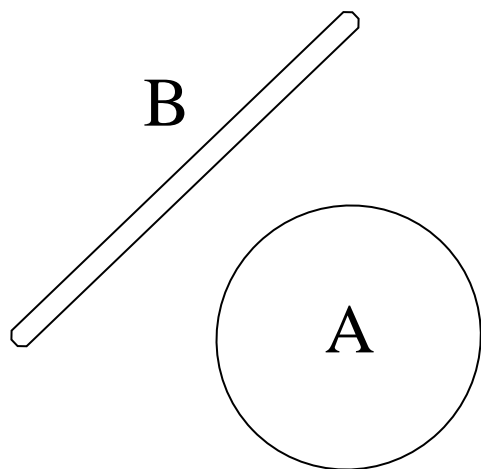
通常在得到实际对象的若干具体特征之后，再由这些原始特征产生出对分类识别最有效、数目最少的特征，这就是特征提取与选择的任务。从本质上讲，我们的目的是使在最小维数特征空间中异类模式点相距较远（类间距离较大），而同类模式点相距较近（类内距离较小）。

5.3 特征选择和特征提取的区别

- 所谓特征选择，就是从 n 个度量值集合 $\{x_1, x_2, \dots, x_n\}$ 中，按某一准则选取出 m 维（ $m < n$ ）供分类用的子集，作为降维的分类特征；
- 所谓特征提取，就是使 (x_1, x_2, \dots, x_n) 通过某种变换，产生 m 个特征 (y_1, y_2, \dots, y_m) （ $m < n$ ），作为新的分类特征（或称为二次特征）；
- 其目的都是为了在尽可能保留识别信息的前提下，降低特征空间的维数，已达到有效的分类。



例：特征选择与特征提取的区别：对一个条形和圆进行识别。



解：[法1]

① 特征抽取：测量三个结构特征

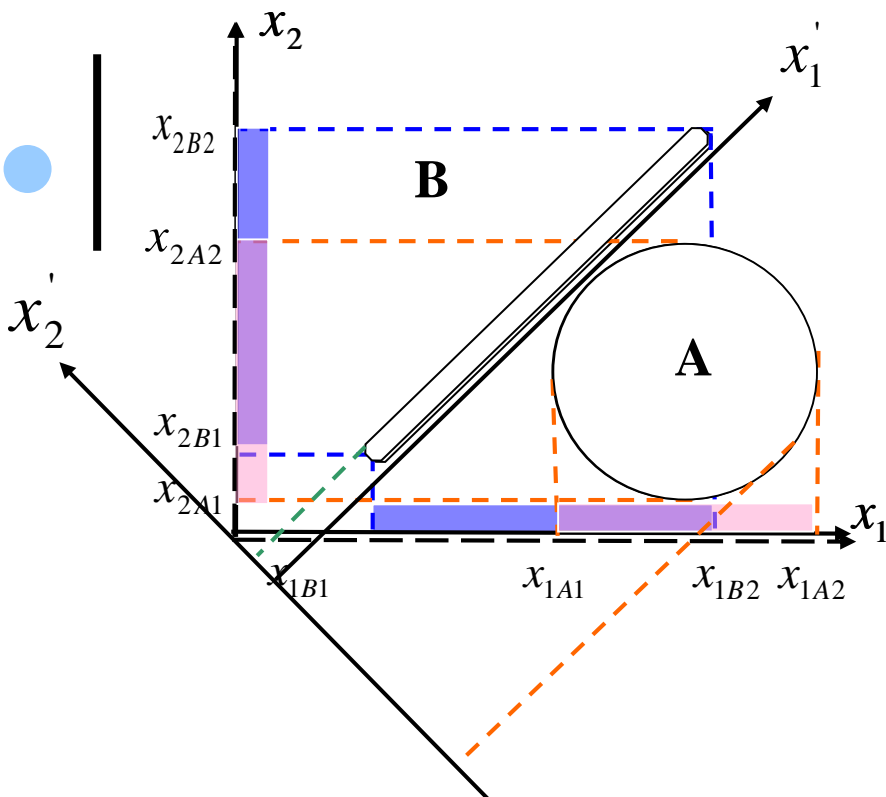
(a) 周长

(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。

5.4 特征提取与选择

特征提取与选择的两个基本途径

(1) **直接选择法**：当实际用于分类识别的特征数目 d 确定后，直接从已获得的 n 个原始特征中选出 d 个特征 x_1, x_2, \dots, x_d ，使可分性判据 J 的值满足下式：

$$J(x_1, x_2, \dots, x_d) = \max [J(x_{i_1}, x_{i_2}, \dots, x_{i_d})]$$

式中 $x_{i_1}, x_{i_2}, \dots, x_{i_d}$ 是 n 个原始特征中的任意 d 个特征，上式表示直接寻找 n 维特征空间中的 d 维子空间。

主要方法有：**分支定界法、用回归建模技术确定相关特征**等方法。



5.4 特征提取与选择

特征提取与选择的两个基本途径

(2) **变换法**，在使判据 J 取最大的目标下，对 n 个原始特征进行变换降维，即对原 n 维特征空间进行坐标变换，然后再取子空间。

主要方法有：**基于可分性判据的特征选择、基于误判概率的特征选择、离散K-L变换法(DKLT)、基于决策界的特征选择等方法。**

5.5 类别可分性判据

构造可分性判据

为确立特征提取和选择的准则：引入类别可分性判据，来刻划特征对分类的贡献。为此希望所构造的可分性判据满足下列要求：

- (1) 与误判概率(或误分概率的上界、下界)有单调关系。
- (2) 当特征相互独立时，判据有可加性，即：

$$J_{ij}(x_1, x_2, \dots, x_d) = \sum_{k=1}^d J_{ij}(x_k)$$

式中， x_1, x_2, \dots, x_d 是对不同种类特征的测量值， $J_{ij}(\cdot)$ 表示使用括号中特征时第*i*类与第*j*类可分性判据函数。

5.5 类别可分性判据

构造可分性判据

(3) 判据具有“距离”的某些特性，即：

$$J_{ij} > 0, \text{ 当 } i \neq j \text{ 时;}$$

$$J_{ij} = 0, \text{ 当 } i = j \text{ 时;}$$

$$J_{ij} = J_{ji}$$

(4) 对特征数目是单调不减，即加入新的特征后，判据值不减。

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



5.5 类别可分性判据

构造可分性判据

值得注意的是：上述的构造可分性判据的要求，即“**单调性**”、“**叠加性**”、“**距离性**”、“**单调不减性**”。在实际应用并不一定能同时具备，但并不影响它在实际使用中的价值。

基于几何距离的可分性判据

一般来讲，不同类的模式可以被区分是由于它们所属类别在特征空间中的类域是不同的区域。显然，区域重叠的部分越小或完全没有重叠，类别的可分性就越好。因此可以用距离或离差测度（散度）来构造类别的可分性判据

5.5 类别可分性判据

基于几何距离的可分性判据

(一) 点与点的距离

$$d(\vec{a}, \vec{b}) = [(\vec{a} - \vec{b})^T (\vec{a} - \vec{b})]^{1/2} = \left[\sum_{k=1}^n (a_k - b_k)^2 \right]^{1/2}$$

(二) 点到点集的距离

用均方欧氏距离表示

$$\bar{d}^2(\vec{x}, \{\vec{a}_k^{(i)}\}) = \frac{1}{N_i} \sum_{k=1}^{N_i} d^2(\vec{x}, \vec{a}_k^{(i)})$$

5.5 类别可分性判据

基于几何距离的可分性判据

(三) 类内及总体的均值矢量

类的均值矢量:
$$\vec{m}^{(i)} = \frac{1}{N_i} \sum_{k=1}^{N_i} \vec{x}_k^{(i)} \quad i = 1, 2, \dots, c$$

各类模式的总体均值矢量
$$\vec{m} = \sum_{i=1}^c P_i \vec{m}^{(i)}$$

P_i 为相应类的先验概率，当用统计量代替先验概率时，总体均值矢量可表示为：

$$\vec{m} = \sum_{i=1}^c P_i \vec{m}^{(i)} = \sum_{i=1}^c \frac{N_i}{N} \vec{m}^{(i)} = \frac{1}{N} \sum_{i=1}^c \sum_{k=1}^{N_i} \vec{x}_k^{(i)} = \frac{1}{N} \sum_{l=1}^N \vec{x}_l$$

5.5 类别可分性判据

基于几何距离的可分性判据

(四) 类内距离

$$\bar{d}^2(\omega_i) = \frac{1}{N_i} \sum_{k=1}^{N_i} (\vec{x}_k^{(i)} - \vec{m}^{(i)})^T (\vec{x}_k^{(i)} - \vec{m}^{(i)})$$

类内均方欧氏距离

类内均方距离也可定义为：

$$\bar{d}_c^2(\omega_i) = \frac{1}{N_i(N_i - 1)} \sum_{k=1}^{N_i} \sum_{l=1}^{N_i} d^2(x_k^{(i)}, x_l^{(i)})$$

5.5 类别可分性判据

基于几何距离的可分性判据

(五) 类内离差矩

阵

$$S_{\omega_i} = \frac{1}{N_i} \sum_{k=1}^{N_i} (\vec{x}_k^{(i)} - \vec{m}^{(i)})(\vec{x}_k^{(i)} - \vec{m}^{(i)})^T$$

(六) 两类之间的距离

$$\bar{d}^2(\omega_i, \omega_j) = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} d^2(\vec{x}_k^{(i)}, \vec{x}_l^{(j)})$$

$$\bar{d}^2(\omega_i, \omega_j) = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} (\vec{x}_k^{(i)} - \vec{x}_l^{(j)})^T (\vec{x}_k^{(i)} - \vec{x}_l^{(j)})$$

5.5 类别可分性判据

基于几何距离的可分性判据

(七) 各类模式之间的总的均方距离

$$\bar{d}^2(\vec{x}) = \frac{1}{2} \sum_{i=1}^c P_i \sum_{j=1}^c P_j \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} d^2(\vec{x}_k^{(i)}, \vec{x}_l^{(j)})$$

当取欧氏距离时，总的均方距离为

$$\bar{d}^2(\vec{x}) = \frac{1}{2} \sum_{i=1}^c P_i \sum_{j=1}^c P_j \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} (\vec{x}_k^{(i)} - \vec{x}_l^{(j)})^T (\vec{x}_k^{(i)} - \vec{x}_l^{(j)})$$

5.5 类别可分性判据

基于几何距离的可分性判据

(八) 多类情况下总的类内、类间及总体离差矩阵

类内离差 $S_W = \sum_{i=1}^c P_i S_{\omega_i} = \sum_{i=1}^c P_i \frac{1}{N_i} \sum_{k=1}^{N_i} (\vec{x}_k^{(i)} - \vec{m}^{(i)})(\vec{x}_k^{(i)} - \vec{m}^{(i)})^T$

类间离差 $S_B = \sum_{i=1}^c P_i (\vec{m}^{(i)} - \vec{m})(\vec{m}^{(i)} - \vec{m})^T$

总体离差 $S_T = \frac{1}{N} \sum_{l=1}^N (\vec{x}_l - \vec{m})(\vec{x}_l - \vec{m})^T = S_W + S_B$

5.5 类别可分性判据

基于几何距离的可分性判据

可以用 S_W 、 S_B 、 S_T 构造不同的可分性判据：

$$J_1 = \text{Tr}[S_W^{-1} S_B] \quad J_2 = \ln \left[\frac{|S_B|}{|S_W|} \right]$$

$$J_3 = \frac{\text{Tr}[S_B]}{\text{Tr}[S_W]} \quad J_4 = \frac{|S_W + S_B|}{|S_W|} = \frac{|S_T|}{|S_W|}$$

可以证明 J_1 、 J_2 和 J_4 在任何非奇异线性变换下是不变的， J_3 与坐标系有关。



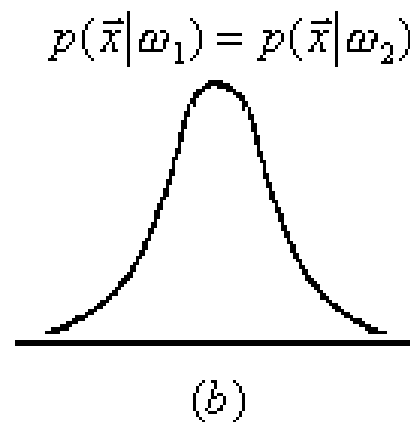
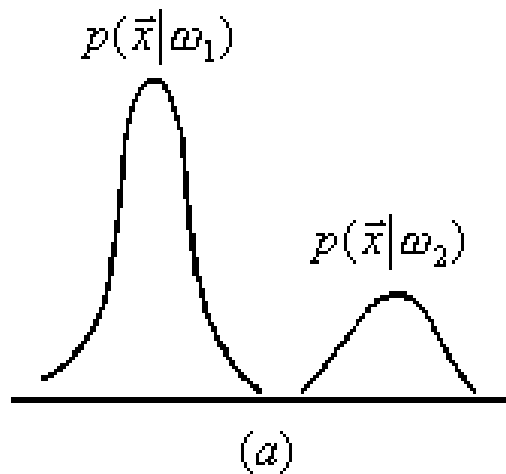
5.5 类别可分性判据

基于几何距离的可分性判据

在特征空间中，当类内模式较密聚，而不同类的模式相距较远时，从直觉上我们知道分类就较容易，由各判据的构造可知，这种情况下所算得的判据值也较大。由判据的构造我们还可以初步了解运用这类判据的原则和方法。

5.5 类别可分性判据

基于类的概率密度函数的可分性判据



考虑两类问题。上图是一维的两类概率分布密度。

(a) 表示两类是完全可分的。

(b) 是完全不可分的。

5.5 类别可分性判据

基于类的概率密度函数的可分性判据

可用两类概密函数的重叠程度来度量可分性，构造基于类概密的可分性判据。此处的所谓重叠程度是指两个概密函数相似的程度。

由类概密构造的可分性判据 J_p 应满足：

- (1) $J_p \geq 0$;
- (2) 当两类概密函数完全不重叠时， $J_p = \max$;
- (3) 当两类概密函数完全重合时， $J_p = 0$;
- (4) 相对两个概密具有“对称性”。

5.5 类别可分性判据

基于类的概率密度函数的可分性判据

(一) Bhattacharyya 判据 (J_B)

受相关概念与应用的启发，我们可以构造B-判据，它的计算式为

$$J_B = -\ln \int_{\Omega} [p(\vec{x}|\omega_1)p(\vec{x}|\omega_2)]^{\frac{1}{2}} d\vec{x}$$

式中 Ω 表示特征空间。在最小误判概率准则下，误判概率有

$$P_0(e) \leq [P(\omega_1)P(\omega_2)]^{\frac{1}{2}} \exp [-J_B]$$

5.5 类别可分性判据

基于类的概率密度函数的可分性判据

(二) Chernoff 判据 (J_C)

我们可以构造比 J_B 更一般的判据, 称其为 C -判据, 其定义式为

$$J_C = -\ln \int_{\Omega} p(\vec{x}|\omega_1)^s p(\vec{x}|\omega_2)^{1-s} d\vec{x}$$

$$\triangleq J_C(\omega_1, \omega_2; s)$$

$$\triangleq J_C(s; x_1, x_2, \dots, x_n) \triangleq J_C(s) \quad 0 < s < 1$$

5.5 类别可分性判据

基于类的概率密度函数的可分性判据

J_C 具有如下性质:

(1) 对一切 $0 < s < 1$, $J_C \geq 0$;

(2) 对一切 $0 < s < 1$, $J_C = 0 \Leftrightarrow p(\vec{x}|\omega_1) = p(\vec{x}|\omega_2)$;

(3) 当参数 s 和 $(1-s)$ 互调时, 有对称性,

$$J_C(\omega_1, \omega_2; s) = J_C(\omega_2, \omega_1; 1-s)$$

(4) 当 \vec{x} 的各分量 x_1, x_2, \dots, x_n 相互独立时,

$$J_C(s; x_1, x_2, \dots, x_n) = \sum_{l=1}^n J_C(s; x_l)$$

5.5 类别可分性判据

基于类的概率密度函数的可分性判据

(5) 当 \bar{x} 的各分量 x_1, x_2, \dots, x_n 相互独立时, 有

$$J_C(s; x_1, x_2, \dots, x_{k-1}) \leq J_C(s; x_1, x_2, \dots, x_{k-1}, x_k) \quad (k \leq n)$$

(6) 最小误判概率

$$P_0(e) \leq P(\omega_1)^s P(\omega_2)^{1-s} \exp[-J_C(\omega_1, \omega_2; s)] \quad (0 < s < 1)$$

5.5 类别可分性判据

基于后验概率的可分性判据

在信息论中，熵 (Entropy) 表示不确定性，熵越大不确定性越大。可以借用熵的概念来描述各类的可分性。

对于c类问题，给定各类的后验概率 $p(\omega_i|\vec{x})$ 可以写成如下形式：

$$\left\{ \begin{array}{cccc} \omega_1 & \omega_2 & \cdots & \omega_c \\ p(\omega_1|\vec{x}) & p(\omega_2|\vec{x}) & \cdots & p(\omega_c|\vec{x}) \end{array} \right\} \triangleq \left\{ \begin{array}{cccc} \omega_1 & \omega_2 & \cdots & \omega_c \\ p_1 & p_2 & \cdots & p_c \end{array} \right\}$$

熵的定义：

$$H(\vec{x}) \triangleq H(\vec{p}) = -\sum_{i=1}^c p_i \log p_i = -\sum_{i=1}^c p(\omega_i|\vec{x}) \log p(\omega_i|\vec{x})$$

由洛必达法则知：当 $p_i = 0$ 时 $p_i \log p_i = 0$

5.5 类别可分性判据

基于后验概率的可分性判据

熵的主要性质:

(1) $H(\vec{p}) \geq 0$, 即 $H(\vec{p})$ 是有下界的非负数。当且仅当某个 i 有 $p_i=1$ 而其余的 $p_j=0 (j \neq i)$ 时等号成立, 即确定场熵最小。

例如: $p(\omega_i|\vec{x}) = 1 \quad p(\omega_j|\vec{x}) = 0, \forall j \neq i$

显然这时能实现完全正确的分类识别

5.5 类别可分性判据

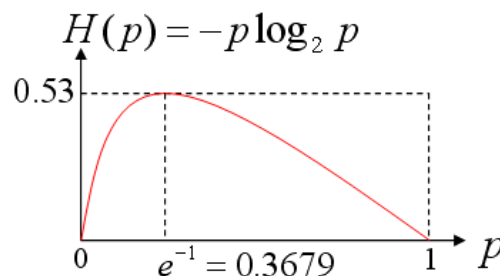
基于后验概率的可分性判据

熵的主要性质:

(2) $H(p_1, \dots, p_c) \leq \log c$, 即 $H(\vec{p})$ 有上界。当且仅当 $p_i = \frac{1}{c}$ ($i = 1, 2, \dots, c$) 时等号成立, 即等概率场熵最大。识别最困难。

(3) $H(\vec{p})$ 是 \vec{p} 的连续上凸函数。

设 Ω_s 是某一降维的特征空间, 如果 $H(\vec{x}) (\vec{x} \in \Omega_s)$ 较大, 则说明 $p(\omega_i | \vec{x}), (i = 1, 2, \dots, c)$ 较接近, 这时分类正确率可能不高。



5.5 类别可分性判据

基于后验概率的可分性判据

熵的主要性质：

$$(4) \quad H(p_1, p_2, \dots, p_c) = H(p_1 + p_2, p_3, \dots, p_c) + (p_1 + p_2)H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$

其中 $p_1 + p_2 > 0$

说明当类别较少时，分类识别的不确定性变小。

从特征选择角度看，我们应选择使熵最小的那些特征用于分类即选用具有最小不确定性的特征进行分类是有益的。

5.5 类别可分性判据

使熵最小的特征利于分类，取熵的期望：

$$J_H = E_{\vec{x}} \left\{ - \sum_{i=1}^c P(\omega_i | \vec{x}) \log P(\omega_i | \vec{x}) \right\} \Rightarrow \min$$

广义熵（具有熵的性质，利于计算）_c 定义

为： $H^{(\alpha)}(p_1, p_2, \dots, p_c) = (2^{1-\alpha} - 1)^{-1} (\sum p_i^\alpha - 1)$

式中 $\alpha > 0$ ， $\alpha \neq 1$ 。不同的 α 值可得不同的可分性度量。

当 $\alpha \rightarrow 1$ 时，由洛必达法则可得**Shannon熵**

$$H^{(1)}(\vec{p}) = - \sum_{i=1}^c p_i \log p_i$$

当 $\alpha = 2$ 时，可得平方熵

$$H^{(2)}(\vec{p}) = 2 \left[1 - \sum_{i=1}^c p_i^2 \right]$$

5.5 类别可分性判据

基于后验概率的可分性判据

同理，我们亦可用点熵在整个特征空间的概率平均

$$J_H^{(\alpha)} = E_{\vec{x}} \left[J^\alpha \left(p(\omega_1 | \vec{x}), p(\omega_2 | \vec{x}), \dots, p(\omega_c | \vec{x}) \right) \right]$$

作为可分性判据。

使用 J_H 判据进行特征提取与选择时，我们的目标是使

$$J_H \Rightarrow \min$$



5.6 离散K-L变换

- 全称：**Karhunen-Loeve**变换（卡洛南-洛伊变换）
- 简单删掉某 $n-k$ 个特征的做法并不十分理想，因为一般来说，原来的 n 个数据各自在不同程度上反映了识别对象的某些特征，简单地删去某些特征可能会丢失较多的有用信息。
- 如果将原来的特征做正交变换，获得的每个数据都是原来 n 个数据的线性组合，然后从新的数据中选出少数几个，使其尽可能多地反映各类模式之间的差异，而这些特征间又尽可能相互独立，则比单纯的选择方法更灵活、更有效。
- **K-L**变换就是一种适用于任意概率密度函数的正交变换。

5.6 离散K-L变换

正交变换

- 对随机变量 $x(t)=[x(1), x(2), \dots, x(n)]^t$ 做离散正交展开, 正交函数为 $\phi_j = [\phi_j(1), \phi_j(2), \dots, \phi_j(n)]^t$, 则对每一模式可分别写成:

$$x = \sum_{j=1}^n a_j \phi_j = \Phi a \quad a = \Phi^t x$$

其中矩阵
而且

$$\Phi = [\phi_1, \phi_2, \dots, \phi_n]$$

$$\phi_j^t \phi_k = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}$$

- 对各个模式类别, 正交函数都是相同的, 但其展开系数向量 a_j 则因类别的不同模式分布而异。

5.6离散K-L变换

利用K-L展开式实现特征选择

- 原理：从 n 个特征向量中取出 m 个组成变换矩阵 Φ ，即
$$\Phi = (\phi_1 \ \phi_2 \ \dots \ \phi_m), \quad m < n$$
此时， Φ 是一个 $n \times m$ 维矩阵， \mathbf{x} 是 n 维向量，经过 $\Phi^t \mathbf{x}$ 变换，即得到降维为 m 的新向量。
- 问题：如何选取变换矩阵 Φ ，使得降维后的新向量在最小均方差条件下接近原来的向量 \mathbf{x} ？

按K-L展开式选择特征

- K-L展开式系数 \mathbf{a}_j 也就是变换后的特征，用 \mathbf{y}_j 表示，写成向量形式： $\mathbf{y} = \Phi^T \mathbf{x}$ 。此时变换矩阵 Φ 用 m 个特征向量组成。

- 为使误差最小，不采用的特征向量，其对应的特征值应尽可能小。因此，将特征值按大小次序标号，即

$$\lambda_1 > \lambda_2 > \dots > \lambda_m > \dots > \lambda_n \geq 0$$

- 若首先采用前面的 m 个特征向量，便可使变换误差最小。此时的变换矩阵为

$$\Phi = (\phi_1 \ \phi_2 \ \dots \ \phi_m)$$

离散K-L变换举例

例 两个模式类的样本分别为

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

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

利用自相关矩阵 \mathbf{R} 作K-L变换，把原样本集压缩成一维样本集。

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 λ_1 对应的特征向量 \mathbf{u}_1 ，归一化后为

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

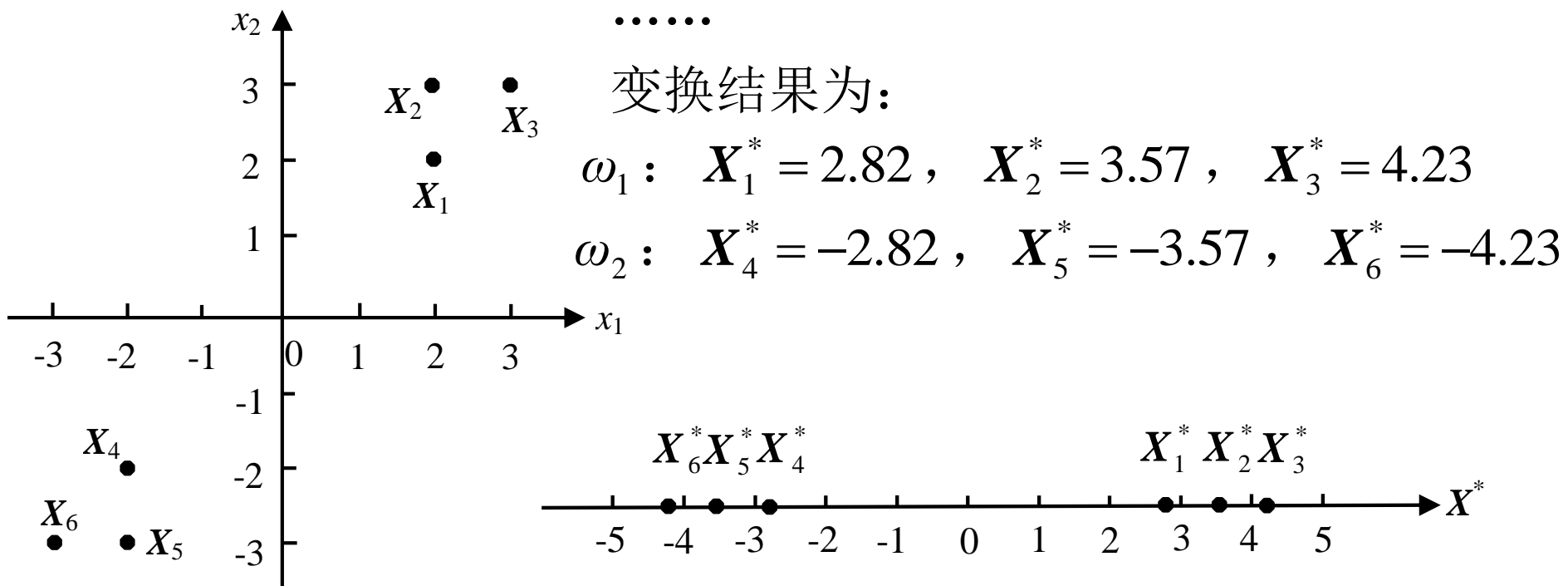


变换矩阵为

$$U = [u_1] = \begin{bmatrix} 0.66 \\ 0.75 \end{bmatrix}$$

第四步：利用 U 对样本集中每个样本进行 K-L 变换。

$$X_1^* = U^T X_1 = [0.66 \ 0.75] \begin{bmatrix} 2 \\ 2 \end{bmatrix} = 2.82$$



K-L变换的应用

- 通过屏幕上的图形显示来大致分析高维模式样本的聚类情况。
- 降维和压缩
 - 对于一幅M行N列的图像，原始的特征空间维数为 $M*N$ 。
 - 如果在K-L变换中只使用30个基，则维数降到30。
 - 降维可实现数据压缩。30个基(基是不是图像？)加上每幅图像的描述参数。
- 人脸识别
 - 收集要识别的人脸，组成人脸图像库；
 - 利用K-L变换确定人脸基图像；
 - 得到每幅图像的参数向量；
 - 对待识别的图像，经预处理后，进行K-L变换，得到参数向量；
 - 将该参数向量与人脸图像库中的参数向量进行比较，最相似参数向量对应的人脸即为要识别的结果。



K-L变换的应用

例：PCA在人脸识别中的应用。

在人脸识别中，PCA是一种常用的特征提取方法。设一幅 $p \times q$ 大小的人脸图像，可以将它看成是一个矩阵 $(f_{ij})_{p \times q}$ ， f_{ij} 为图像在该点的灰度。若将该矩阵按列相连构成一个 $p \times q$ 维向量 $x = (f_{11}, f_{21}, \dots, f_{p1}, f_{21}, f_{22}, \dots, f_{p2}, \dots, f_{1q}, f_{2q}, \dots, f_{pq})^T$ 。设训练样本集为 $X = \{x_1, x_2, \dots, x_N\}$ ，包含N幅图像。

K-L变换的应用

N幅图像的协方差矩阵为：

$$\mathbf{R} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

其中，

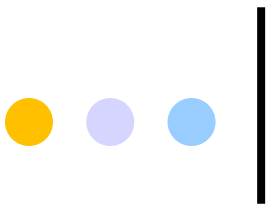
$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$$

求出矩阵 \mathbf{R} 的前 m 个最大特征值 $\lambda_1, \lambda_2, \dots, \lambda_m$ 及其对应的正交化、归一化特征向量 $\alpha_1, \alpha_2, \dots, \alpha_m$ 。分别将这 m 个特征向量化为 $p \times q$ 矩阵，得到 m 幅图像，称为“特征脸”(eigenface)。

下图显示的是对应前最大特征值的特征向量的图像。



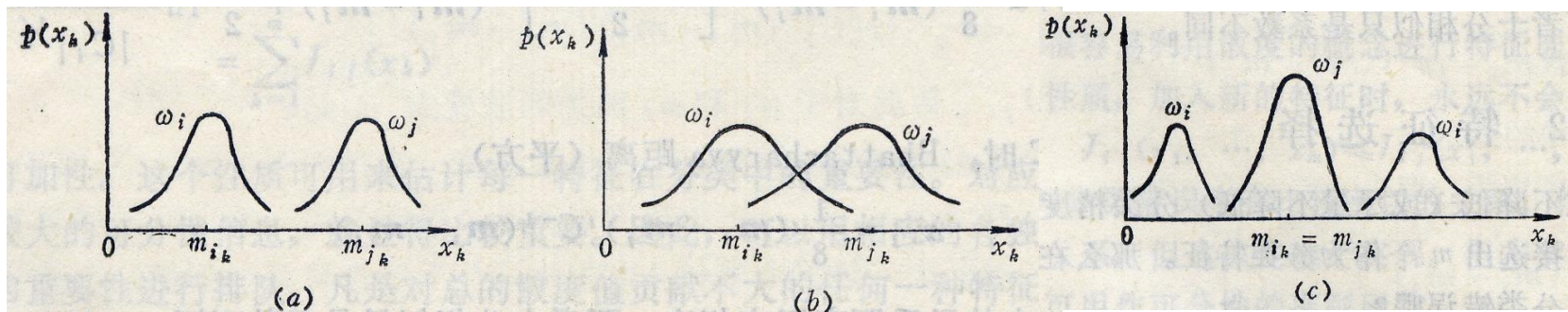
“特征脸”图像



将每一幅人脸图像投影到由 a_1, a_2, \dots, a_m 张成的子空间中，对应于该子空间的一个点，该点的坐标系数对应于图像在子空间的位置，可以作为识别人脸的依据。对于任意待识别样本 \mathbf{x} ，可通过向“特征脸”子空间投影获得系数向量 $\mathbf{y}=(a_1, a_2, \dots, a_m)^T \mathbf{x}$ 。

5.7 特征选择

- 所谓特征选择，就是从 n 个度量值集合 $\{x_1, x_2, \dots, x_n\}$ 中，按某一准则选取出供分类用的子集，作为降维（ m 维， $m < n$ ）的分类特征。
- 穷举法。
- 独立特征的选择准则
 - 适用场合：原始特征测量值统计独立，且类概率密度函数近似正态分布
 - 可分性准则函数为



5.7 特征选择

○ 一般特征的选择准则

- 散布矩阵准则
- 散度准则
- Battacharyya距离准则

○ 散布矩阵准则

- 类内散布矩阵 S_w ，类间散布矩阵 S_b ，总体散布矩阵 S_t 。
- 散布矩阵准则的形式

$$J_1 = tr(S_w^{-1} S_b) \quad J_2 = tr(S_t^{-1} S_b)$$

$$J_3 = \ln\left(\frac{|S_b|}{|S_w|}\right) \quad J_4 = \frac{tr(S_b)}{tr(S_w)}$$

$$J_5 = \frac{(S_w + S_b)}{S_w}$$

作业

1 假定 ω_i 类的样本集为 $\{X_1, X_2, X_3, X_4\}$ ，它们分别为

$$X_1 = [2, 2]^T, \quad X_2 = [3, 2]^T, \quad X_3 = [3, 3]^T, \quad X_4 = [4, 2]^T$$

- (1) 求类内散布矩阵；
- (2) 求类内散布矩阵的特征值和对应的特征向量；
- (3) 求变换矩阵 A ，将二维模式变换为一维模式。

2 给定先验概率相等的两类，其均值向量分别为： $\mu_1 = [1, 3, -1]^T$ 和 $\mu_2 = [-1, -1, 1]^T$ ，协方差矩阵是

$$J_2 = \text{tr}(S_w^{-1} S_b) \quad \Sigma_1 = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

求用 J_2 判据的最优特征提取。

作业

2 给定先验概率相等的两类,其均值向量分别为: $\mu_1=[1,3,-1]^T$ 和 $\mu_2=[-1,-1,1]^T$,协方差矩阵是

$$J_5 = \frac{|S_w + S_b|}{|S_w|} \quad \Sigma_1 = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

求用 J_5 判据的最优特征提取。