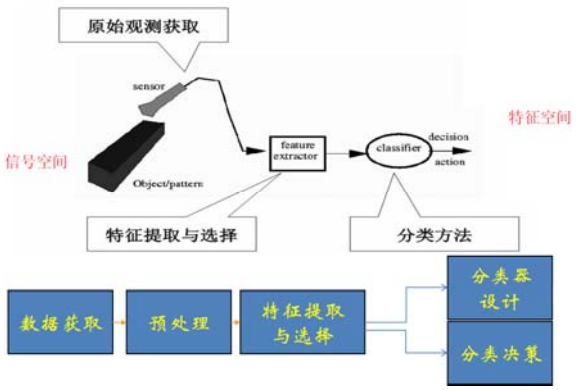
第七章
特征的选择与提取

2009-11-24

7.1 引言



7.1 引言

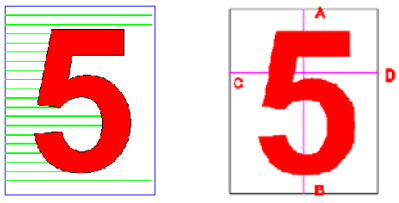
- 特征的选择与提取是模式识别中重要而困难的一个环节：
  - 分析各种特征的有效性并选出最有代表性的特征是模式识别的关键一步；
  - 降低特征维数在很多情况下是有效设计分类器的重要课题；
- 三大类特征：物理、结构和数学特征
  - **物理和结构特征**：易于为人的直觉感知，但有时难于定量描述，因而不易用于机器判别
  - **数学特征**：易于用机器定量描述和判别，如基于统计的特征

7.1 引言

- 确定特征空间的不同层次
  - **物理量的获取与转换 ⇨ 形成原始特征（测量）**
    - 实例：
      - 数字图像中的各像素灰度值；
      - 人体的各种生理指标；
    - 原始特征分析：
      - 原始测量不能反映对象本质；
      - 高维原始特征不利于分类器设计：计算量大，冗余，样本分布十分稀疏。

7.1 引言

- 确定特征空间的不同层次
  - **描述事物方法的选择与设计**

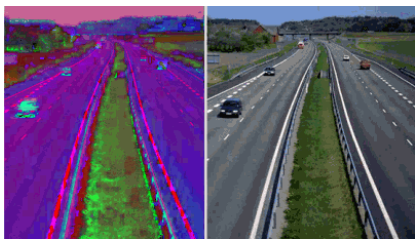


距离　　切割次数

7.1 引言

- 确定特征空间的不同层次
  - **特征空间的优化、降维**

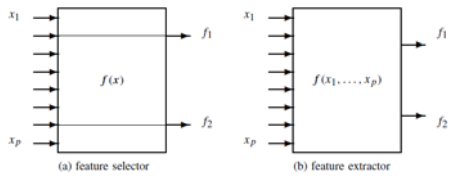
HSI颜色空间



原始图像

7.1 引言

- **优化特征空间的两种基本方法**
  - 特征选择 (selection)：从原始特征中挑选出最有代表性，分类性能最好的特征；
  - 特征提取 (extraction)：用映射（或变换）的方法把原始特征变换为较少的新特征。



7.1 引言

- **优化特征空间的两种基本方法**
  - 特征选择 (selection)：从原始特征中挑选出最有代表性，分类性能最好的特征；
  - 特征提取 (extraction)：用映射（或变换）的方法把原始特征变换为较少的新特征。
  - 特征的选择与提取与具体问题有很大关系，目前没有理论能给出对任何问题都有效的特征选择与提取方法。

7.1 引言

- 特征选择与提取举例 — 细胞自动识别
  - **原始测量**：（正常与异常）细胞的数字图像；
  - **原始特征**（特征的形成，找到一组代表细胞性质的特征）：细胞面积，胞核面积，形状系数，光密度，核内纹理，和浆比；
  - **压缩特征**：原始特征的维数仍很高，需压缩以便于分类；
    - 特征选择：挑选最有分类信息的特征；
    - 特征提取：数学变换；
      - 傅立叶变换或小波变换；
      - 用PCA方法作特征压缩。

7.2 类别可分离性判据

- 类别可分离性判据：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；
- 常见类别可分离性判据：
  - 基于距离的可分性判据；
  - 基于概率分布的判据；
  - 熵函数的可分性判据。

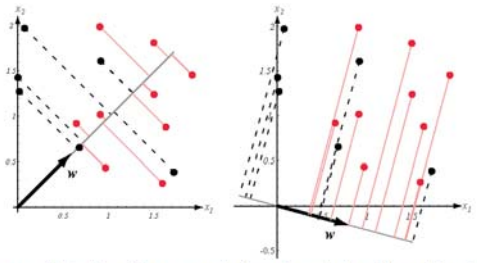
7.2 类别可分离性判据

- 实际的类别可分离性判据应满足的条件：

1. 度量特性：$J_{ij} > 0$, if $i \neq j$; $J_{ij} = 0$, if $i = j$; $J_{ij} = J_{ji}$;
2. 与错误率有单调关系：J_{ij} 大 $\Rightarrow$ P_e 小；
3. 当特征独立时有可加性：$J_{ij}(x_1, x_2, \ldots, x_d) = \sum_{k=1}^{d} J_{ij}(x_k)$;
4. 单调性：$J_{ij}(x_1, x_2, \ldots, x_d) \leq J_{ij}(x_1, x_2, \ldots, x_d, x_{d+1})$.

7.3.1 基于距离的可分性判据

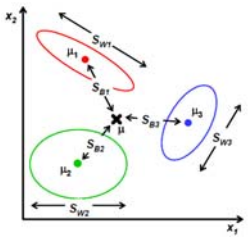
- **实质**：综合考虑不同类样本的类内聚合程度与类间离散程度两个因素。



Figure： Projection of the same set of samples onto two different lines in the directions marked as w. The figure on the right shows greater separation between the red and black projected points.

7.3.1 基于距离的可分性判据

- 描述母体的离散程度



母体类均值：$\boldsymbol{\mu}_i = E_i[\mathbf{x}];$

母体总体均值：$\boldsymbol{\mu} = E[\mathbf{x}];$

母体类间离散度矩阵：$\mathbf{S}_b = \sum_{i=1}^{c} P_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T;$

母体类内离散度矩阵：$\mathbf{S}_w = \sum_{i=1}^{c} P_i E_i(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T;$

7.3.1 基于距离的可分性判据

- 有限样本集下离散度矩阵的估计

样本类均值向量：$\mathbf{m}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} \mathbf{x}_k^{(i)};$

样本总体均值向量：$\mathbf{m} = \sum_{i=1}^{c} P_i \mathbf{m}_i;$

样本类间离散度矩阵：$\tilde{\mathbf{S}}_b = \sum_{i=1}^{c} P_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T;$

样本类内离散度矩阵：

$$\tilde{\mathbf{S}}_w = \sum_{i=1}^{c} P_i \frac{1}{n_i}\sum_{k=1}^{n_i}(\mathbf{x}_k^{(i)} - \mathbf{m}_i)(\mathbf{x}_k^{(i)} - \mathbf{m}_i)^T$$

$$= \sum_{i=1}^{c} P_i \Sigma_i \quad (\Sigma_i: \text{样本类内协方差阵}).$$

7.3.1 基于距离的可分性判据

- 基于各类特征向量间平均距离的判据

$$J_d(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j \frac{1}{n_i n_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j} \delta(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)});$$

其中，$\mathbf{x}_k^{(i)} \in \omega_i,\ k = 1, \cdots, n_i,$

$\mathbf{x}_l^{(j)} \in \omega_j,\ l = 1, \cdots, n_j,$

P_i, P_j：先验概率，

$\delta(\mathbf{x}_k, \mathbf{x}_l)$：$\mathbf{x}_k$和$\mathbf{x}_l$之间的距离度量;

度量ω_i，ω_j两类之间的（平均）分离程度

- 采用欧式距离：$\delta(\mathbf{x}_k, \mathbf{x}_l) = (\mathbf{x}_k - \mathbf{x}_l)^T(\mathbf{x}_k - \mathbf{x}_l),$

➔J_D为各类之间的平均平方距离。

7.3.1 基于距离的可分性判据

- 基于各类特征向量间平均距离的判据
  - 代入欧式距离及 $\mathbf{m}_i, \mathbf{m}$：

$$J_d(\mathbf{x}) = \sum_{i=1}^{c} P_i \left[\frac{1}{n_i}\sum_{k=1}^{n_i}(\mathbf{x}_k^{(i)} - \mathbf{m}_i)^T(\mathbf{x}_k^{(i)} - \mathbf{m}_i)\right]$$

$$+ \sum_{i=1}^{c} P_i \left[(\mathbf{m}_i - \mathbf{m})^T(\mathbf{m}_i - \mathbf{m})\right];$$

类内平均平方距离

各类均值与总体均值的平方距离

$$= \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j (\mathbf{m}_i - \mathbf{m}_j)^T(\mathbf{m}_i - \mathbf{m}_j),$$

即各类均值向量的平均平方距离。

7.3.1 基于距离的可分性判据

- 基于各类特征向量间平均距离的判据
  - 代入欧式距离及 $\mathbf{m}_i, \mathbf{m}$：

$$J_d(\mathbf{x}) = \sum_{i=1}^{c} P_i \left[\frac{1}{n_i}\sum_{k=1}^{n_i}(\mathbf{x}_k^{(i)} - \mathbf{m}_i)^T(\mathbf{x}_k^{(i)} - \mathbf{m}_i)\right]$$

$$+ \sum_{i=1}^{c} P_i \left[(\mathbf{m}_i - \mathbf{m})^T(\mathbf{m}_i - \mathbf{m})\right];$$

$$\rightarrow \quad J_d(\mathbf{x}) = \mathrm{tr}(\tilde{\mathbf{S}}_w + \tilde{\mathbf{S}}_b).$$

7.3.1 基于距离的可分性判据

- 常用的基于类内类间距离的可分性判据

$$J_1(\mathbf{x}) = tr(\mathbf{S}_w + \mathbf{S}_b);$$

$$J_2(\mathbf{x}) = tr(\mathbf{S}_w^{-1}\mathbf{S}_b); \qquad J_3(\mathbf{x}) = \ln\frac{|\mathbf{S}_b|}{|\mathbf{S}_w|};$$

$$J_4(\mathbf{x}) = \frac{tr(\mathbf{S}_b)}{tr(\mathbf{S}_w)}; \qquad J_5(\mathbf{x}) = \frac{|\mathbf{S}_w + \mathbf{S}_b|}{|\mathbf{S}_w|}.$$

基于距离的准则概念直观，计算方便；但与错误率没有直接联系；当各类协差相差不大时，用此判据较好。

7.3.1 基于距离的可分性判据

- 常见的距离度量
  - s阶 Minkowski 度量

$$\delta(x_k,x_l)=\left[\sum_{i=1}^{d}|x_{ki}-x_{li}|^s\right]^{\frac{1}{s}}$$

x_k、x_l为d维向量，x_{ki}、x_{li}为其第i个分量，$i=1,\cdots,d$

  - 城市块（City block）

$$\delta(x_k,x_l)=\sum_{i=1}^{d}|x_{ki}-x_{li}|$$

7.3.1 基于距离的可分性判据

- 常见的距离度量
  - 欧式距离

$$\delta(x_k,x_l)=\left[(x_k-x_l)^T(x_k-x_l)\right]^{\frac{1}{2}}$$

  - Chebychev 距离

$$\delta(x_k,x_l)=\max_i|x_{ki}-x_{li}|$$

7.3.1 基于距离的可分性判据

- 常见的距离度量
  - 平方距离

$$\delta(x_k,x_l)=(x_k-x_l)^T Q(x_k-x_l)$$

Q 是给定的正定标尺矩阵;

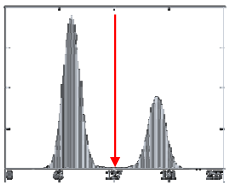
  - 非线性距离度量

$$\delta(x_k,x_l)=\begin{cases}H & \text{iff } \delta_E(x_k,x_l)\geq T\\ 0 & \text{if } \delta_E(x_k,x_l)<T\end{cases}$$

H,T 是非线性度量参数，$\delta(\mathbf{x}_k,\mathbf{x}_l)$ 是上述任何一种度量。

7.3.2 直接应用举例

- **图像二值化**：大津灰度图像阈值法 (Otsu thresholding)
  - 实质是两类分类问题：确定一个灰度（特征值）阈值将图像中的像素分类;
  - 算法基本思想：用类间离散度（方差）作为可分性判据;



7.3.2 直接应用举例

- Otsu thresholding
  - 设图像有 L 个灰度级，n_i为灰度值为 i 的像素数目，图像总像素 $N = n_1 + n_2 + \cdots + n_L$;
  - 灰度为 i 的像素概率（估计）：$p_i = \frac{n_i}{N}$;
  - 当前阈值 T 将图像分为两个部分，计算类间方差:

$$J=\sigma_b^2(T)=\omega_0(\mu_0-\mu_T)^2+\omega_1(\mu_1-\mu_T)^2,$$

其中：$\omega_0(T)=\sum_{i=1}^{T}p_i,\quad \omega_1(T)=\sum_{i=T+1}^{L}p_i=1-\omega_0(T),$

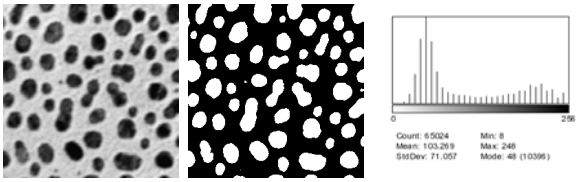
$$\mu_0(T)=\sum_{i=1}^{T}ip_i\Big/\omega_0,\ \mu_1(T)=\sum_{i=T+1}^{L}ip_i\Big/\omega_1,\ \mu_T=\omega_0\mu_0+\omega_1\mu_1=\sum_{i=1}^{L}ip_i.$$

7.3.2 直接应用举例

- Otsu thresholding
  - 求解阈值:

$$T^*=\arg\max_{T=1}^{L}\sigma_b^2(T);$$

  - 演示:



7.3.3 按距离度量的特征提取方法

- **特征提取**：把 D 个特征（原始特征 $\mathbf{x}$）通过变换变为 d 个新特征（$\mathbf{y}$），即 $\mathbf{y}=\mathbf{A}(\mathbf{x})$;
  - **目的**：更好的分类和/或减少计算量;
  - **线性变换**:

$$\mathbf{y}=\mathbf{A}^T\mathbf{x};$$

    - $\mathbf{A}$ 是 $D\times d$ 维矩阵，通常 $D>d$ ⇨ 特征压缩，特征变换;
  - 非线性情况下，可能希望 $D\leqslant d$，如广义线性分类器 SVM，NN。

7.3.3 按距离度量的特征提取方法

- **按欧式距离的特征提取准则函数**

$$J_1(\mathbf{A})=tr\left(\mathbf{A}^T(\mathbf{S}_w+\mathbf{S}_b)\mathbf{A}\right);$$

$$J_2(\mathbf{A})=tr\left[\left(\mathbf{A}^T\mathbf{S}_w\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{S}_b\mathbf{A}\right)\right];$$

$$J_3(\mathbf{A})=\ln\frac{\left|\mathbf{A}^T\mathbf{S}_b\mathbf{A}\right|}{\left|\mathbf{A}^T\mathbf{S}_w\mathbf{A}\right|};$$

$$J_4(\mathbf{A})=\frac{tr\left(\mathbf{A}^T\mathbf{S}_b\mathbf{A}\right)}{tr\left(\mathbf{A}^T\mathbf{S}_w\mathbf{A}\right)};$$

$$J_5(\mathbf{A})=\frac{\left|\mathbf{A}^T\Sigma\mathbf{A}\right|}{\left|\mathbf{A}^T\mathbf{S}_w\mathbf{A}\right|},\quad \Sigma=\mathbf{S}_w+\mathbf{S}_b.$$

目标：求$\mathbf{A}^*$, 使得

$$J(\mathbf{y})=\max_{\{\mathbf{A}\}}J\left(\mathbf{A}^T\mathbf{x}\right).$$

7.3.3 按距离度量的特征提取方法

- **基于J_2判据的特征提取 — 线性判别分析 (LDA)**
  - 结论：J_2 对线性非奇异变换具有不变性。

$$\begin{aligned}J_2(\mathbf{y})&=tr(\mathbf{S}_w^{*-1}\mathbf{S}_b^*)\\&=tr\left[\left(\mathbf{A}^T\mathbf{S}_w\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{S}_b\mathbf{A}\right)\right]\\&=tr\left[\mathbf{A}^{-1}\mathbf{S}_w^{-1}\left(\mathbf{A}^T\right)^{-1}\mathbf{A}^T\mathbf{S}_b\mathbf{A}\right]\\&=tr\left[\mathbf{A}^{-1}\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{A}\right]\\&=tr\left[\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{A}^{-1}\mathbf{A}\right]\\&=tr(\mathbf{S}_w^{-1}\mathbf{S}_b)=J_2(\mathbf{x}).\end{aligned}$$

7.3.3 按距离度量的特征提取方法

- 基于 J_2 判据的特征提取 — LDA
  - **求 J_2 的最大值**

$$J_2(\mathbf{A})=tr\left[\left(\mathbf{A}^T\mathbf{S}_w^x\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{S}_b^x\mathbf{A}\right)\right]$$

$$\frac{\partial J_2(\mathbf{A})}{\partial\mathbf{A}}=0\Rightarrow$$

$$-2\mathbf{S}_w^x\mathbf{A}\left(\mathbf{A}^T\mathbf{S}_w^x\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{S}_b^x\mathbf{A}\right)\left(\mathbf{A}^T\mathbf{S}_w^x\mathbf{A}\right)^{-1}+2\mathbf{S}_b^x\mathbf{A}\left(\mathbf{A}^T\mathbf{S}_w^x\mathbf{A}\right)^{-1}=0$$

$$\Rightarrow\left[\left(\mathbf{S}_w^x\right)^{-1}\mathbf{S}_b^x\right]\mathbf{A}=\mathbf{A}\left[\left(\mathbf{S}_w^y\right)^{-1}\mathbf{S}_b^y\right];$$

7.3.3 按距离度量的特征提取方法

- 基于 J_2 判据的特征提取 — LDA
  - **求 J_2 的最大值**
    - 对称矩阵 $\mathbf{S}_w^y$, $\mathbf{S}_b^y$ 可被一个非奇异线性变换 $\mathbf{B}$（$d\times d$ 维）同时对角化，且

$$\mathbf{B}^T\mathbf{S}_w^y\mathbf{B}=\mathbf{I},\quad \mathbf{B}^T\mathbf{S}_b^y\mathbf{B}=\mathbf{\Lambda};$$

其中，$\mathbf{I}$ 和 $\mathbf{\Lambda}$ 分别是将 $\mathbf{y}$ 空间进行线性变换后所得特征空间 $\mathbf{z}$ 的类内离散度矩阵和类间离散度矩阵，即

$$\mathbf{z}=\mathbf{B}^T\mathbf{y}=\mathbf{B}^T\mathbf{A}^T\mathbf{x};$$

且有 $J_2(\mathbf{z})=J_2(\mathbf{y})$.

7.3.3 按距离度量的特征提取方法

- 基于 J_2 判据的特征提取 — LDA
  - **求 J_2 的最大值**

$$\Rightarrow\left[\left(\mathbf{S}_w^x\right)^{-1}\mathbf{S}_b^x\right]\mathbf{A}=\mathbf{A}\left(\mathbf{B}\mathbf{\Lambda}\mathbf{B}^{-1}\right)\quad\Rightarrow\left[\left(\mathbf{S}_w^x\right)^{-1}\mathbf{S}_b^x\right]\mathbf{AB}=(\mathbf{AB})\mathbf{\Lambda}$$

$$\Rightarrow\left[\left(\mathbf{S}_w^x\right)^{-1}\mathbf{S}_b^x\right]\boldsymbol{\varphi}_i=\boldsymbol{\varphi}_i\lambda_i,\quad i=1,2,\cdots,D$$

$\Rightarrow\lambda_i,\boldsymbol{\varphi}_i$分别是 $\left(\mathbf{S}_w^x\right)^{-1}\mathbf{S}_b^x$ 的特征值和特征向量;

$\Rightarrow\mathbf{\Lambda}$ 的对角元素是$\left(\mathbf{S}_w^x\right)^{-1}\mathbf{S}_b^x$的$d$个特征值;

矩阵 $(\mathbf{AB})$ 由相应的d个特征向量组成;

7.3.3 按距离度量的特征提取方法

- 基于 J_2 判据的特征提取 — LDA
  - **求 J_2 的最大值**

令 $\mathbf{C}=\mathbf{AB}$，

$$\Rightarrow \left[\left(\mathbf{S}_w^x\right)^{-1}\mathbf{S}_b^x\right]\mathbf{C}=\mathbf{C}\boldsymbol{\Lambda} \quad \Rightarrow \mathbf{C}^T\mathbf{S}_b^x\mathbf{C}=\mathbf{C}^T\mathbf{S}_w^x\mathbf{C}\boldsymbol{\Lambda}$$

$$\Rightarrow J_2(\mathbf{y})=tr\left[\left(\mathbf{C}^T\mathbf{S}_w^x\mathbf{C}\right)^{-1}\left(\mathbf{C}^T\mathbf{S}_b^x\mathbf{C}\right)\right]=tr(\boldsymbol{\Lambda})=\sum_{i=1}^{d}\lambda_i^*;$$

$\Rightarrow$ 当 $\lambda_1^*,\cdots,\lambda_d^*$ 是 $\left(\mathbf{S}_w^x\right)^{-1}\mathbf{S}_b^x$ 的D个特征值中最大的d个，$J_2(\mathbf{y})$ 达到最大值。

7.3.3 按距离度量的特征提取方法

- 基于 J_2 判据的特征提取 — LDA

设矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的特征值为 $\lambda_1,\lambda_2,\cdots,\lambda_D$，

且 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_D$，

则取前 d 个特征值对应的特征向量 $\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,\cdots,\boldsymbol{\varphi}_d$，

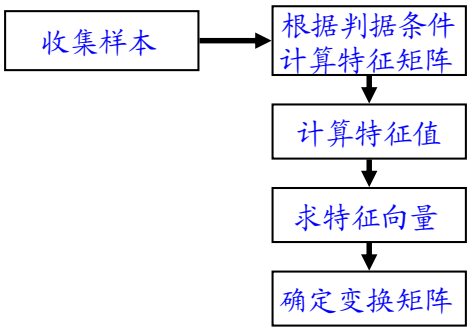
组成的线性变换矩阵 $\mathbf{A}=[\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,\cdots,\boldsymbol{\varphi}_d]$，

则 $\mathbf{A}$ 能使降维后的特征空间的 $J_2(\mathbf{y})$ 判据值最大

- J_3, J_5 对线性非奇异变换具有不变性，J_1, J_4 与坐标系相关，但求解极值的结论相同。

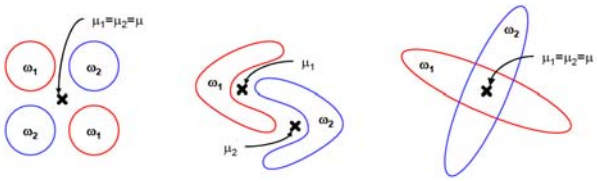
7.3.3 按距离度量的特征提取方法

- **特征提取流程图**



7.3.3 按距离度量的特征提取方法

- LDA 的局限性



7.3.3 按距离度量的特征提取方法

- 例：给定先验概率相等的两类，其均值向量分别为：$\boldsymbol{\mu}_1=[1,3,-1]^T$，$\boldsymbol{\mu}_2=[-1,-1,-1]^T$，协方差矩阵是：

$$\boldsymbol{\Sigma}_1=\begin{bmatrix}4&1&0\\1&4&0\\0&0&1\end{bmatrix},\quad \boldsymbol{\Sigma}_2=\begin{bmatrix}2&1&0\\1&2&0\\0&0&1\end{bmatrix};$$

求用 J_2 判据的最优特征提取。

- 思路：应先求 $\mathbf{S}_w^{-1}\mathbf{S}_b$，再求此矩阵的特征矩阵。

7.3.3 按距离度量的特征提取方法

- 解：

混合均值：$\boldsymbol{\mu}=\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)=[0,1,0]^T$；

类间离散度矩阵：$\mathbf{S}_b=\frac{1}{2}\sum_{i=1}^{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T=\frac{1}{4}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T$；

类内离散度矩阵：$\mathbf{S}_w=\frac{1}{2}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)=\begin{bmatrix}3&1&0\\1&3&0\\0&0&1\end{bmatrix}$；

$$\mathbf{S}_w^{-1}=\frac{1}{8}\begin{bmatrix}3&-1&0\\-1&3&0\\0&0&8\end{bmatrix};$$

7.3.3 按距离度量的特征提取方法

- 注意：$\mathbf{S}_w^{-1}\mathbf{S}_b$ 的秩是1，故 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 只有一个非零特征值；即所求线性变换矩阵 $\mathbf{A}$ 是 $D\times 1$ 矩阵，是一个向量，求解 $\mathbf{a}$ 需解方程

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{a}=\lambda_1\mathbf{a}$$

$$\Rightarrow \frac{1}{4}\mathbf{S}_w^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^T\mathbf{a}=\lambda_1\mathbf{a}$$

$$\because\ \frac{1}{4}(\mu_1-\mu_2)^T\mathbf{a}\text{是标量}$$

$$\therefore\ \mathbf{a}=\mathbf{S}_w^{-1}(\mu_1-\mu_2)=\frac{1}{8}[1,5,-8]^T.$$