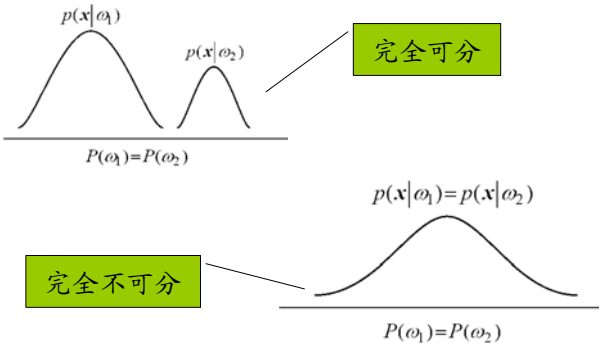
第七章
特征的选择与提取

2009-12-1

7.4 基于概率分布的可分性判据

- 考查两类分布密度之间的交叠程度



7.4 基于概率分布的可分性判据

- **定义**：两个密度函数之间的距离

$$J(\cdot)=\int g\left[p(\mathbf{x}\mid\omega_1),p(\mathbf{x}\mid\omega_2),P_1,P_2\right]d\mathbf{x}$$

  - 须满足如下条件：

1. J_p是非负，即$J_p \geqslant 0$;

2. 当两类完全不交叠时，J_p达到其最大值，即

   if $p(\mathbf{x}\mid\omega_1)p(\mathbf{x}\mid\omega_2)=0,\forall\mathbf{x}$, then $J_p=J_{\max}$;

3. 当两类分布密度相同时，$J_p=0$;

   if $p(\mathbf{x}\mid\omega_1)=p(\mathbf{x}\mid\omega_2),\forall\mathbf{x}$, then $J_p=0$.

7.4.1 常用的概率距离度量

- **Bhattacharyya 距离**

$$J_B=-\ln\int[p(\mathbf{x}\mid\omega_1)p(\mathbf{x}\mid\omega_2)]^{1/2}d\mathbf{x}$$

1) 两类完全重合时，$J_B=0$;

2) 两类完全不交叠时，$J_B=\infty$;

3) 与错误概率的上界有直接关系：

$$P_e\leq\left[P(\omega_1)P(\omega_2)\right]^{1/2}\exp(-J_B).$$

7.4.1 常用的概率距离度量

- **Chernoff 界**

$$J_C=-\ln\int p^s(\mathbf{x}\mid\omega_1)p^{1-s}(\mathbf{x}\mid\omega_2)d\mathbf{x},\ s\in[0,1];$$

1) $s=0.5,\quad J_C=J_B$;

2) for $\forall s\in[0,1]\quad J_C\geq 0$;

3) for $\forall s\in[0,1]\quad J_C=0\Leftrightarrow p(\mathbf{x}\mid\omega_1)=p(\mathbf{x}\mid\omega_2),\forall\mathbf{x}$;

4) 当 $\mathbf{x}$ 的各分量彼此独立时，

$$J_C(s;x_1,x_2,\cdots,x_n)=\sum_{i=1}^{n}J_C(s;x_i);$$

5) 当 $\mathbf{x}$ 的各分量彼此独立时

$$J_C(s;x_1,x_2,\cdots x_k)\leq J_C(s;x_1,x_2,\cdots x_k,x_{k+1}).$$

7.4.1 常用的概率距离度量

- **散度** (Divergence)
  - 利用对数似然比（或似然比），对于某特征值$\mathbf{x}$

$$l_{ij}(\mathbf{x})=\ln\frac{p(\mathbf{x}\mid\omega_i)}{p(\mathbf{x}\mid\omega_j)};$$

    - $p(\mathbf{x}\mid\omega_1)=p(\mathbf{x}\mid\omega_2)\Rightarrow l_{ij}(\mathbf{x})=0$;
    - $p(\mathbf{x}\mid\omega_1)$和$p(\mathbf{x}\mid\omega_2)$ 差异越大，$\left|l_{ij}(\mathbf{x})\right|$ 越大；
  - 考虑整个特征空间概率分布的差异，分别定义对ω_i类及对 ω_j 类的平均可分性信息

$$I_{ij}=E[l_{ij}(\mathbf{x})]=\int_X p(\mathbf{x}\mid\omega_i)\ln\frac{p(\mathbf{x}\mid\omega_i)}{p(\mathbf{x}\mid\omega_j)}d\mathbf{x},\quad I_{ji}=E[l_{ji}(\mathbf{x})].$$

7.4.1 常用的概率距离度量

- **散度 (Divergence)**
  - 定义区分 ω_i类和 ω_j 类的总的平均信息（散度）：

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

1）$J_D \geq 0$;

2）$J_D = 0 \Leftrightarrow p(\mathbf{x}|\omega_1) = p(\mathbf{x}|\omega_2), \forall \mathbf{x}$;

3）当 $\mathbf{x}$ 的各分量彼此独立时，$J_D(x_1, x_2, \cdots, x_n) = \sum_{i=1}^{n} J_D(x_i)$;

4）当 $\mathbf{x}$ 的各分量彼此独立时，

$$J_D(x_1, x_2, \cdots x_k) \leq J_D(x_1, x_2, \cdots x_k, x_{k+1});$$

5）$J_D(\omega_1, \omega_2) = J_D(\omega_2, \omega_1)$.

7.4.1 常用的概率距离度量

- **正态分布下的散度**
  - 设两类别都是 d 维正态分布，分别表示为 $\omega_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$，$\omega_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$，则

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)],$$

$$p(\mathbf{x}|\omega_j) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)];$$

$\Rightarrow$ 对数似然比

$$l_{ij} = \frac{1}{2}\ln|\frac{\boldsymbol{\Sigma}_j}{\boldsymbol{\Sigma}_i}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j);$$

7.4.1 常用的概率距离度量

- **正态分布下的散度**
  - 利用矩阵迹的性质 tr(**BA**^{T})=tr(**A**^{T}**B**) (=**A**^{T}**B**，如**A**，**B**是向量)，似然比可改写成

$$l_{ij} = \frac{1}{2}\ln|\frac{\boldsymbol{\Sigma}_j}{\boldsymbol{\Sigma}_i}| - \frac{1}{2}tr[\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T]$$

$$+\frac{1}{2}tr[\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)(\mathbf{x}-\boldsymbol{\mu}_j)^T];$$

$\Rightarrow$

$$I_{ij} = \frac{1}{2}\ln|\frac{\boldsymbol{\Sigma}_j}{\boldsymbol{\Sigma}_i}| + \frac{1}{2}tr[\boldsymbol{\Sigma}_i(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1})]$$

$$+\frac{1}{2}tr[\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T];$$

7.4.1 常用的概率距离度量

- **正态分布下的散度**

$$J_D = \frac{1}{2}tr[\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i - 2\mathbf{I}]$$

$$+\frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T(\boldsymbol{\Sigma}_i^{-1}+\boldsymbol{\Sigma}_j^{-1})(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j);$$

  - 如果两类协方差矩阵相等，则

$$J_D = (\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) = J_M;$$

Mahalanobis 距离的平方

在协方差矩阵相等条件下散度与 J_d 很相似，都是对样本在特征空间分散程度的描述。

7.4.1 常用的概率距离度量

- **正态分布下的 Bhattacharyya 距离**

$$J_B = \frac{1}{8}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\left(\frac{\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j}{2}\right)^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

$$+\frac{1}{2}\ln\frac{|(\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j)/2|}{\left[|\boldsymbol{\Sigma}_i||\boldsymbol{\Sigma}_j|\right]^{1/2}};$$

  - 如果两类协方差矩阵相等，则

$$8J_B = (\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) = J_D = J_M.$$

7.4.2 概率距离判据下的特征提取

- **基本方法**：类似于距离度量下的特征提取
  - 设

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

求映射后的判据表达式（J_B, J_C, J_D）对 **A** 的各分量的偏导数并令其为零，得到所需的方程式组，然后用相应方法求解。

  - 注意：原空间中一个矩阵 W 经映射后变为

$$\mathbf{W}^* = \mathbf{A}^T\mathbf{W}\mathbf{A}.$$

7.4.2 概率距离判据下的特征提取

- **讨论：两类别问题，正态分布及相同的协方差阵**

$$J_D=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$
$$=tr\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\right]$$
$$=tr\left[\boldsymbol{\Sigma}^{-1}\mathbf{M}\right]\quad(\text{其中，}\mathbf{M}=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T);$$

$$J_D(\mathbf{A})=tr\left[\left(\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{M}\mathbf{A}\right]\quad(\text{具有非奇异变换不变性})$$

$$\frac{\partial J_D(\mathbf{A})}{\partial\mathbf{A}}=-2\boldsymbol{\Sigma}\mathbf{A}\left(\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{M}\mathbf{A}\right)\left(\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}\right)^{-1}+2\mathbf{M}\mathbf{A}\left(\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}\right)^{-1}=0$$

$$\Rightarrow\mathbf{M}\mathbf{A}-\boldsymbol{\Sigma}\mathbf{A}\left(\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{M}\mathbf{A}\right)=0;$$

7.4.2 概率距离判据下的特征提取

- **讨论：两类别问题，正态分布及相同的协方差阵**
  - 设矩阵 $(\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{M}\mathbf{A}$ 的特征值矩阵与特征向量矩阵分别是 $\boldsymbol{\Lambda}$ 和 $\mathbf{U}$，即有

$$\left(\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{M}\mathbf{A}\right)\mathbf{U}=\mathbf{U}\boldsymbol{\Lambda};$$

$\Rightarrow\boldsymbol{\Sigma}^{-1}\mathbf{M}\mathbf{A}\mathbf{U}-\mathbf{A}\mathbf{U}\boldsymbol{\Lambda}=\mathbf{0};$

令 $\mathbf{B}=\mathbf{A}\mathbf{U}$，则 $\mathbf{B}$ 是 $\boldsymbol{\Sigma}^{-1}\mathbf{M}$ 的特征向量矩阵；

对于两类别问题，$rank\left(\boldsymbol{\Sigma}^{-1}\mathbf{M}\right)=1$，即 $\boldsymbol{\Sigma}^{-1}\mathbf{M}\mathbf{A}=\lambda\mathbf{A}$，

$\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\mathbf{A}=\lambda\mathbf{A}\Rightarrow\mathbf{A}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2).$

7.5 基于熵函数的可分性判决

- **熵**：事件不确定性的度量
  - **A** 事件的不确定性大（熵大），则对 **A** 事件的观察所提供的信息量大。
- **思路**：
  - 把各类 ω_i 看作一系列事件，把后验概率 $P(\omega_i|\mathbf{x})$ 看作特征 $\mathbf{x}$ 上出现 ω_i 的概率；
  - 如从 $\mathbf{x}$ 能确定 ω_i，则对 ω_i 的观察不提供信息量，熵为0 ⇨ 特征 $\mathbf{x}$ 有利于分类；
  - 如从 $\mathbf{x}$ 完全不能确定 ω_i，则对 ω_i 的观察信息量大，熵大 ⇨ 特征 $\mathbf{x}$ 无助于分类。

7.5 基于熵函数的可分性判决

- **熵函数**：$H=J_c[P(\omega_1|\mathbf{x}),\cdots,P(\omega_c|\mathbf{x})]$

性质：

①规一化 $J_c\left(\frac{1}{c},\cdots,\frac{1}{c}\right)=1$

$0\le J_c(P_1,\cdots,P_c)\le J_c\left(\frac{1}{c},\cdots,\frac{1}{c}\right)=1$

②对称性 $J_c(P_1,\cdots,P_c)=J_c(P_2,\cdots,P_1)$

③确定性 $J_c(1,0,\cdots,0)=J_c(0,1,\cdots,0)=\cdots=0$

④扩张性 $J_c(P_1,\cdots,P_c)=J_{c+1}(P_1,\cdots,P_c,0)$

⑤连续性 $P(\omega_i|x)$ 的连续函数

⑥分枝性（综合性） 一分为二，则熵增加；二合为一，则熵减小。

7.5 基于熵函数的可分性判决

- 常用的熵函数：
  - **Shannon 熵**：$H_c^1=-\sum_{i=1}^{c}P(\omega_i|\mathbf{x})\log_2P(\omega_i|\mathbf{x});$
  - **平方熵**：$H_c^2=2\left[1-\sum_{i=1}^{c}P^2(\omega_i|\mathbf{x})\right].$
- **熵可分离性判据**：

$$J_e=\int H(\mathbf{x})p(\mathbf{x})d\mathbf{x},$$

  - J_e大，则不同类样本重叠性大，可分性不好；
  - J_e小，则可分性好。

7.5.1 基于判别熵最小化的特征提取

- **相对熵**：表示某一种分布偏离给定标准分布的程度，两种分布重合时最大（0），

$$V(p,q)=-\sum p(\mathbf{x}_i)\log\frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}\le0;$$

- **判别熵**：表示两类分布之间的差别

$$W(p,q)=V(p,q)+V(q,p)$$
$$=-\sum p(\mathbf{x}_i)\log p(\mathbf{x}_i)-\sum q(\mathbf{x}_i)\log q(\mathbf{x}_i)$$
$$+\sum p(\mathbf{x}_i)\log q(\mathbf{x}_i)+\sum q(\mathbf{x}_i)\log p(\mathbf{x}_i);$$

7.5.1 基于判别熵最小化的特征提取

- 为方便计算，判别熵可被代替

$$U(p,q)=-\sum_i (p_i-q_i)^2 \le 0;$$

- 目标:

$$\min W(p,q) \Rightarrow \min U(p,q)=-\sum_i (p_i-q_i)^2 \le 0$$

- 结论：使 U 最小的坐标系统（变换矩阵）是由矩阵 $\mathbf{A}=\boldsymbol{G}^{(1)}-\boldsymbol{G}^{(2)}$ 的前 d 个（按特征值平方大小）特征向量组成的，其中$\boldsymbol{G}^{(1)}$、$\boldsymbol{G}^{(2)}$是两类各自的协方差矩阵（估计）。

7.6 PCA 特征提取方法与K-L变换

- **PCA (Principle Component Analysis) 方法**:
  进行特征降维变换，不能完全地表示原有的对象，能量总会有损失。希望找到一种能量最为集中的的变换方法使损失最小。
  - 通过变换，用较少的特征 $(y_1,y_2,\cdots,y_m)^T$ 近似表示原来的对象 $x=(x_1,x_2,\cdots,x_n)^T$，($m<<n$)，并且使得误差尽可能的小。
- **K-L (Karhunen-Loeve) 变换**：最优正交线性变换，相应的特征提取方法被称为 PCA 方法。

7.6 PCA 特征提取方法与K-L变换

- **正交变换**
  - 给定 n 维空间中的一组标准正交基 $\phi_1,\phi_2,\cdots,\phi_n$，它诱导了一个线性变换:

$$L:\mathbf{x}\to\mathbf{y} \quad L(\mathbf{x})=\mathbf{y}=(y_1,y_2,\cdots,y_n)^T,$$

$$y_i=\mathbf{x}^T\phi_i, \quad i=1,2,\cdots,n,$$

$$\mathbf{x}=\sum_{i=1}^{n} y_i\phi_i, \quad \text{正交展开。}$$

  - 反之，任何一个正交变换也确定了一组正交基。

7.6 PCA 特征提取方法与K-L变换

- **误差**
  - 用 m 个分量表示带来的误差

$$\Delta\mathbf{x}(m)=\mathbf{x}-\sum_{i=1}^{m} y_i\phi_i=\sum_{i=m+1}^{n} y_i\phi_i;$$

  - 目标：误差平方的期望最小

$$e^2(m)=E\left[\|\Delta\mathbf{x}(m)\|^2\right]=E\left[\sum_{i=m+1}^{n} y_i^2\right],$$

7.6 PCA 特征提取方法与K-L变换

- **求解最小均方误差正交基**:
  - 首先假定随机特征向量为零均值（期望）的，否则减掉均值即可

$$E\mathbf{x}=0;$$

  - 找 n 个正交基 $\phi_1,\phi_2,\cdots,\phi_n$，使得对任意一组正交基 $\varphi_1,\varphi_2,\cdots,\varphi_n$，和所有的 $m\leqslant n$，

$$e_\phi^2(m)=E\left[\sum_{i=m+1}^{n}\left(\mathbf{x}^T\phi_i\right)^2\right]\le e_\varphi^2(m)=E\left[\sum_{i=m+1}^{n}\left(\mathbf{x}^T\varphi_i\right)^2\right];$$

7.6 PCA 特征提取方法与K-L变换

- **求解最小均方误差正交基**:
  - 目标函数：对于一个固定的 m

$$\min e_\phi^2(m)=\min E\left[\sum_{i=m+1}^{n}\left(\phi_i^T\mathbf{x}\mathbf{x}^T\phi_i\right)\right]$$

$$=\min\sum_{i=m+1}^{n}\left(\phi_i^T\boldsymbol{\Sigma}\phi_i\right),$$

$$s.t. \quad \|\phi_i\|^2=1, i=m+1,m+2,\cdots,n;$$

$\boldsymbol{\Sigma}=E(\mathbf{x}\mathbf{x}^T)$，$\mathbf{x}$ 的协方差矩阵

7.6 PCA 特征提取方法与K-L变换

- **求解最小均方误差正交基:**
  - 用 Lagrange 乘子法

$$g(\phi)=\sum_{i=m+1}^{n}\left(\phi_i^T\Sigma\phi_i\right)-\sum_{i=m+1}^{n}\left[\lambda_i\left(\phi_i^T\phi_i-1\right)\right];$$

令 $\dfrac{\partial g(\phi)}{\partial\phi_i}=0,\quad i=m+1,\cdots,n,$

$\Rightarrow \Sigma\phi_i=\lambda_i\phi_i,\quad i=m+1,\cdots,n;$

$\Rightarrow \phi_i$ 是Σ的特征向量，λ_i 是特征根;

且，$e_\phi^{\,2}(m)=\sum_{i=m+1}^{n}\left(\phi_i^T\Sigma\phi_i\right)=\sum_{i=m+1}^{n}\lambda_i.$

7.6 PCA 特征提取方法与K-L变换

- 协方差矩阵的所有特征根是实数，特征向量也是实的，所有 n 个特征向量构成一组标准正交基，记作 $\xi_1,\xi_2,\cdots,\xi_n$，分别对应特征根 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_n$，

$$\Sigma=(\xi_1,\xi_2,\cdots,\xi_n)\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n\end{pmatrix}\begin{pmatrix}\xi_1^T\\ \xi_2^T\\ \vdots\\ \xi_n^T\end{pmatrix}.$$

- 选择协方差矩阵的特征向量 $\xi_1,\xi_2,\cdots,\xi_m$ 作为正交基，可以使得均方误差最小。

7.6 PCA 特征提取方法与K-L变换

- **总结**
  - 当取协方差矩阵 Σ 的 m 个最大特征值对应的特征向量来展开 $\mathbf{x}$ 时，此时的截断均方误差最小。这 m 个特征向量组成的正交坐标系称作 $\mathbf{x}$ 所在的 n 维空间的 m 维 Karhunen-Loeve (K-L) 变换坐标系，$\mathbf{x}$ 在 K-L 坐标系上的展开系数向量 $\mathbf{y}$ 称作 $\mathbf{x}$ 的 K-L 变换。
  - 实际应用中，协方差矩阵是未知的，用样本协方差矩阵代替。

7.6 PCA 特征提取方法与K-L变换

- 特征向量常被叫做“主分量”，每个样本被它在前几个主分量上的投影近似表示

$$\mathbf{x}=\sum_{i=1}^{n}y_i\xi_i\approx\sum_{i=1}^{m}y_i\xi_i=\sum_{i=1}^{m}\left(\mathbf{x}^T\xi_i\right)\xi_i;$$

- 特征值标记着相应特征向量上的能量;
- $\xi_1,\xi_2,\cdots,\xi_m$ 张成的空间称为原空间的子空间，PCA 实际上是在子空间上的投影，并且

$$\|\mathbf{x}_1-\mathbf{x}_2\|^2=\left\|\sum_{i=1}^{n}y_{1i}\xi_i-\sum_{i=1}^{n}y_{2i}\xi_i\right\|^2\approx\left\|\sum_{i=1}^{m}y_{1i}\xi_i-\sum_{i=1}^{m}y_{2i}\xi_i\right\|^2.$$

7.6 PCA 特征提取方法与K-L变换

- **K-L 变换的性质**
  - 变换后空间中的各特征是互不相关的

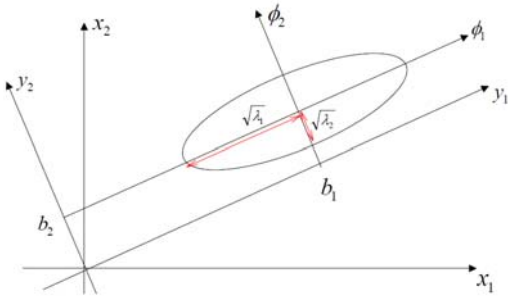
$$E\left[y_iy_j\right]=E\left[\xi_i^T\mathbf{x}\mathbf{x}^T\xi_j\right]=\lambda_i\xi_i^T\xi_j=\lambda_i\delta_{ij},$$

$$E\left[\mathbf{y}\mathbf{y}^T\right]=E\left[\xi^T\mathbf{x}\mathbf{x}^T\xi\right]=\xi^T\Sigma\xi=\Lambda;\qquad \Lambda=\begin{bmatrix}\lambda_1 & & 0\\ & \lambda_2 & \\ & \ddots & \\ 0 & & \lambda_d\end{bmatrix};$$

➔K-L 坐标系把矩阵 Σ 对角化，即通过 K-L 变换消除原有向量 $\mathbf{x}$ 的各分量间的相关性，从而有可能去掉那些带有较少信息的分量以达到降低特征维数的目的。

7.6 PCA 特征提取方法与K-L变换

- 例子



7.6 PCA 特征提取方法与K-L变换

- **K-L 变换的产生矩阵**
  - 数据集 $K_N=\{\mathbf{x}_i\}$ 的 K-L 变换的产生矩阵由数据的二阶统计量决定，即 K-L 坐标系的基向量为某种基于数据 $\mathbf{x}$ 的二阶统计量的产生矩阵的特征向量。
  - K-L 变换的产生矩阵可以有多种选择:
    - $\mathbf{x}$ 的相关函数矩阵 $R=E[\mathbf{x}\mathbf{x}^T]$;
    - $\mathbf{x}$ 的协方差矩阵 $C=E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$;
    - 样本总类内离散度矩阵:

$$S_w = \sum_{i=1}^{c} P_i \boldsymbol{\Sigma}_i, \quad \boldsymbol{\Sigma}_i = E\left[(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T\right], \ \mathbf{x} \in \omega_i.$$

7.6 PCA 特征提取方法与K-L变换

- PCA 的问题
  - 由于用样本协方差矩阵代替协方差矩阵，主分量与训练数据有着很大关系，用一批训练数据得到的主成分，可能不反映其另外一批数据的特征。

7.6 PCA 特征提取方法与K-L变换

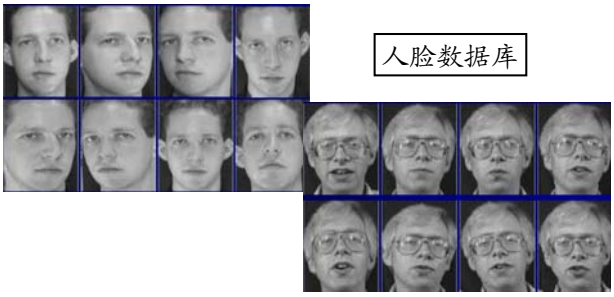
- **非监督的 K-L 特征提取**
  - 训练样本的类别未知，无法定义可分离性指标;
  - 可用方差作衡量指标 – 选择或提取总体未知样本方差越大，越有利于分类。
  - 可用总体协方差矩阵作为 K-L 产生矩阵:

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^T, \text{其中 } \mathbf{m} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i;$$

    - 把特征值从大到小排序 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$，选前 m 个对应的特征向量组成特征提取器 ⇨ 在均方误差意义下，用 $m<n$ 维对 n 维样本空间的最佳表示。

7.6 PCA 特征提取方法与K-L变换

- K-L 变换在人脸识别中的应用 — **Eigenface**
  - 简介，详细内容阅读教材 ch9.9.

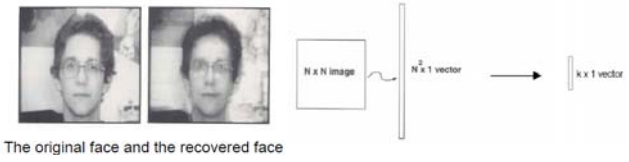


7.6 PCA 特征提取方法与K-L变换

- K-L 变换在人脸识别中的应用 — **Eigenface**

7.6 PCA 特征提取方法与K-L变换

- K-L 变换在人脸识别中的应用 — **Eigenface**
  - 把原图像表示成"特征脸"的线性组合（即特征脸空间中的点）;
  - 按照特征值从大到小排序，并从前向后取对应的"特征脸"，即构成对原图像的最佳的降维表示。



The original face and the recovered face

7.6 PCA 特征提取方法与K-L变换

- 基于 PCA 的人脸识别方法
  - 读取每个人的前 M 幅图像，构造矩阵 $\mathbf{X}$;
  - 计算: $\mathbf{\Sigma} = \mathbf{X}^T\mathbf{X}$;
  - 计算: $[\mathbf{V},\mathbf{D}] = svd(\mathbf{X})$;
  - 计算: $\xi = \mathbf{XVD}^{-1/2}$;
  - 按特征值从大到小排序，选择前几个最大的特征值对应的 ξ_i 作为变换矩阵 $\mathbf{W}$。
  - 把所有训练样本做变换 $\mathbf{y}=\mathbf{W}^T\mathbf{x}$，保留系数 $\mathbf{y}$。
  - 对新样本也作变换，看与哪个 $\mathbf{y}$ 最接近。
  - 与实际比较确定是否识别正确，统计识别率。

$$\mathbf{X} = [\mathbf{x}_1 - \boldsymbol{\mu}, \cdots, \mathbf{x}_M - \boldsymbol{\mu}];$$

7.7 特征选择

- **特征选择**：从原始特征中挑选出一些最有代表性、分类性能最好的特征来进行分类；
- 每个特征的状态是离散的：选或不选；
- 从 N 个特征中选取 k 个，共 C_N^k 种组合；如不限定个数，则共有 2^N 种组合 —— 典型的优化组合问题。

7.7 特征选择

- 特征选择的方法大体可分两大类：
  - **Filter 方法**：不考虑所使用的学习算法。通常给出一个独立于分类器的指标 J 来评价所选择的特征子集 S，然后在所有可能的特征子集中搜索出使得 J 最大的特征子集作为最优特征子集。
  - **Wrapper 方法**：将特征选择和分类器结合在一起，即特征子集的好坏标准是由分类器决定的，在学习过程中表现优异的的特征子集会被选中。

7.7 特征选择

- 一种 Filter 算法：FOCUS
  - 该算法致力于寻找一个能正确区分所有类别的最小特征集合；
  - 例如：如区分每个人的特征包括姓名、性别、籍贯、工作单位、身份证号……则该算法将选择：身份证号。
  - 搜索时会检测一个特征能否正确区分样本；如不能，则考察两个特征……以此类推。

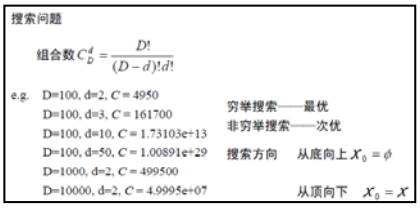
7.7 特征选择

- 一种 Wrapper 算法：OBLIVION
  - 该方法与最近邻法结合，根据特征子集的分类表现来选择特征;
  - 用**顺序后退法**搜索特征子集:

    从全体特征开始，每次剔除一个特征，使得所保留的特征集合有最大的分类识别率（基于最近邻法）。依次迭代，直至识别率开始下降为止。
  - 用 leave-one-out 方法估计平均识别率：用 $N-1$ 个样本判断余下一个的类别，N 次取平均。

7.7 特征选择

- 许多特征选择算法力求解决搜索问题，经典搜索算法有：
  - 分支定界法 ：最优搜索，效率比盲目穷举法高;
  - 单独最优特征组合法：次优搜索;
  - 顺序前进法
  - 顺序后退法
  - 模拟退火法
  - Tabu 搜索法
  - 遗传算法



7.7 特征选择

- **单独最优特征组合**
  - 计算各特征单独使用时的可分性判据并加以排序，取前 d 个作为选择结果。
  - 不一定是最优结果；
  - 当可分性判据对各特征具有（广义）可加性，该方法可以选出一组最优的特征；例如：
    - 各类具有正态分布；
    - 各特征统计独立；
    - 可分性判据基于 Mahalanobis 距离；

$$J_{ij}(x_1,x_2,\cdots,x_d)=\sum_{k=1}^{d}J_{ij}(x_k),\ J_D(\mathbf{x})=(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j).$$

7.7 特征选择

- **顺序前进法**
  - 自下而上的搜索方法；
  - 每次从未入选的特征中选择一个特征，使得它与已入选的特征组合在一起时所得的准则值为最大，直至特征数目增加到 d 为止；
  - 该方法考虑了所选特征与已入选特征之间的相关性。

7.7.1 特征选择 — 遗传算法

- 该算法受进化论启迪，根据“物竞天择，适者生存”规则演变；
- 术语：
  - 基因链码：使用遗传算法时要把问题的每个解编码成一个基因链码。

    比如要从 D 个特征中挑选 d个，就用一个 D 位的 0 或 1 组成的字符串表示一种特征组合； 1 表示该特征被选中。

    每个基因链码代表一个解，称作一个“个体”，其中的每一位看作是一个“基因”。

7.7.1 特征选择 — 遗传算法

- 术语
  - 群体：若干个体的集合，也就是一些解的集合；
  - 交叉：选择群体中的两个个体，以这两个个体为双亲作基因链码的交叉，从而产生两个新的个体，作为后代；

  - 变异：对某个体，随机选取其中一位，将其翻转

  - 适应度：对每个解，以给定的优化准则来评价其性能的优劣，作为其适应度。

7.7.1 特征选择 — 遗传算法

- **遗传算法的基本框架**

1. 初始化进化代数 t=0；
2. 给出初始化群体 $P(t)$，并令 $\mathbf{x}_g$ 为任意一个体；
3. 对 $P(t)$ 中每个个体估值，并将群体中最优解 $\mathbf{x}'$ 与 $\mathbf{x}_g$ 比较，若优于 $\mathbf{x}_g$，则令 $\mathbf{x}_g=\mathbf{x}'$；
4. 如果终止条件满足，则算法结束，$\mathbf{x}_g$ 为最终结果。否则，转步骤 5；
5. 从 $P(t)$ 选择个体并进行交叉和变异操作，得到新一代个体 $P(t+1)$，令 $t=t+1$，转步骤 3。

7.7.1 特征选择 — 遗传算法

- 关于遗传算法的说明
  - 由步骤 3 保证了最终解是所搜索过的最优解；
  - 常用的终止条件是群体的世代数超过一个给定值，或连续数个世代都没有得到更优解；
  - 群体的大小和演化代数是值得重视的参数；在一定范围内，这两个参数大些能得到更好的解；
  - 对交叉的亲本选择可采用如下规则：个体的性能越好，被选中的可能性也越大。

7.7.2 特征选择举例

- 例：用于癌症分类的基因选择
  - 根据癌症患者与正常人的基因表达数据，挑选出与癌症相关的基因；
  - 这种相关基因很可能就是治病基因，它可以帮助我们查找病源，进而可以指导设计药物；
  - 在选出的基因上作病患识别，可以提高识别率，有助于临床诊断。

7.7.2 特征选择举例

- 基因选择的困难
  - 人类大约有 3 万个左右的基因，但与某种疾病有关的基因不多；
  - 基因数（成千上万）远远大于实验样本数（几十）。

7.7.2 特征选择举例

- **单基因选择算法**
  - 基于某种准则给每个基因打分，把得分低的基因滤掉，选取那些得分高的基因组成特征子集。
  - 如 G-S 算法：以 Fisher 判别指标对每个特征打分，即根据每维特征上两类的距离和访查来评价该特征的分类能力：
    - 准则函数：$GS\_correlation(gene\_g)=\left|\frac{\mu_1^g-\mu_2^g}{\sigma_1^g+\sigma_2^g}\right|$

      其中 $\mu_1^g,\sigma_1^g,\mu_2^g,\sigma_2^g$ 分别是基因 g 在训练样本中第一类和第二类的均值和标准差。
    - 在分类时，该分值可作为每个特征的分类权重。

7.7.2 特征选择举例

- **多基因选择算法**
  - 与分类器相联系，采用各种搜索算法或优化算法进行特征选择；
  - SVM Recursive Feature Elimination (SVM-RFE) 算法：根据训练得到的 SVM 线形分类器的系数来判断每个特征分量的重要性和分类能力，假设由 SVM 得到的分类器为 $f(\mathbf{x})=\sum w_i x_i+b$,
    - 当 w_i 较大时，第 i 个特征对分类器的影响较大；
    - 当 w_i 较小时，第 i 个特征对分类器的影响较小；
    - 当 w_i 为0时，第 i 个特征对分类器几乎没有影响。