IPL

计算机学院

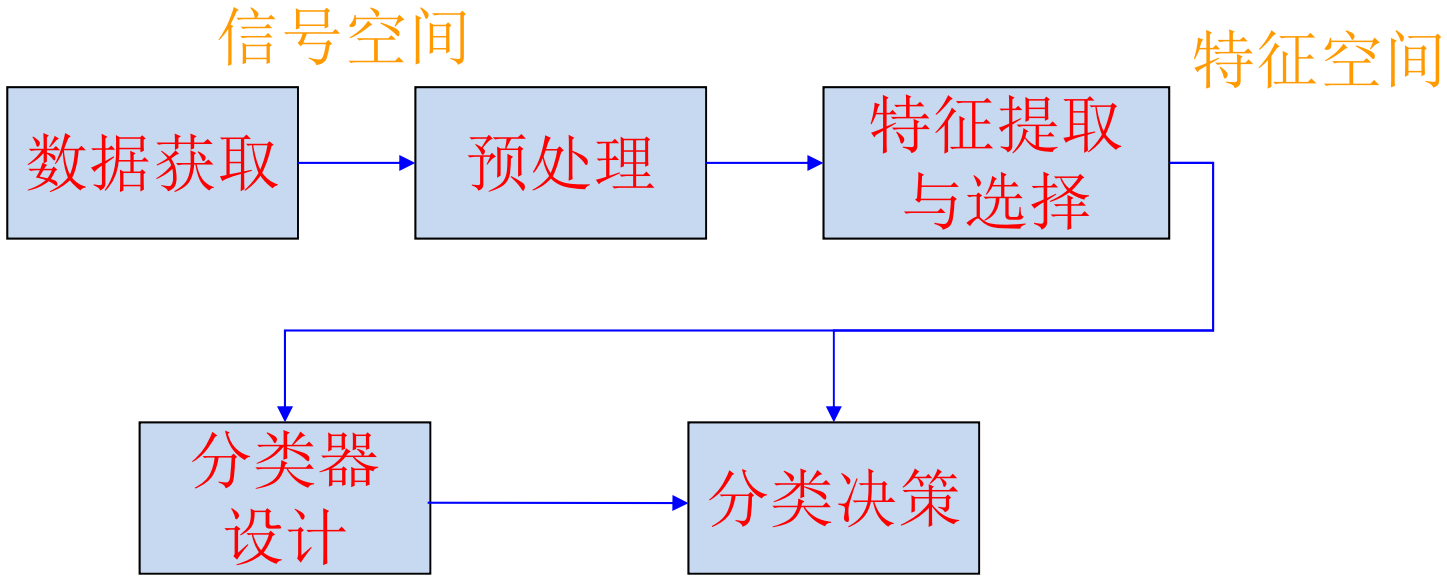
第六章 特征的选择与提取

孔万增 Kong Wanzeng, Ph.D
Tel: 15967146928
Email: kongwanzeng@hdu.edu.cn

Table of Contents

6.1 基本概念

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。

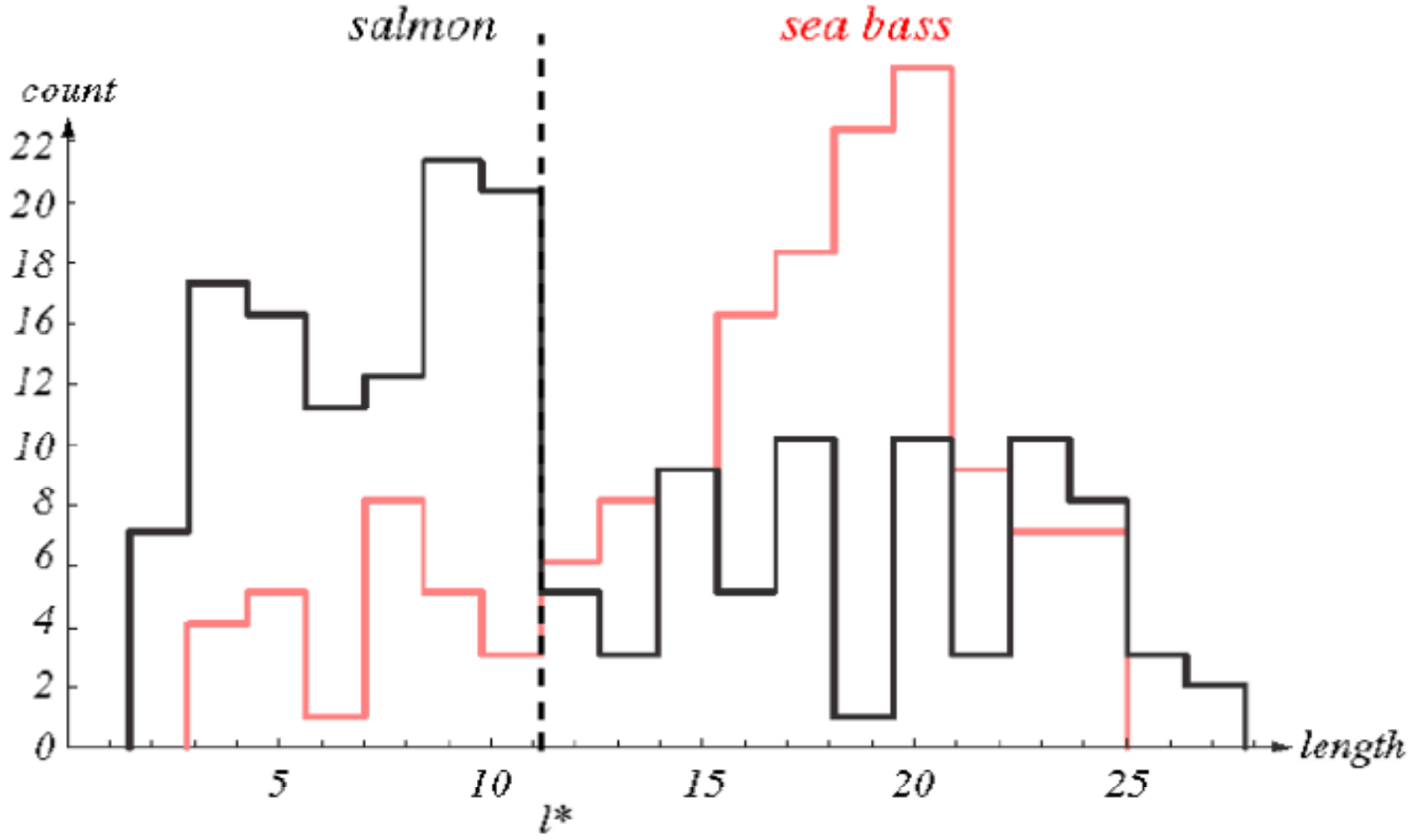
三大类特征

- 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。

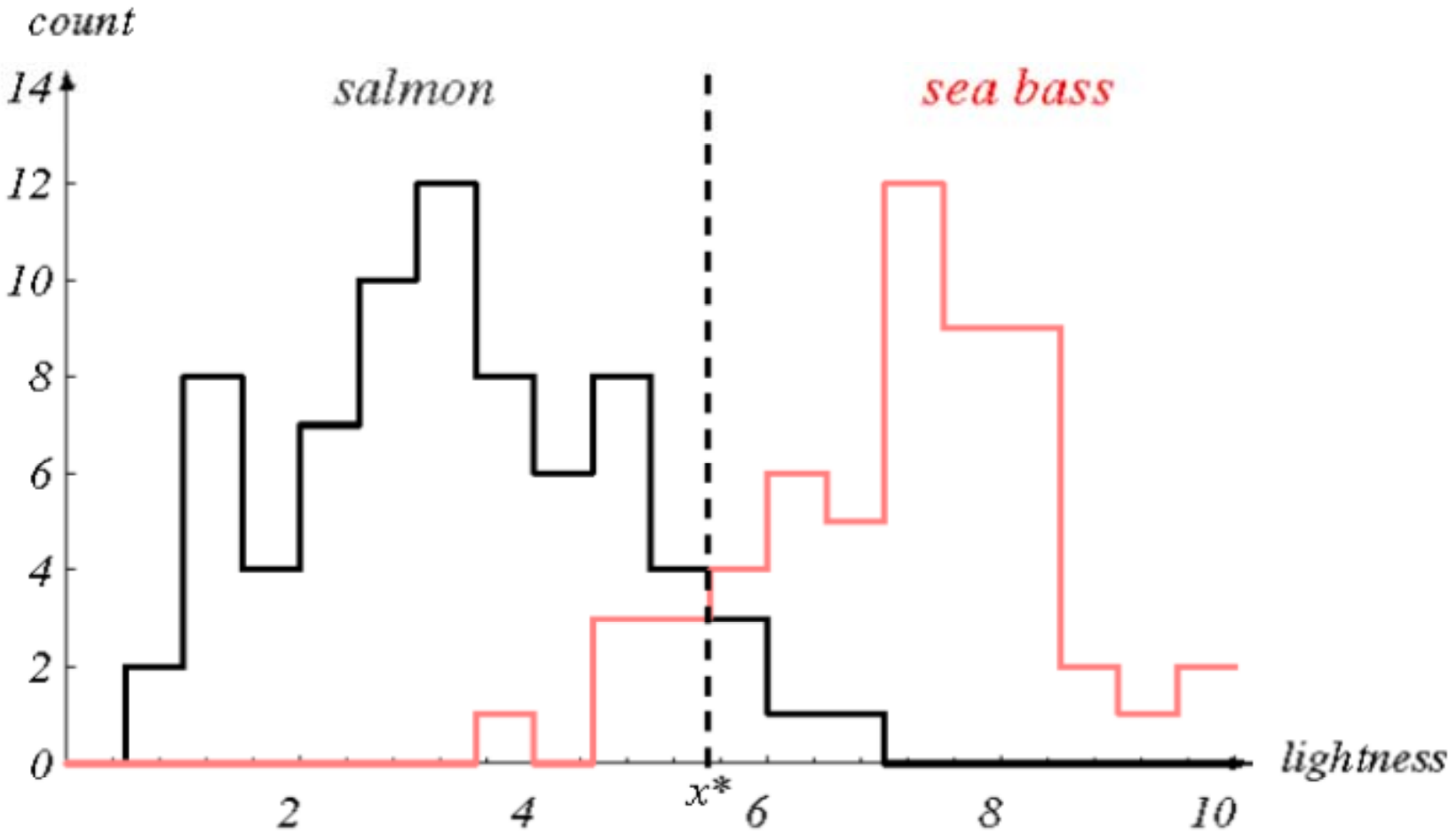
一个例子：鱼分拣

- 两类鱼
  - Sea bass
  - Salmon
- Pattern Classification, 2001

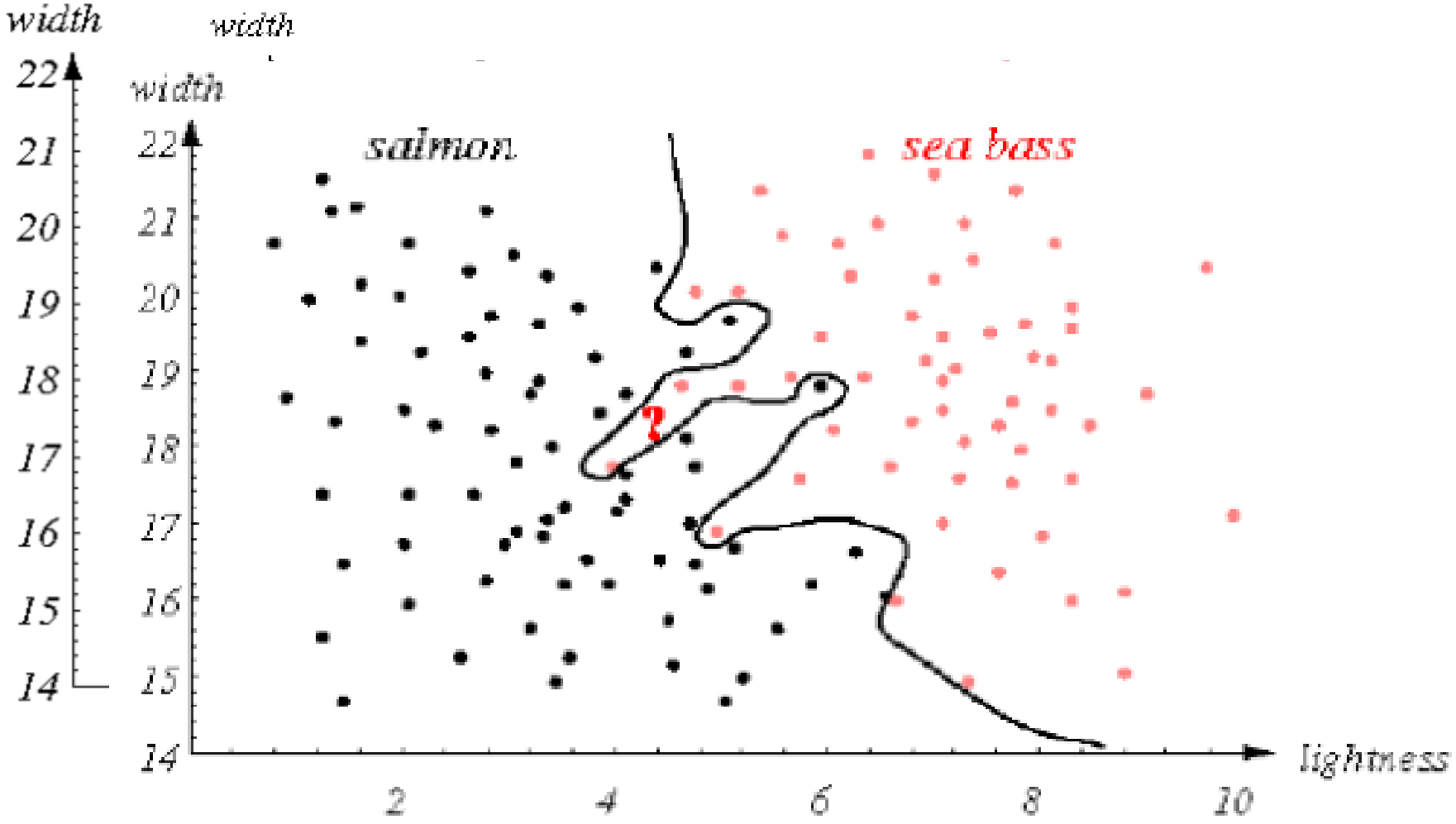
特征1：长度


salmon
sea bass
count
22
20
18
16
12
10
8
6
4
2
0
5
10
15
20
25
length
l*

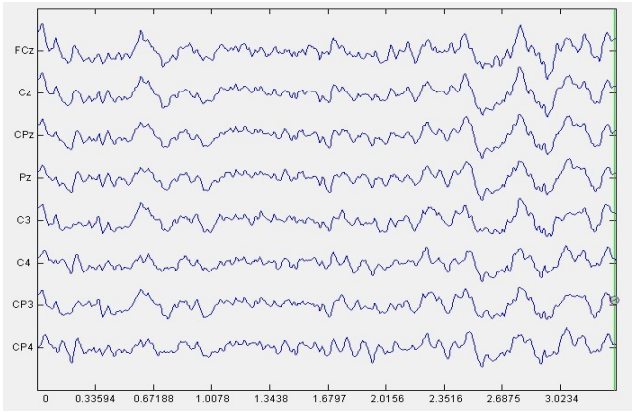
特征2：亮度

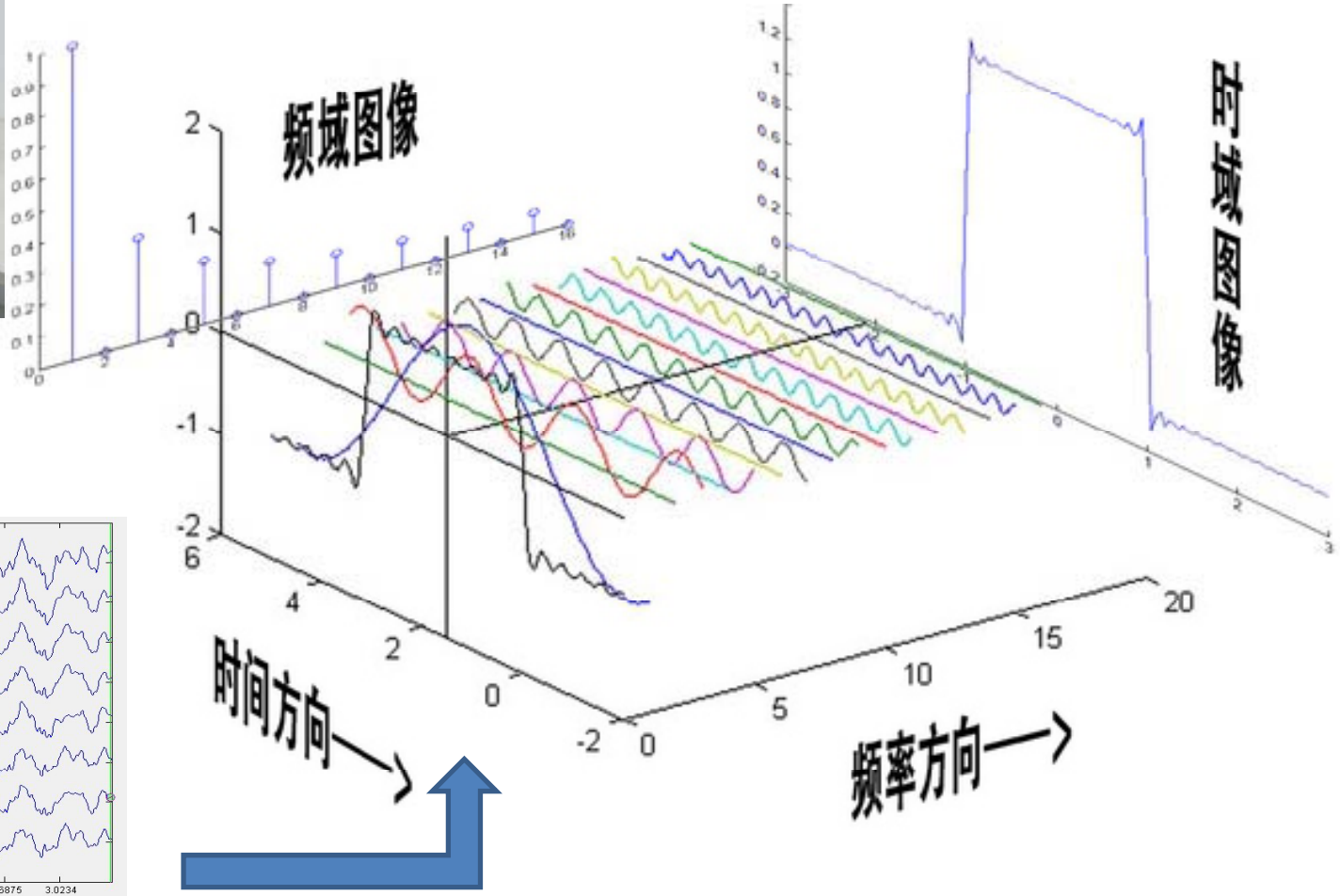


模式分类：线性、二次、最近邻分类器

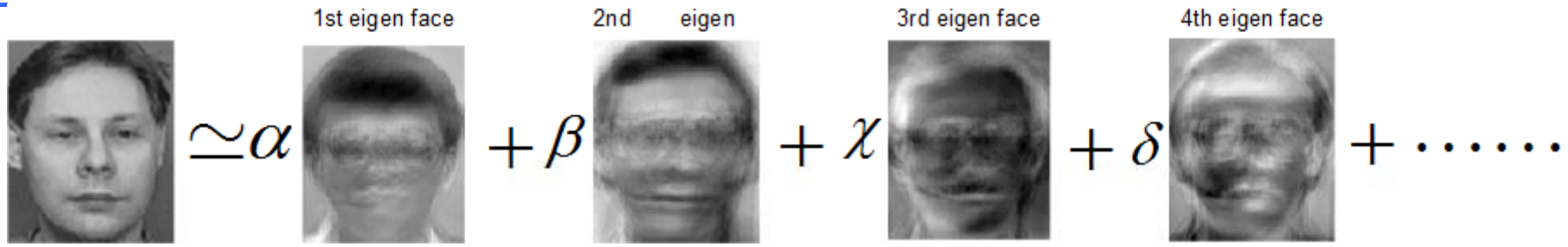


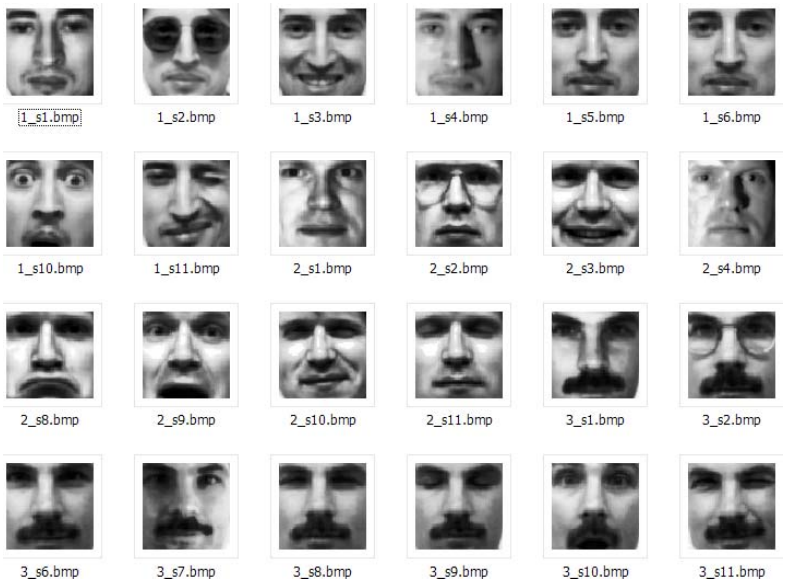
特征提取之傅里叶变化（抽象）



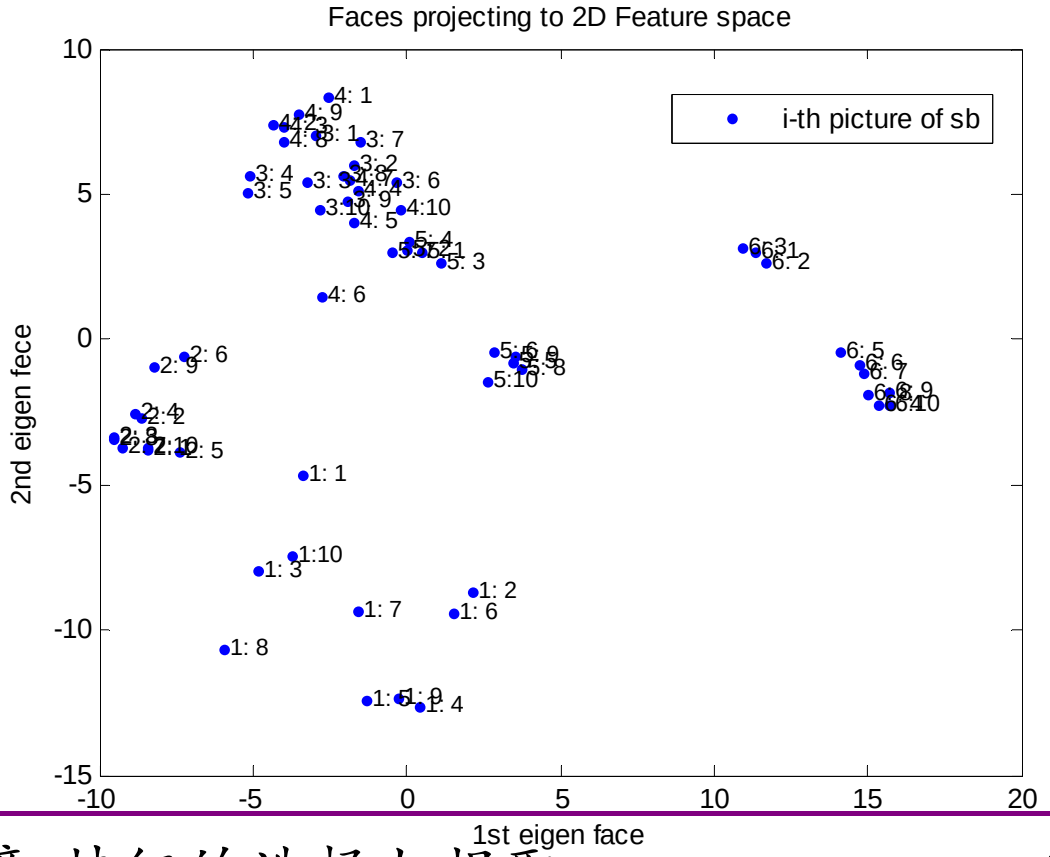


人脸识别中的特征提取（本征脸）





人脸库部分样本



特征的形成

- 特征形成 (acquisition)：
  - 信号获取或测量→原始测量
  - 原始特征
- 实例：
  - 数字图象中的各像素灰度值
  - 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。

特征的选择与提取

- 两类提取有效信息、压缩特征空间的方法：特征提取和特征选择
- **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。
- **特征选择**(selection)：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。
- 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。

特征的选择与提取举例

- 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，核浆比
  - 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：专家知识，数学方法
    - 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

6.2 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：某组特征使分类器的错误率最小
- 实际的类别可分离性判据应满足的条件：
  - 度量特性：$J_{ij} > 0, \text{ if } i \neq j;\ J_{ij} = 0, \text{ if } i = j;\ J_{ij} = J_{ji}$
  - 与错误率有单调关系
  - 当特征独立时有可加性：$J_{ij}(x_1, x_2, \ldots, x_d) = \sum_{k=1}^{d} J_{ij}(x_k)$
  - 单调性：$J_{ij}(x_1, x_2, \ldots, x_d) \leq J_{ij}(x_1, x_2, \ldots, x_d, x_{d+1})$
- 常见类别可分离性判据：基于距离、概率分布、熵函数

基于距离的可分性判据

可分性判据

- 类间可分性:=所有样本间的平均距离：

$$J_d(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j \frac{1}{n_i n_j} \sum_{k=1}^{n_i}\sum_{l=1}^{n_j} \delta(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)}) \qquad (8\text{-}1)$$

$$\delta(\mathbf{x}_k^{(i)}, x_l^{(j)}) = (\mathbf{x}_k^{(i)} - x_l^{(j)})^T (\mathbf{x}_k^{(i)} - x_l^{(j)})$$

squared Euclidian

$$\mathbf{m}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} \mathbf{x}_k^{(i)} \qquad \mathbf{m} = \sum_{i=1}^{c} P_i \mathbf{m}_i$$

类内平均距离

类间距离

$$J_d(\mathbf{x}) = \sum_{i=1}^{c} P_i \left[\frac{1}{n_i}\sum_{k=1}^{n_i} \delta(\mathbf{x}_k^{(i)}, \mathbf{m}_i) + \delta(\mathbf{m}_i, \mathbf{m}) \right] \qquad (8\text{-}5)$$

$$\sum_{i=1}^{c} P_i \delta(\mathbf{m}_i, \mathbf{m}) = \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j \delta(\mathbf{m}_i, \mathbf{m}_j) \qquad (8\text{-}6)$$

基于距离的可分性判据矩阵形式

可分性判据

样本类间离散度矩阵

$$\tilde{S}_b = \sum_{i=1}^{c} P_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$$

样本类内离散度矩阵

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

类间可分离性判据

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

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

特征可分性评价判据

可分性判据

FEATEVAL Evaluation of feature set for classification

J = FEATEVAL(A,CRIT,T) J = FEATEVAL(A,CRIT,N)

A input dataset

CRIT string name of a method or untrained mapping

T validation dataset (optional)

N number of cross-validations (optional)

DESCRIPTION

Evaluation of features by the criterion CRIT for classification, using objects in the dataset A. The larger J, the better. Resulting J-values are incomparable over the various methods.

基于概率的可分性判据

- 基于概率的可分性判据：用概率密度函数间的距离（交叠程度）来度量

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

- 散度：区分i，j两类总的平均信息

$$J_D(\mathbf{x}) = I_{ij} + I_{ji} = \int_{\mathbf{x}}\left[p(\mathbf{x}\mid\omega_i) - p(\mathbf{x}\mid\omega_j)\right]\ln\frac{p(\mathbf{x}\mid\omega_i)}{p(\mathbf{x}\mid\omega_j)}d\mathbf{x}$$

$$l_{ij}(\mathbf{x}) = \frac{p(\mathbf{x}\mid\omega_i)}{p(\mathbf{x}\mid\omega_j)} \qquad I_{ij}(\mathbf{x}) = E\left[l_{ij}(\mathbf{x})\right] = \int_{\mathbf{x}} p(\mathbf{x}\mid\omega_i)\ln\frac{p(\mathbf{x}\mid\omega_i)}{p(\mathbf{x}\mid\omega_j)}d\mathbf{x}$$

正态分布条件下的散度判据

可分性判据

- 正态分布条件下的散度判据可以用分布参数表示，特别是

$$\text{if } \omega_i \sim N(\boldsymbol{\mu}_i, \Sigma_i), \omega_j \sim N(\boldsymbol{\mu}_j, \Sigma_j), \Sigma_i = \Sigma_j = \Sigma$$

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

Mahalanobis

- 一维正态分布：

$$J_D(x) = \frac{(\mu_i - \mu_j)^2}{\sigma^2}$$

基于熵函数的可分性判据

可分性判据

- 熵函数：衡量后验概率分布的集中程度

$$H = J_c\left[P(\omega_1 \mid \mathbf{x}),\ldots,P(\omega_c \mid \mathbf{x})\right]$$

- Shannon熵：

$$J_c^1 = -\sum_{i=1}^{c} P(\omega_i \mid \mathbf{x}) \log_2 P(\omega_i \mid \mathbf{x})$$

- 平方熵：

$$J_c^2 = 2\left[1 - \sum_{i=1}^{c} P^2(\omega_i \mid \mathbf{x})\right]$$

- 熵函数期望表征类别的分离程度：

$$J(\bullet) = E\left\{J_c\left[P(\omega_1 \mid \mathbf{x}),\ldots,P(\omega_c \mid \mathbf{x})\right]\right\}$$

类别可分离性判据应用举例

可分性判据

- 图像分割：Otsu灰度图像阈值算法 (Otsu thresholding)
- 图像有L阶灰度，n_i是灰度为i的像素数，图像总像素数 $N= n_1+n_2+ \ldots + n_L$
  - 灰度为i的像素概率：$p_i = n_i/N$
  - 类间方差：

$$\mu_1 = \sum_{i=1}^{k} i p_i, \mu_2 = \sum_{i=k+1}^{L} i p_i, \mu = \sum_{i=1}^{L} i p_i$$

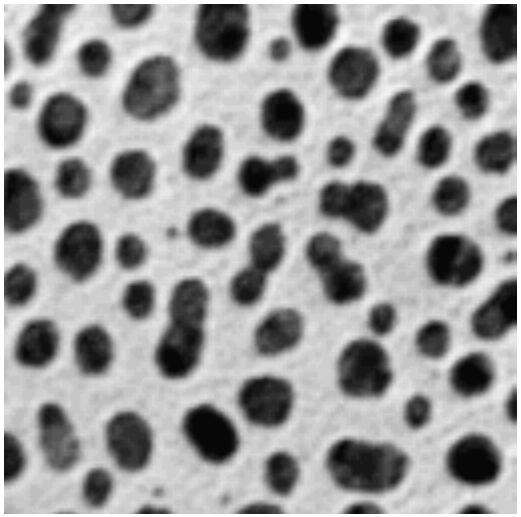
$$\omega_1 = \sum_{i=1}^{k} p_i, \omega_2 = \sum_{i=k+1}^{L} p_i = 1 - \omega_1$$

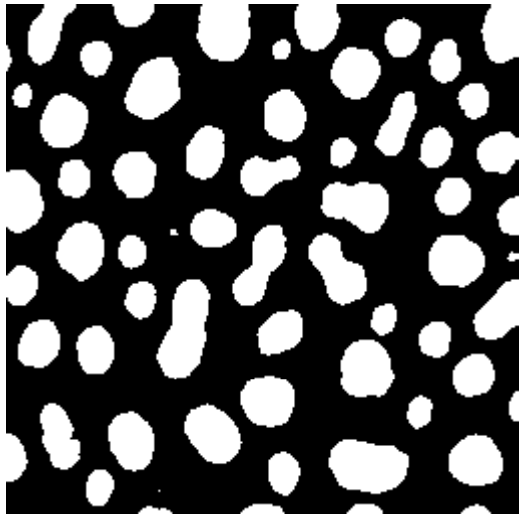
$$\sigma_B^2(k) = \omega_1(\mu_1 - \mu)^2 + \omega_2(\mu_2 - \mu)^2$$

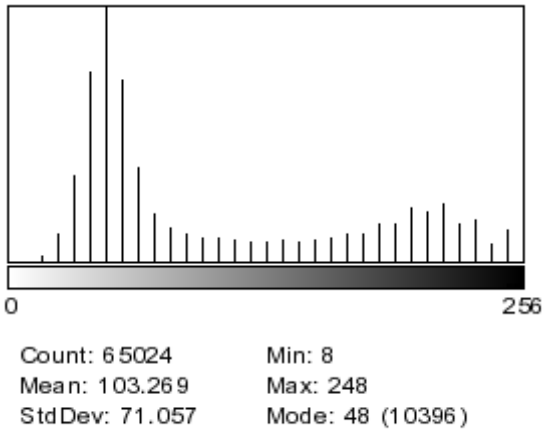
Otsu thresholding

可分性判据

- 灰度图像阈值: $t = \arg\max_{k=1}^{L} \sigma_B^2(k)$
- Otsu灰度图像二值化算法演示及程序分析:







6.3 特征提取与K-L变换

- **特征提取**：用映射（或变换）的方法把原始特征变换为较少的新特征 $J(\mathbf{x}^*) = \max_{\mathbf{x}} J(\mathbf{x})$
- 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。
- 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

K-L变换

- 离散K-L变换：对向量**x**用确定的完备正交归一向量系$\mathbf{u}_j$展开

$$\mathbf{x} = \sum_{j=1}^{\infty} y_j \mathbf{u}_j$$

$$\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$$

$$\mathbf{x} \rightarrow \mathbf{y} \qquad y_j = \mathbf{u}_j^T \mathbf{x}$$

离散K-L变换的均方误差

- 用有限项估计$\mathbf{x}$：$\hat{\mathbf{x}} = \sum_{j=1}^{d} y_j \mathbf{u}_j \qquad y_j = \mathbf{u}_j^T \mathbf{x}$
- 该估计的均方误差：$\varepsilon = E\left[(\mathbf{x} - \hat{\mathbf{x}})^T (\mathbf{x} - \hat{\mathbf{x}})\right]$

$$\varepsilon = E\left[\sum_{j=d+1}^{\infty} y_j^2\right] = E\left[\sum_{j=d+1}^{\infty} \mathbf{u}_j^T \mathbf{x}\mathbf{x}^T \mathbf{u}_j\right]$$

$$\mathbf{R} = \left[r_{ij} = \mathrm{E}(x_i x_j)\right] = E\left[\mathbf{x}\mathbf{x}^T\right]$$

$$\varepsilon = \sum_{j=d+1}^{\infty} \mathbf{u}_j^T E\left[\mathbf{x}\mathbf{x}^T\right]\mathbf{u}_j = \sum_{j=d+1}^{\infty} \mathbf{u}_j^T \mathbf{R}\,\mathbf{u}_j$$

求解最小均方误差正交基

- 用Lagrange乘子法：

$$\text{if } \mathbf{R}\mathbf{u}_j = \lambda_j \mathbf{u}_j \text{ then } \varepsilon = \sum_{j=d+1}^{\infty} \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j \text{取得极值}$$

- 结论：以相关矩阵R的d个本征向量为基向量来展开x时，其均方误差为：

$$\varepsilon = \sum_{j=d+1}^{\infty} \lambda_j$$

- K-L变换：当取矩阵R的d个最大本征值对应的本征向量来展开x时，其截断均方误差最小。这d个本征向量组成的正交坐标系称作x所在的D维空间的d维K-L变换坐标系，x在K-L坐标系上的展开系数向量y称作x的K-L变换

K-L变换的表示

- K-L变换的向量展开表示：

$$\mathbf{x} = \sum_{j=1}^{d} y_j \mathbf{u}_j \qquad y_j = \mathbf{u}_j^T \mathbf{x}$$

- K-L变换的矩阵表示：

$$\mathbf{x} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d]\mathbf{y} = \mathbf{U}\mathbf{y}$$

$$\mathbf{y} = \mathbf{U}^T \mathbf{x}$$

K-L变换的性质

- y的相关矩阵是对角矩阵：

$$E\left[y_i y_j\right] = E\left[\mathbf{u}_i^T \mathbf{x}\mathbf{x}^T \mathbf{u}_j\right] = \mathbf{u}_i^T E\left[\mathbf{x}\mathbf{x}^T\right]\mathbf{u}_j$$

$$= \mathbf{u}_i^T R\,\mathbf{u}_j = \mathbf{u}_i^T \lambda_j \mathbf{u}_j = \lambda_i \delta_{ij}$$

$$R_{\mathbf{y}} = E\left[\mathbf{y}\mathbf{y}^T\right] = E\left[U^T \mathbf{x}\mathbf{x}^T U\right]$$

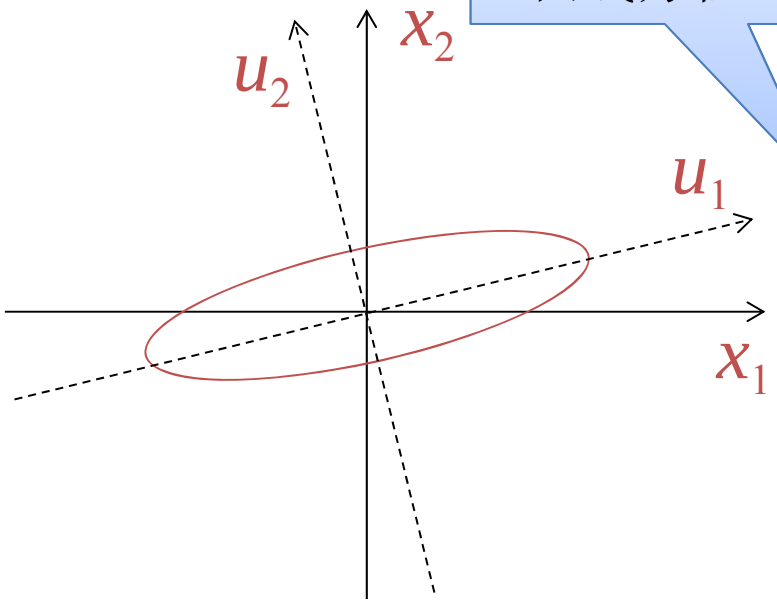
$$= U^T \mathbf{R}\, U = \Lambda$$

K-L变换的性质

- K-L坐标系将相关函数矩阵对角化，即通过K-L变换消除原有向量**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。

$$R_y = \Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_d \end{bmatrix}$$

K-L变换图解



二次曲线方程

$$f(x_1, x_2, \cdots, x_n)$$

$$= \sum_{i,j=1}^{n} r_{ij} x_i x_j$$

$$\overset{\mathbf{x}=U\mathbf{y}}{=} \mathbf{x}'\mathbf{R}\mathbf{x} = \mathbf{y}'(\mathbf{U}'\mathbf{R}\mathbf{U})\mathbf{y} = \mathbf{y}'\Lambda\mathbf{y}$$

$$= \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots \lambda_n y_n^2$$

标准二次曲线方程

K-L变换的数据压缩图解

- 取2x1变换矩阵U=[$\mathbf{u}_1$]，则$\mathbf{x}$的K-L变换$\mathbf{y}$为：

$$\mathbf{y} = \mathrm{U}^{\mathrm{T}}\mathbf{x} = \mathbf{u}_1{}^{\mathrm{T}}\mathbf{x} = y_1$$

- 变换的能量损失为

$$\frac{\lambda_2^2}{\lambda_1^2 + \lambda_2^2} = \frac{1}{4^2 + 1^2} = 5.9\%$$

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本征向量
- 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：

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

未知类别样本的K-L变换

特征提取

- 用总体样本的协方差矩阵C=E[(**x**-**μ**) (**x**-**μ**)^T]进行K-L变换，K-L坐标系U=[$\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_d$]按照C的本征值的下降次序选择
- 例：设一样本集的协方差矩阵是：$C=\begin{bmatrix}19.5 & 9.5\\ 9.5 & 7.5\end{bmatrix}$
  求最优2x1特征提取器U
  解答：计算特征值及特征向量[V, D]=eig(C);
  特征值D=[24.736, 2.263]^T,特征向量：$V=\begin{bmatrix}0.875 & -0.482\\ 0.482 & 0.875\end{bmatrix}$
  由于$\lambda_1>\lambda_2$，故最优2x1特征提取器 $U=[\mathbf{u}_1]=\begin{bmatrix}0.875\\ 0.482\end{bmatrix}$
  此时的K-L变换式为：

$$\mathbf{y}=U^T\mathbf{x}=\mathbf{u}^T\mathbf{x}=\begin{bmatrix}0.875 & 0.482\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}$$

6.4 特征的选择

- 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进行分类。
- 从*D*个特征中选取*d*个,共C^d_D种组合。若不限定特征选择个数，则共2^D种组合
  — 典型的组合优化问题
- 特征选择的方法：
  - 是否直接考虑分类器性能
    - Filter方法：根据独立于分类器的指标*J*来评价所选择的特征子集S，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。不考虑所使用的学习算法。
    - Wrapper方法：将特征选择和分类器结合在一起，在分类过程中表现优异的的特征子集会被选中。
  - 选择特征的顺序：
    - 自下而上：特征数从零逐步增加到d。
    - 自上而下：特征数从D开始逐步减少到d。

经典特征选择算法

特征选择

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

单独最优特征组合

- 计算各特征单独使用时的可分性判据*J*并加以排队，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来，例：
  - 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

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

[W,R] = FEATSELI(A,CRIT,K,T)

特征选择

INPUT

A Training dataset

CRIT Name of the criterion or untrained mapping
(default: 'NN', i.e. the 1-Nearest Neighbor error)

K Number of features to select (default: sort all features)

T Tuning dataset (optional)

OUTPUT

W Feature selection mapping

R Matrix with criterion values

DESCRIPTION

Individual selection of K features using the dataset A. CRIT sets the criterion used by the feature evaluation routine FEATEVAL. If the dataset T is given, it is used as test set for FEATEVAL. For K = 0 all features are selected, but reordered according to the criterion. The result W can be used for selecting features using B*W.

顺序前进法Sequential forward selection

特征选择

- 自下而上搜索方法。
- 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可分性或分类识别率为最大，直至特征数增加到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。

[W,R] = FEATSELF(A,CRIT,K,T,FID)

Forward selection of K features using the dataset A. CRIT sets the criterion used by the feature evaluation routine FEATEVAL.

顺序后退法Sequential backw. selection

特征选择

- 该方法根据特征子集的分类表现来选择特征
- 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可分性或分类识别率。
- 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。

[W,R] = FEATSELB(A,CRIT,K,T,FID)

Backward selection of K features using the dataset A. CRIT sets the criterion used by the feature evaluation routine FEATEVAL.

模拟退火法

- 来源于统计力学。材料粒子从高温开始，非常缓慢地降温(退火)，粒子就可在每个温度下达到热平衡。
- 假设材料在状态i的能量为$E(i)$，那么材料在温度T时从状态i进入状态j遵循如下规律：
  - 如果$E(j) \le E(i)$，接受该状态被转换。
  - 如果$E(j)>E(i)$，则状态转换以如下概率被接受：

$$e^{\frac{E(i)-E(j)}{KT}}$$

模拟退火法(II)

- 在某一温度下，进行了充分转换后，材料达到热平衡，这时材料处于状态i的概率满足：

$$P_T(x=i)=\frac{e^{-\frac{E(i)}{KT}}}{\sum_{j\in S} e^{-\frac{E(j)}{KT}}}$$

- 所有状态在高温下具有相同概率。

$$\lim_{T\to\infty}\frac{e^{-\frac{E(i)}{KT}}}{\sum_{j\in S} e^{-\frac{E(j)}{KT}}}=\frac{1}{|S|}$$

模拟退火法(III)

- 当温度降至很低时，材料会以很大概率进入最小能量状态。

$$\lim_{T\to 0}\frac{e^{-\frac{E(i)-E_{\min}}{KT}}}{\sum_{j\in S} e^{-\frac{E(j)-E_{\min}}{KT}}}=\begin{cases}\dfrac{1}{|S_{\min}|} & i\in S_{\min}\\ 0 & otherwise\end{cases}$$

- 模拟退火优化法：f: x→R^+，其中x∈S，表示优化问题的一个可行解。N(x)≤S表示x的一个邻域集合。

模拟退火法(IV)

- 首先给定初始温度T_0和初始解$x(0)$，以概率P生成下一个新解x'：

$$P(x(0 \to x')) = \begin{cases} 1 & f(x') < f(x(0)) \\ e^{-\frac{f(x')-f(x(0))}{T_0}} & otherwise \end{cases}$$

- 对于温度T_i和该优化问题的解$x(k)$，可以生成新解x'。
- 经过多次转换，降低温度得到$T_{i+1}<T_i$。在T_{i+1}下重复上述过程。
- 优化即是交替寻找新解和缓慢降低温度，最终的解是对该问题寻优的结果。

模拟退火法(V)

- 经过有限次转换，在温度T_i下的平衡态x_i的分布为：

$$P_i(T_i) = \frac{e^{-\frac{f(x_i)}{T}}}{\sum_{j \in S} e^{-\frac{f(x_j)}{T}}}$$

- 当温度T降为0时，x_i的分布为：

$$P_{i^*} = \begin{cases} \dfrac{1}{|S_{\min}|} & x_i \in S_{\min} \\ 0 & otherwise \end{cases} \qquad \sum_{x_i \in S_{\min}} P_{i^*} = 1$$

特征选择的模拟退火法

- Step1: 令i=0, k=0, 给出初始温度T_0和初始特征组合$x(0)$。
- Step2: 在$x(k)$的邻域$N(x(k))$中选择一个状态x'，即新特征组合。计算其可分性判据$J(x')$，并按概率P接受$x(k+1)=x'$。
- Step3: 如果在T_i下还未达到平衡，则转到Step2。
- Step4: 如果T_i已经足够低，则结束，当时的特征组合即为算法的结果。否则继续。
- Step5: 根据温度下降方法计算新的温度T_{i+1}。转到Step2。

遗传算法

- 从生物进化论得到启迪。遗传，变异，自然选择。基于该思想发展了遗传优化算法。
- 基因链码：待解问题的解的编码，每个基因链码也称为一个个体。对于特征选择，可用一个D位的0/1构成的串表示一种特征组合。
- 群体：若干个个体的集合，即问题的一些解的集合。
- 交叉：由当前两个个体的链码交叉产生新一代的个体。
- 变异：由一个链码随机选取某基因使其翻转。

遗传算法

- 适应度：每个个体x_i的函数值f_i，个体x_i越好，f_i越大。新一代群体对环境的平均适应度比父代高。
- 遗传算法的基本框架：

- Step1: 令进化代数t=0。
- Step2: 给出初始化群体P(t)，令x_g为任一个体。
- Step3: 对P(t)中每个个体估值，并将群体中最优解x'与x_g比较，如果x'的性能优于x_g，则x_g=x'
- Step4: 如果终止条件满足，则算法结束，x_g为算法的结果。否则继续。
- Step5: 从P(t)中选择个体并进行交叉和变异操作，得到新一代群体P(t+1)。令t=t+1,转到Step3。

6.5 讨论

- 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、结构和数学特征
  - 物理和结构特征：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

习题

1. 试推导(8-6)式，即：

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

2. 试由(8-1)式推导(8-5)式，即：

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

3. 习题8.1

9. 习题9.1