

模式识别

Pattern Recognition

第4讲 特征选择与特征提取

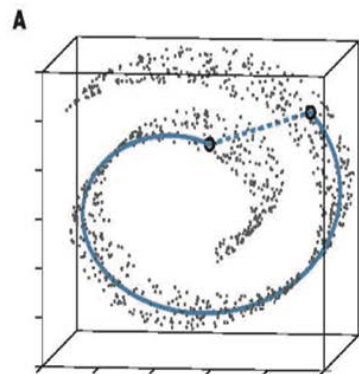
维数的诅咒 (Curse of Dimensionality)

维数增高带来困难：增大了分类学习过程和识别过程计算和存储的复杂程度，降低了分类器的效率；识别特征维数过大使得分类器过于复杂

维数 VS 样本数：在样本数量一定的条件下，估计参数的数量越少准确度越高，用少量样本估计过多的参数则是一个不可靠的过程！

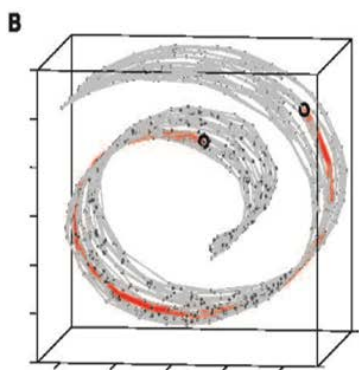
例：1维特征，估计均值方差（2个参数），20个样本具有一定可信度。
100维特征，估计均值协方差矩阵（5150个参数），需要多少样本？

降维(Dimensionality Reduction)



从高维空间映射到低维空间

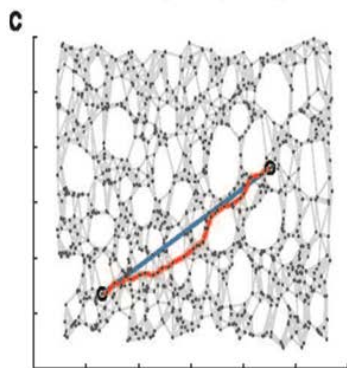
$$\mathbf{y} = f(\mathbf{x}), \quad \mathbf{x} \in R^m, \mathbf{y} \in R^{m'} \text{ 其中 } m' < m$$



降低计算的复杂度:

降低存储器的占用;

提高分类器的识别速度;



提高分类器的性能:

降低分类器的复杂度, 提高泛化能力;

有可能丢失可分性信息, 降低分类准确率;

降维的基本方法

如何选出有效特征？

从应用的角度
可以分为：

特征选择
特征提取

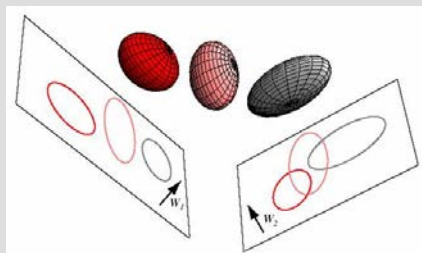
特征选择 (Feature Selection)：从原始的特征中直接挑选出对分类最有价值的特征

特征提取 (Feature Extraction) 将原始特征变换得到一组新的低维特征

如何计算出更有效特征？

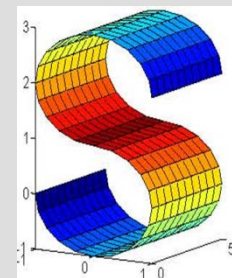
从数学的角度
可以分为：

线性降维
非线性降维

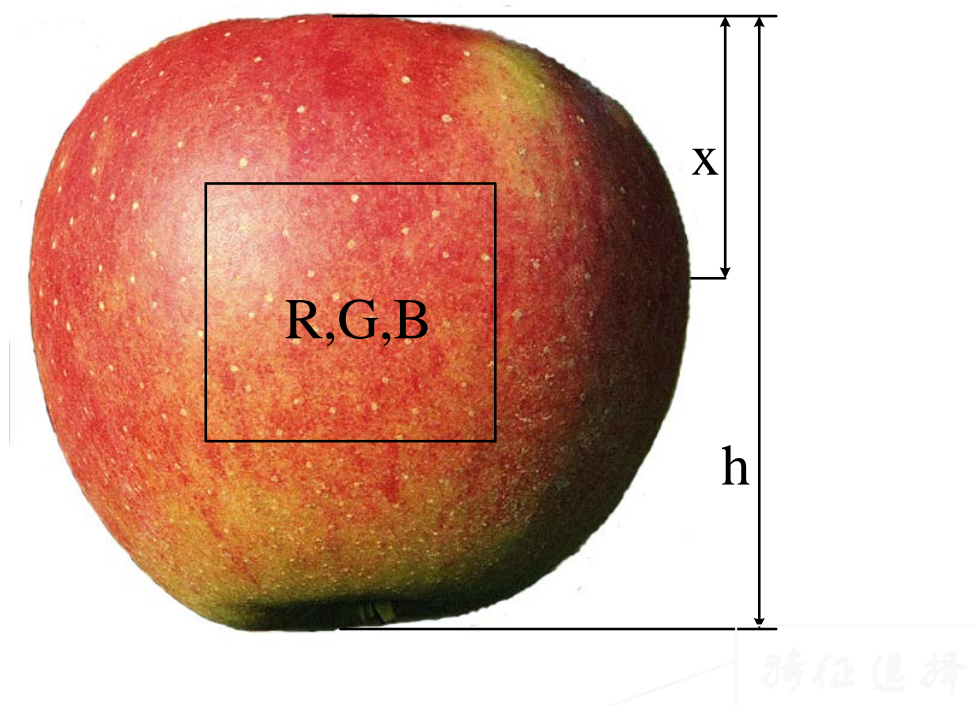


线性降维：样本分布在一个嵌入的子空间中， $y = f(x) = W^t x$ 是线性变换

非线性降维：样本分布在一个嵌入的非线性流形上， $y = f(x)$ 是非线性变换



特征选择、提取举例：



原始5维特征 (R, G, B, x, h)

蓝色分量B对分类作用低，去掉后得到4维特征 (R, G, x, h)

计算红绿比值和x、h比值得到2维特征 $(\frac{R}{G}, \frac{x}{h})$

特征提取

对于复杂问题如何选择和提取特征？

□特征选择（Feature Selection）

——判断特征对于分类的有效性

- 类别可分性判据
- 分支定界法

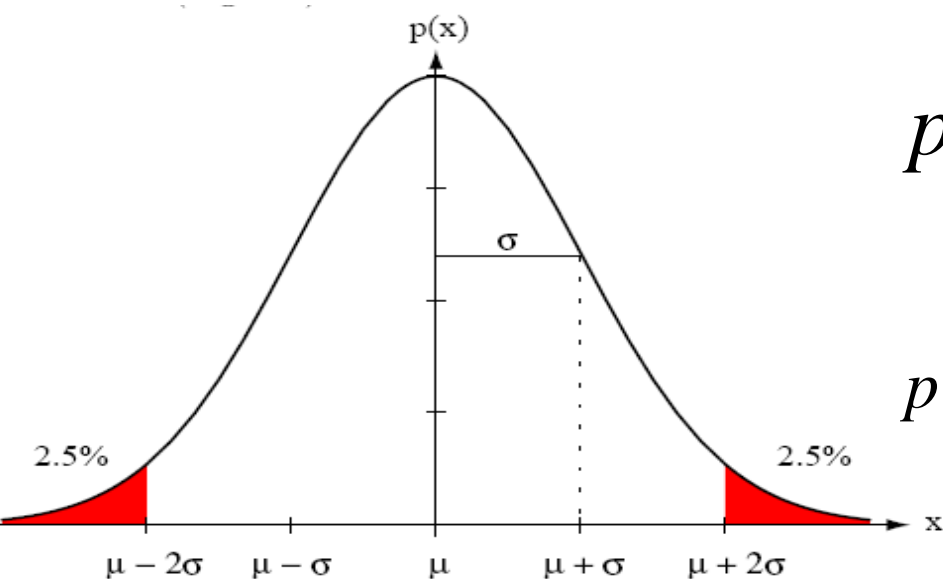
□特征提取

——根据各维特征间的统计关系，对特征进行变换

- 主成分分析
- Fisher判别分析

正态分布

- 单变量正态分布密度函数（高斯分布）：



$$p(x) \sim N(\mu, \sigma^2)$$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

多元正态分布函数

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

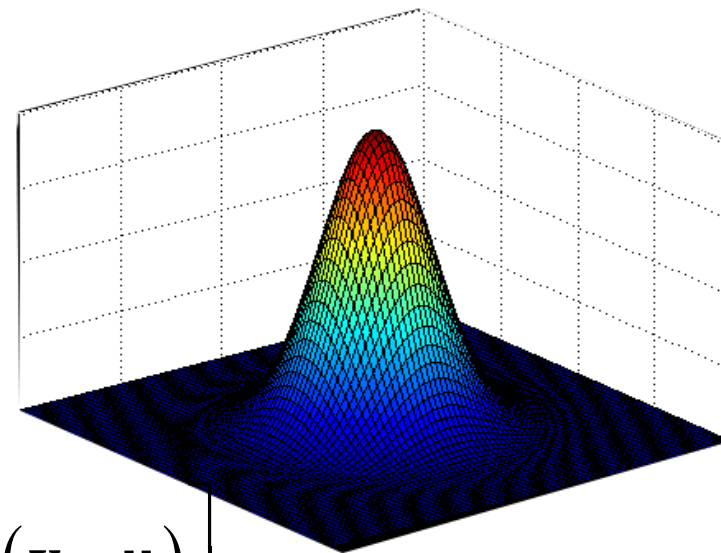
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

$$\text{均值: } \boldsymbol{\mu} = E(\mathbf{x}) = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

$$\mu_i = E(x_i)$$

$$\text{协方差矩阵: } \boldsymbol{\Sigma} = E\left((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t\right) = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t p(\mathbf{x}) d\mathbf{x}$$

$$\sigma_{ij} = E\left((x_i - \mu_i)(x_j - \mu_j)\right)$$



多元正态分布函数

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

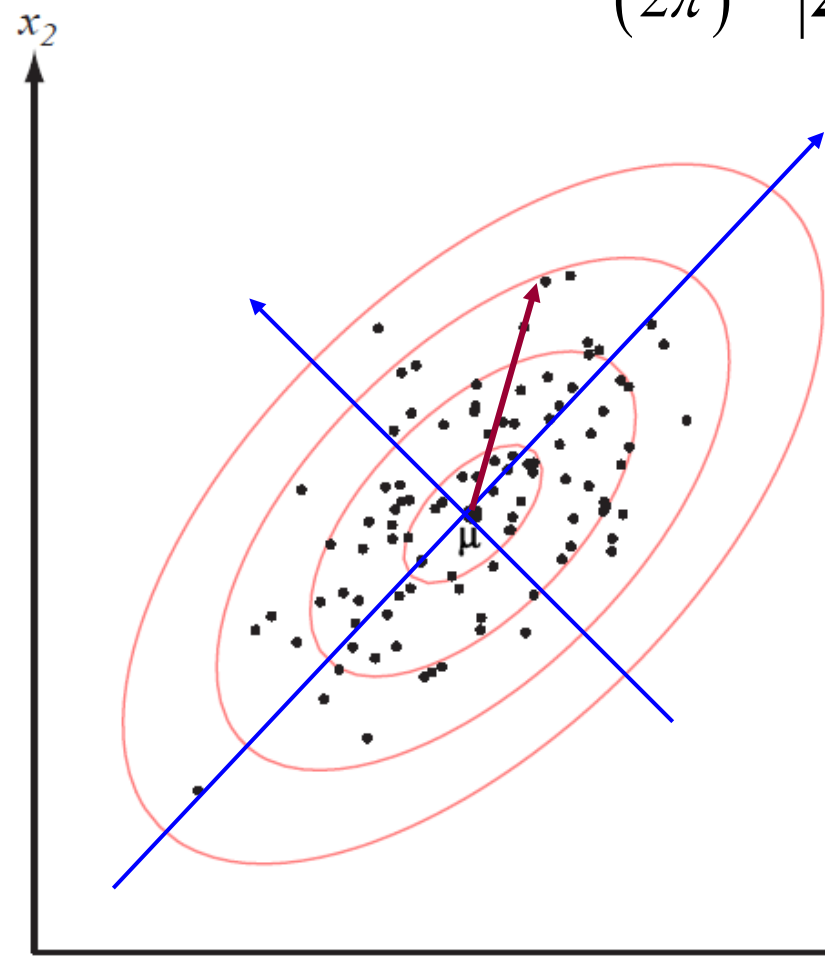
$$r = -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

$$V = V_d |\Sigma|^{1/2} r^d$$

$$V_d = \begin{cases} \pi^{d/2} / (d/2)! & d \text{ 为偶数} \\ 2^d \pi^{(d-1)/2} / ((d-1)!/2)! & d \text{ 为奇数} \end{cases}$$

主轴方向: 由 Σ 的特征向量确定

主轴长度: 由 Σ 的特征值决定



马氏距离：(Mahalanobis Distance)

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}$$

| 欧氏距离 | 马氏距离 |
|---------------------------|-------------------------------|
| 适用于各向同性 (isotropic) 空间 | 不受量纲的影响,排除 变量之间的相关性的干 扰 |
| 不适宜进行线性变换 | 夸大了变化微小的变量的 作用, 需要大量样本 |

正态分布的参数估计

- Gauss分布的参数由均值矢量 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 构成，其参数估计结果为：

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^t$$

5.1 类别可分性判据

3 聚类——怎样才算一个好的划分?



可分性判据——怎样才算一组好的特征?

□思路：类内散布程度低、类间散布程度高

□模型：

- 1) 类内距离准则：用每个样本与其所属类别中心之间的距离平方和来度量

$$J_W(C_1, \dots, C_k) = \frac{1}{n} \sum_{j=1}^k \sum_{\mathbf{x} \in C_j} \|\mathbf{x} - \mathbf{m}_j\|^2 \quad \mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$$

- 2) 类间距离准则：用每个类别的中心到样本整体中心之间的加权距离平方和度量

$$J_B(C_1, \dots, C_k) = \sum_{j=1}^k \frac{n_j}{n} \|\mathbf{m}_j - \mathbf{m}\|^2 \quad \mathbf{m} = \frac{1}{n} \sum_{\mathbf{x} \in D} \mathbf{x}$$

附录D——多元正态分布与散布矩阵

样本 \mathbf{x} 的“散布矩阵”

$$\mathbf{S} = \frac{1}{n} \sum_{\mathbf{x} \in D} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t$$

$$= \frac{1}{n} \sum_{\mathbf{x} \in D} \left[(x_i - m_i)(x_j - m_j) \right]_{d \times d}$$

$$= \left[\frac{1}{n} \sum_{\mathbf{x} \in D} (x_i - m_i)(x_j - m_j) \right]_{d \times d}$$

下标表示向量的第
 i 、 j 维特征

散布矩阵-----协方差矩阵的估计!!!

回顾——3.1.4 聚类问题的描述

类内、类间“散布矩阵”

□第 j 类内散布矩阵:

$$S_W^j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} (\mathbf{x} - \mathbf{m}_j)(\mathbf{x} - \mathbf{m}_j)^t$$

协方差矩阵估计

□总的类内散布矩阵:

$$S_W = \sum_{j=1}^k \frac{n_j}{n} S_W^j$$

□类间散布矩阵:

$$S_B = \sum_{j=1}^k \frac{n_j}{n} (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^t$$

□总体散布矩阵:

$$\mathbf{S}_T = \sum_{\mathbf{x} \in D} \frac{1}{n} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t = \mathbf{S}_W + \mathbf{S}_B$$

例：证明散布准则 $J_W(C_1, \dots, C_k) = \text{tr}(S_W)$

• 基于迹的准则：

$$\text{tr}[S_W] = \sum_{i=1}^c \text{tr}[S_i]$$

$$= \sum_{i=1}^c \sum_{\mathbf{x} \in D_i} \text{tr}[(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t] = \sum_{i=1}^c \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2 = J_W$$

对角线元素和，代表散布半径的平方

等价于误差平方和准则

$$\underline{S_T = S_W + S_B}$$


$$\underline{\text{tr}[S_T] = \text{tr}[S_W] + \text{tr}[S_B]}$$

与具体划分无关

最小化类内准则的同时，也最大化了类间准则

回顾——3.1.4 聚类问题的描述

详见模式分类第10章，
习题25-29



4) 类内、类间距离准则

□基于迹的距离准则: $J_W = \text{tr}[S_w], J_B = \text{tr}[S_B]$

➤ 误差平方和准则，最小化类内准则的同时，最大化类间准则

□基于行列式的散布准则: $J_d = |S_w|$

➤ 行列式（特征值的积）反映散布体积的平方

□基于不变量的散布准则: $J_{wb} = \text{tr}[S_T^{-1}S_W] = \sum_{i=1}^d \frac{1}{1 + \lambda_i}$

➤ $S_W^{-1}S_B$ 的特征值 $\lambda_1, \dots, \lambda_d$ 在非奇异变换下是一个不变量。

$S_T^{-1}S_B$ 的特征值与 $S_W^{-1}S_B$ 的特征值满足关系 $v_i = \frac{1}{1 + \lambda_i}$

衡量类间散布和类内散布在对应特征向量方向上的比值

散布准则

$$J_1(\mathcal{X}) = \text{tr}(\mathbf{S}_w^{-1} \mathbf{S}_b)$$

$$J_2(\mathcal{X}) = \frac{\text{tr}(\mathbf{S}_b)}{\text{tr}(\mathbf{S}_w)}$$

$$J_3(\mathcal{X}) = \frac{|\mathbf{S}_b|}{|\mathbf{S}_w|} = |\mathbf{S}_w^{-1} \mathbf{S}_b|$$

$$J_4(\mathcal{X}) = \frac{|\mathbf{S}_t|}{|\mathbf{S}_w|}$$

例：已知两类样本，计算3维特征中1、2维的类别

可分性判据 $J_1(\mathcal{X}) = \text{tr}(\mathbf{S}_w^{-1} \mathbf{S}_b)$

$$S_W^j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} (\mathbf{x} - \mathbf{m}_j)(\mathbf{x} - \mathbf{m}_j)^t$$

$$\text{其中 } S_W = \sum_{j=1}^k \frac{n_j}{n} S_W^j$$

$$S_B = \sum_{j=1}^k \frac{n_j}{n} (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^t$$

$$\omega_1 : \mathbf{x}_1 = (0, 0, 0)^t, \mathbf{x}_2 = (1, 0, 0)^t, \mathbf{x}_3 = (2, 2, 1)^t, \mathbf{x}_4 = (1, 1, 0)^t$$

$$\omega_2 : \mathbf{x}_5 = (0, 0, 1)^t, \mathbf{x}_6 = (0, 2, 0)^t, \mathbf{x}_7 = (0, 2, 1)^t, \mathbf{x}_8 = (1, 1, 1)^t$$

解：首先计算第 1、2 维特征上每个类别的均值和样本的总体均值：

$$\boldsymbol{\mu}_1 = \frac{1}{4} \sum_{i=1}^4 \mathbf{x}_i = (1.00, 0.75)^t, \boldsymbol{\mu}_2 = \frac{1}{4} \sum_{i=5}^8 \mathbf{x}_i = (0.25, 1.25)^t$$

$$\boldsymbol{\mu} = \frac{1}{8} \sum_{i=1}^8 \mathbf{x}_i = (0.625, 1.000)^t$$

计算类内散布矩阵：

$$\begin{aligned} \mathbf{S}_w &= \frac{1}{2} \left[\frac{1}{4} \sum_{i=1}^4 (\mathbf{x}_i - \boldsymbol{\mu}_1)(\mathbf{x}_i - \boldsymbol{\mu}_1)^t + \frac{1}{4} \sum_{i=5}^8 (\mathbf{x}_i - \boldsymbol{\mu}_2)(\mathbf{x}_i - \boldsymbol{\mu}_2)^t \right] \\ &= \begin{pmatrix} 0.3438 & 0.2188 \\ 0.2188 & 0.6875 \end{pmatrix} \end{aligned}$$

计算类间散布矩阵：

$$\mathbf{S}_b = \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu})(\boldsymbol{\mu}_1 - \boldsymbol{\mu})^t + \frac{1}{2} (\boldsymbol{\mu}_2 - \boldsymbol{\mu})(\boldsymbol{\mu}_2 - \boldsymbol{\mu})^t = \begin{pmatrix} 0.1406 & -0.0938 \\ -0.0938 & 0.0625 \end{pmatrix}$$

学号尾数为单：计算3维特征中1、3维可分性判据
学号尾数为双：计算3维特征中2、3维可分性判据

$$J_1(\mathcal{X}) = \text{tr}(\mathbf{S}_w^{-1} \mathbf{S}_b)$$

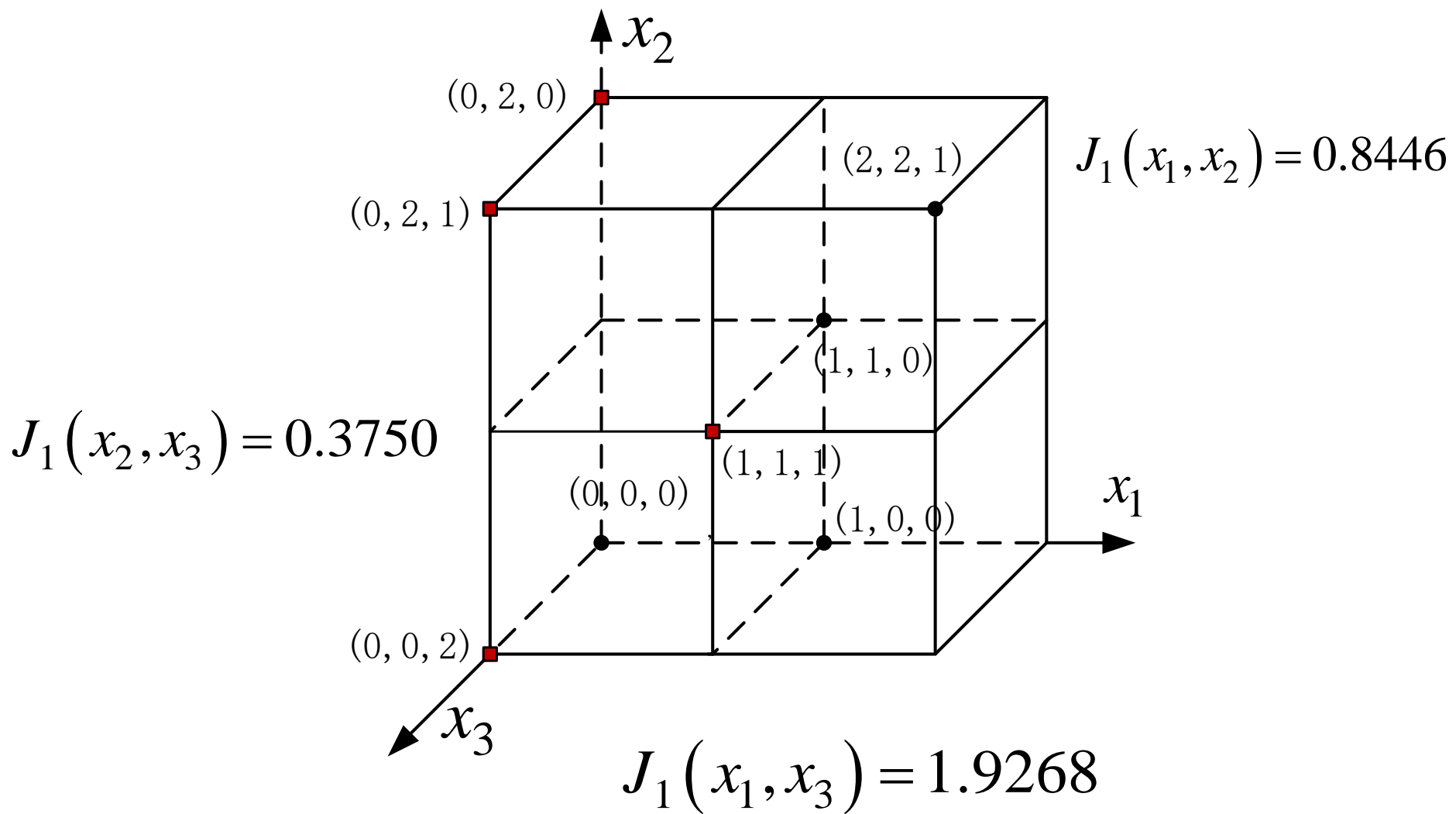
$$S_W^j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} (\mathbf{x} - \mathbf{m}_j)(\mathbf{x} - \mathbf{m}_j)^t$$

$$\text{其中 } S_W = \sum_{j=1}^k \frac{n_j}{n} S_W^j$$

$$S_B = \sum_{j=1}^k \frac{n_j}{n} (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^t$$

$$\omega_1 : \mathbf{x}_1 = (0, 0, 0)^t, \mathbf{x}_2 = (1, 0, 0)^t, \mathbf{x}_3 = (2, 2, 1)^t, \mathbf{x}_4 = (1, 1, 0)^t$$

$$\omega_2 : \mathbf{x}_5 = (0, 0, 1)^t, \mathbf{x}_6 = (0, 2, 0)^t, \mathbf{x}_7 = (0, 2, 1)^t, \mathbf{x}_8 = (1, 1, 1)^t$$



5.2 特征选择

从原始的特征集合 \mathcal{X} 中挑选出一组最有利于分类的特征 \mathcal{X}'

$$\mathcal{X}' = \arg \max_{\tilde{\mathcal{X}} \subset \mathcal{X}} J(\tilde{\mathcal{X}})$$

思路 1: 用可分性判据 J 分别评价每一个特征,
然后根据判据值的大小对特征重新排序, 使得:

$$J(x_1) \geq J(x_2) \geq \cdots \geq J(x_d)$$

特征之间相互独立时才能够保证解的最优性

思路 2: 对所有 $\tilde{\mathcal{X}} \subset \mathcal{X}$ 的特征组合进行穷举,
计算每一种组合的判据值, 选择出最优组合。

从100个特征中选择10个时组合数则变为 $C_{100}^{10} = 17310309456440$

5.2.1 分支定界法

判据单调性：对于两个特征子集 \mathcal{X}_1 和 \mathcal{X}_2 来说：

$$\mathcal{X}_1 \subset \mathcal{X}_2 \Rightarrow J(\mathcal{X}_1) \leq J(\mathcal{X}_2)$$

满足单调性：

$$J_1(\mathcal{X}) = \text{tr}(\mathbf{S}_w^{-1} \mathbf{S}_b)$$

$$J_3(\mathcal{X}) = \frac{|\mathbf{S}_b|}{|\mathbf{S}_w|} = |\mathbf{S}_w^{-1} \mathbf{S}_b|$$

不满足单调性：

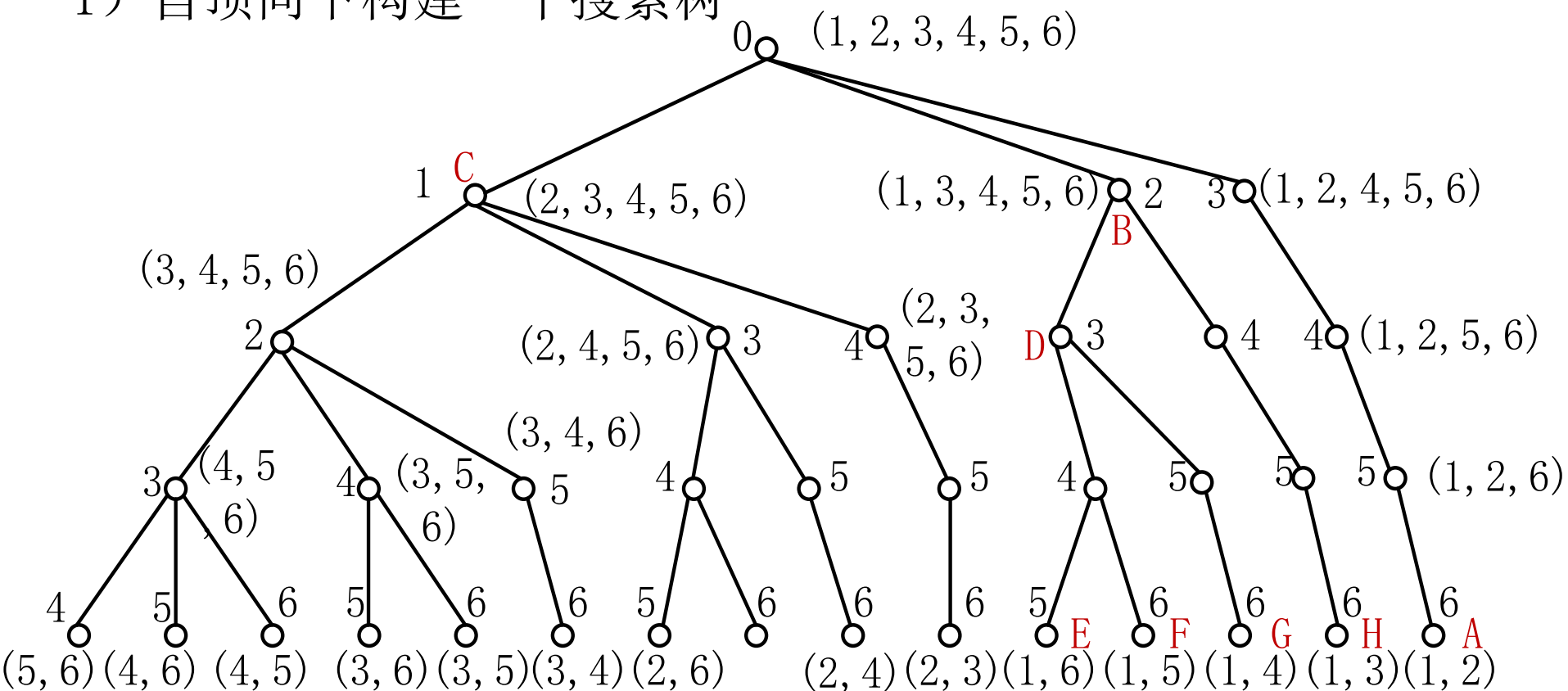
$$J_2(\mathcal{X}) = \frac{\text{tr}(\mathbf{S}_b)}{\text{tr}(\mathbf{S}_w)}$$

$$J_4(\mathcal{X}) = \frac{|\mathbf{S}_t|}{|\mathbf{S}_w|}$$

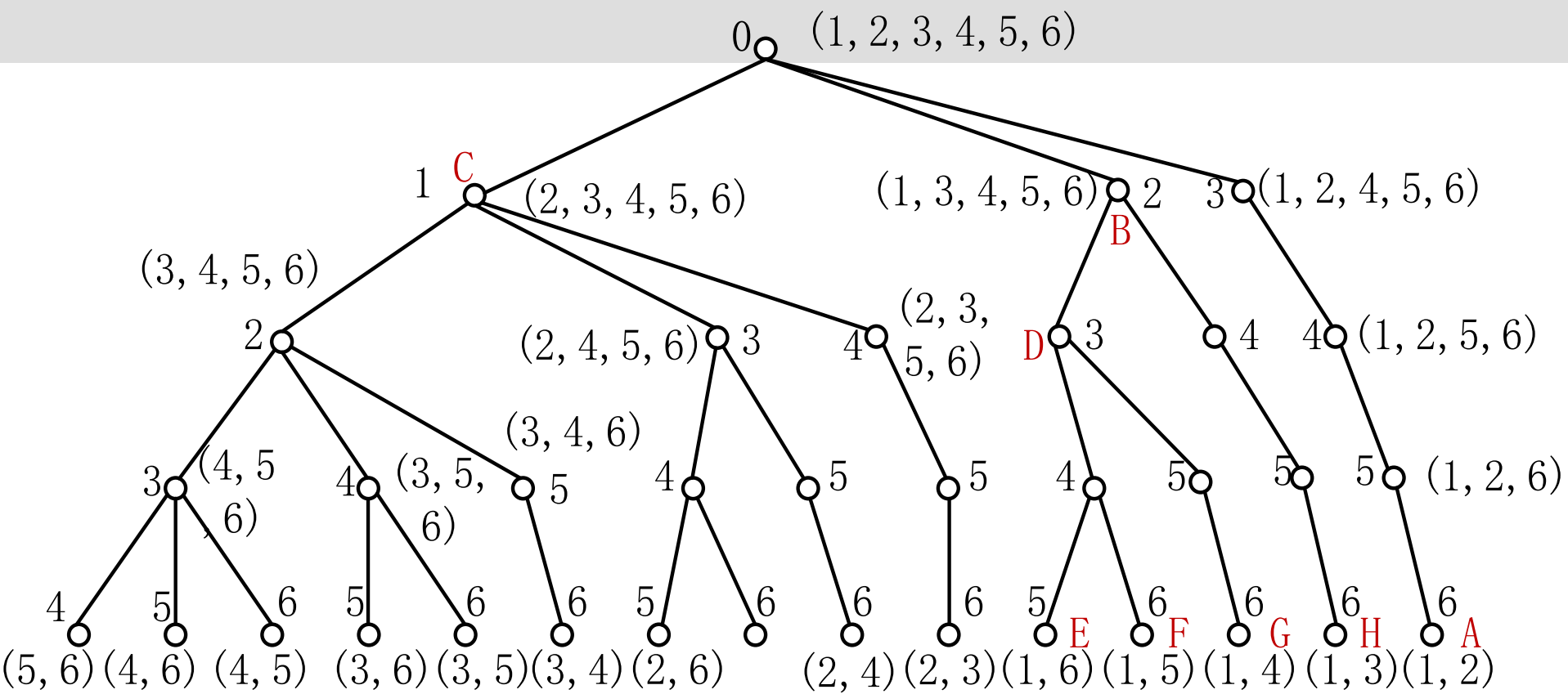
当可分性判据满足单调性时，分支定界法才能够保证搜索到最优特征组合

例：原始6维特征 $\mathcal{X} = \{x_1, x_2, \dots, x_6\}$ 中选出2维特征组合

1) 自顶向下构建一个搜索树



不关心删除的先后次序，只关心删除结果。每一条路径构成一种删除方式



2 由右至左深度优先的方式进行搜索，

首先计算节点 A 对应特征组合的判据值 $J(A)$ 作为当前的最优结果；

回溯搜索至节点 B，计算判据值 $J(B)$ ；如果 $J(B) < J(A)$ ，

由于 E, F, G, H 均为 B 的后继节点，对应的特征组合均为 B 的子集，根据判据 J 的单调性，有： $J(E), J(F), J(G), J(H) \leq J(B) < J(A)$

因此在 B 的后继节点中不可能存在最优节点，无需继续搜索；

如果 $J(B) > J(A)$ ，那么需要继续向下搜索节点 H 和节点 D。

分支定界法存在的问题：

- **可分性判据必须具有单调性。**不具有单调性则不能保证得到最优选择；
- **最优解分支位置决定计算复杂度。**
 - 如果最优解分支在最右端并且根节点的子节点判据值均小于最优解，则搜索效率最高；
 - 如果每个分支的可分性判据都大于其左端分支的可分性判据，实际的计算复杂度会超过穷举法。
- **计算量仍然可观。**当原始的特征维数很大时，搜索到最优解需要的计算量仍然可观

5.2.2 次优搜索算法

顺序前进法 (Sequential Forward Selection, SFS)

从一个空集开始每次向选择的特征集合中加入一个特征直到特征集合中包含 d' 个特征为止,

每次选择加入特征的原则: 加入特征集后能够使得可分性判据最大。

每一轮迭代只需计算将每一个未被选择的特征加入 \mathcal{X}' 之后的判据值, 选择出 d' 个特征需要计算判据值的次数为:

$$\sum_{i=0}^{d'-1} (d-i) = \frac{d'(2d-d'+1)}{2}$$

顺序后退法 (Sequential Backward Selection, SBS)

每一轮从特征集中选择一个最差的特征删除，
选择特征的原则是将其删除之后使得判据值下降最小。

$$\sum_{i=0}^{d-d'-1} (d-i) = \frac{(d-d')(d+d'+1)}{2}$$

顺序后退法 (Sequential Backward Selection, SBS)

每一轮从特征集中选择一个最差的特征删除，
选择特征的原则是将其删除之后使得判据值下降最小。

$$\sum_{i=0}^{d-d'-1} (d-i) = \frac{(d-d')(d+d'+1)}{2}$$

广义顺序前进(后退)法 **Generalized Sequential Forward(Backward) Selection,**

每次增加或删除 r 个特征

$$\sum_{i=0}^{k-1} C_{d-i \times r}^r = \frac{1}{r!} \times \sum_{i=0}^{k-1} \frac{(d-i \times r)!}{(d-i \times r - r)!}$$

增 l —减 r 法（ $l-r$ 法）

先采用顺序前进法向选择特征集合 \mathcal{X}' 加入 l 个特征，
然后采用顺序后退法从 \mathcal{X}' 中删除 r 个特征（ $l > r$ ），
循环这个过程直到 \mathcal{X}' 中包含 d' 个特征为止。

5.3特征提取

- 特征选择：从原始 \mathbf{x} 中选出若干维特征
- 特征提取：对原始特征进行函数变换得到新特征

$$\mathbf{y} = f(\mathbf{x})$$

- 线性特征提取方法：

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

- 主成分分析
- Fisher判别分析

5.3.1 主成分分析 (PCA, Principal Component Analysis)

- ▣ **PCA**是一种最常用的线性成分分析方法;
- ▣ **PCA**的主要思想是寻找到数据的主轴方向, 由主轴构成一个新的坐标系(维数可以比原维数低), 然后数据由原坐标系向新的坐标系投影。
- ▣ **PCA**的其它名称: 离散**K-L**变换, **Hotelling**变换;

问题：有 n 个 d 维样本， $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ ，如何仅用一个样本 \mathbf{x}_0 代表这些样本，使误差准则函数最小？

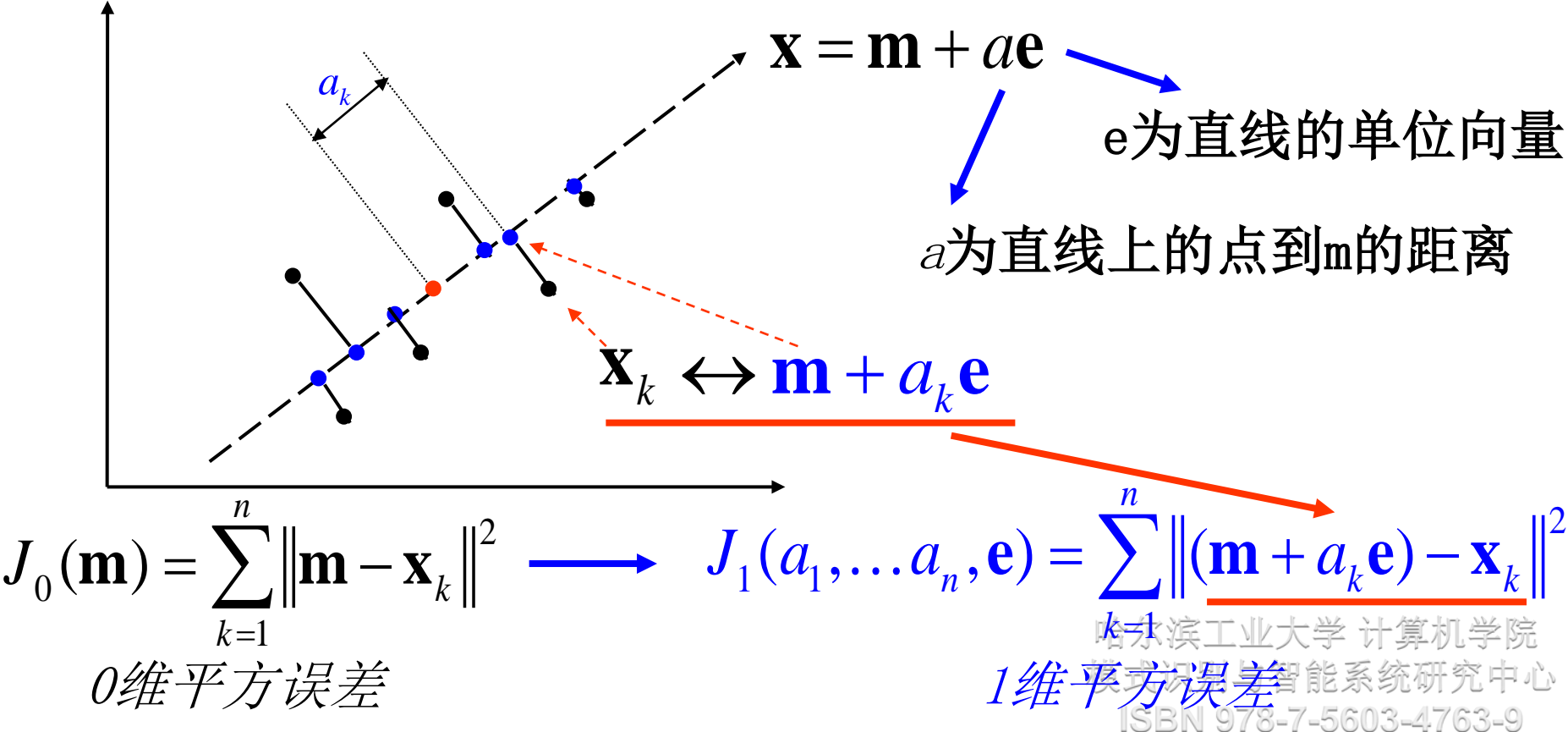
$$\begin{aligned}
 J_0(\mathbf{x}_0) &= \sum_{k=1}^n \|\mathbf{x}_0 - \mathbf{x}_k\|^2 \quad \longrightarrow \quad \mathbf{x}_0 = \mathbf{m} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \\
 &= \sum_{k=1}^n \|(\mathbf{x}_0 - \mathbf{m}) - (\mathbf{x}_k - \mathbf{m})\|^2 \\
 &= \sum_{k=1}^n \|\mathbf{x}_0 - \mathbf{m}\|^2 - 2 \sum_{k=1}^n (\mathbf{x}_0 - \mathbf{m})^t (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^n \|\mathbf{x}_k - \mathbf{m}\|^2 \\
 &= \sum_{k=1}^n \|\mathbf{x}_0 - \mathbf{m}\|^2 - 2(\mathbf{x}_0 - \mathbf{m})^t \underbrace{\sum_{k=1}^n (\mathbf{x}_k - \mathbf{m})}_{=0} + \underbrace{\sum_{k=1}^n \|\mathbf{x}_k - \mathbf{m}\|^2}_{\text{不依赖于 } \mathbf{x}_0}
 \end{aligned}$$

$\mathbf{x}_0 = \mathbf{m}$ 时取得最小值

样本均值是样本数据集的零维表达。
将样本数据集的空间分布，压缩为一个均值点。 → 简单，但不能反映样本间的差异

零维表达改为“一维”表达，将数据集空间，压缩为一条过均值点的线。 ← 样本间的差异

每个样本在直线上存在不同的投影，可以反映样本间的差异

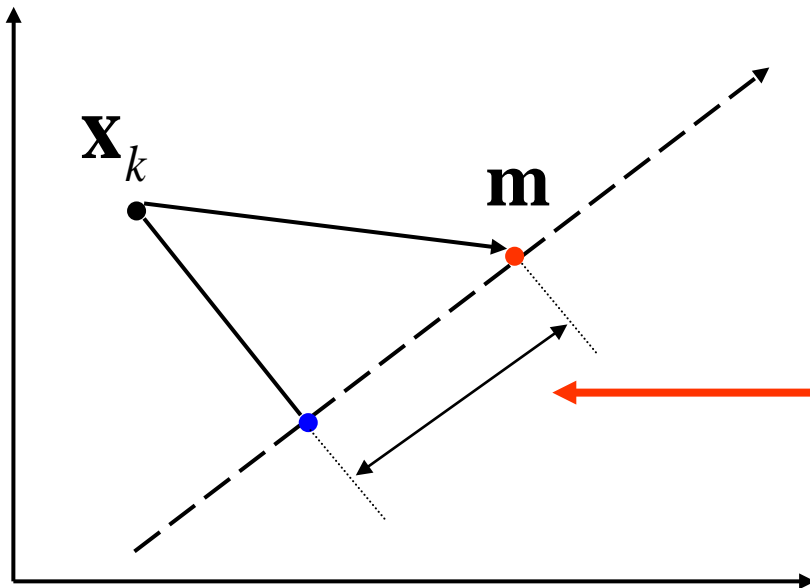


$$\begin{aligned}
 J_1(a_1, \dots, a_n, \mathbf{e}) &= \sum_{k=1}^n \|(\mathbf{m} + a_k \mathbf{e}) - \mathbf{x}_k\|^2 = \sum_{k=1}^n \|a_k \mathbf{e} - (\mathbf{x}_k - \mathbf{m})\|^2 \\
 &= \sum_{k=1}^n a_k^2 \underbrace{\|\mathbf{e}\|^2}_{=1} - 2 \sum_{k=1}^n a_k \mathbf{e}^t (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^n \|\mathbf{x}_k - \mathbf{m}\|^2
 \end{aligned}$$

$$\frac{\partial J_1(a_1, \dots, a_n, \mathbf{e})}{\partial a_k} = 2a_k - 2\mathbf{e}^t (\mathbf{x}_k - \mathbf{m}) = 0$$



$$a_k = \mathbf{e}^t (\mathbf{x}_k - \mathbf{m})$$



只需把向量 \mathbf{x}_k 向过 \mathbf{m} 的直线垂直投影就能得到最小方差



如何找到直线的最优方向？

$$\begin{aligned}
J_1(a_1, \dots, a_n, \mathbf{e}) &= \sum_{k=1}^n \|(\mathbf{m} + a_k \mathbf{e}) - \mathbf{x}_k\|^2 = \sum_{k=1}^n \|a_k \mathbf{e} - (\mathbf{x}_k - \mathbf{m})\|^2 \\
&= \sum_{k=1}^n a_k^2 \underbrace{\|\mathbf{e}\|^2}_{=1} - 2 \sum_{k=1}^n a_k \mathbf{e}^t (\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^n \|\mathbf{x}_k - \mathbf{m}\|^2 \\
J_1(\mathbf{e}) &= \sum_{k=1}^n a_k^2 - 2 \sum_{k=1}^n a_k + \sum_{k=1}^n \|\mathbf{x}_k - \mathbf{m}\|^2 \quad a_k = \mathbf{e}^t (\mathbf{x}_k - \mathbf{m}) \\
&= - \sum_{k=1}^n \left(\mathbf{e}^t (\mathbf{x}_k - \mathbf{m}) \right)^2 + \sum_{k=1}^n \|\mathbf{x}_k - \mathbf{m}\|^2 \\
&= - \sum_{k=1}^n \mathbf{e}^t (\mathbf{x}_k - \mathbf{m}) (\mathbf{x}_k - \mathbf{m})^t \mathbf{e} + \sum_{k=1}^n \|\mathbf{x}_k - \mathbf{m}\|^2 \quad \text{协方差矩阵的 } n-1 \text{ 倍: 散布矩阵} \\
&= -\mathbf{e}^t \mathbf{S} \mathbf{e} + \sum_{k=1}^n \|\mathbf{x}_k - \mathbf{m}\|^2 \quad \mathbf{S} = \sum_{k=1}^n (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t
\end{aligned}$$

最小化 $J_1(\mathbf{e})$ \longrightarrow 最大化 $\mathbf{e}^t \mathbf{S} \mathbf{e}$, 约束条件为: $\|\mathbf{e}\|=1$

最大化 $\mathbf{e}^t \mathbf{S} \mathbf{e}$ ，约束条件为： $\|\mathbf{e}\|=1$ \longrightarrow Lagrange 乘子法

$$u = \mathbf{e}^t \mathbf{S} \mathbf{e} - \lambda \mathbf{e}^t \mathbf{e}$$

$$\frac{\partial u}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda \mathbf{e} = 0 \quad \longrightarrow \quad \underline{\mathbf{S}\mathbf{e}} = \underline{\lambda \mathbf{e}}$$

\nearrow 散布矩阵
 \nearrow 散布矩阵的特征值

$$\mathbf{e}^t \mathbf{S} \mathbf{e} = \mathbf{e}^t \lambda \mathbf{e} = \lambda$$

为了最大化 $\mathbf{e}^t \mathbf{S} \mathbf{e}$

选取散布矩阵最大特征值 λ_{\max}

选取 λ_{\max} 对应的特征向量作为投影直线 \mathbf{e} 的方向

PCA算法——从0维，1维到 d' 维

有 n 个 d 维样本， $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$,

零维表达：仅用一个样本 \mathbf{x}_0 代表这些样本，使误差最小？

$$\mathbf{x}_0 = \mathbf{m} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$$

简单，但不能反映样本间的差异

一维表达：将这些样本，映射到过 \mathbf{m} 的一条直线上使误差最小？

1, 选取散布矩阵 $\mathbf{S} = \sum_{k=1}^n (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t$ 最大特征值 λ_{\max}

2, 选取 λ_{\max} 对应的特征向量作为直线方向 $\mathbf{x} = \mathbf{m} + a\mathbf{e}$

3, 将样本向直线做垂直投影

d' 维表达：将这些样本，映射到以 \mathbf{m} 为原点的 d' 维空间中，使误差准则函数最小？

PCA算法 d' 维表达:

有样本集合 $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ ，其中 $\mathbf{x} = (x_1, \dots, x_d)^t$ ，以样本均值 \mathbf{m} 为坐标原点建立新的坐标系，则有：
$$\mathbf{x} = \mathbf{m} + \sum_{i=1}^d a_i \mathbf{e}_i$$
，其中 $\{\mathbf{e}_i\}$ 为标准正交向量基：因此有：

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

$$a_i = \mathbf{e}_i^t (\mathbf{x} - \mathbf{m})$$

将特征维数降低到 $d' < d$ ，则有对 \mathbf{x} 的近似：
$$\hat{\mathbf{x}} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i$$
误差平方和准则函数：

$$\begin{aligned} J(\mathbf{e}) &= \sum_{k=1}^n \|\mathbf{x}_k - \hat{\mathbf{x}}_k\|^2 = \sum_{k=1}^n \left\| \sum_{i=1}^d a_{ik} \mathbf{e}_i - \sum_{i=1}^{d'} a_{ik} \mathbf{e}_i \right\|^2 = \sum_{k=1}^n \left\| \sum_{i=d'+1}^d a_{ik} \mathbf{e}_i \right\|^2 \\ &= \sum_{k=1}^n \sum_{i=d'+1}^d a_{ik}^2 = \sum_{i=d'+1}^d \sum_{k=1}^n \left[\mathbf{e}_i^t (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m}) \mathbf{e}_i^t \right] \end{aligned}$$

PCA算法 d' 维表达:

$$\begin{aligned}
 J(\mathbf{e}) &= \sum_{k=1}^n \|\mathbf{x}_k - \hat{\mathbf{x}}_k\|^2 = \sum_{k=1}^n \left\| \sum_{i=1}^d a_{ik} \mathbf{e}_i - \sum_{i=1}^{d'} a_{ik} \mathbf{e}_i \right\|^2 = \sum_{k=1}^n \left\| \sum_{i=d'+1}^d a_{ik} \mathbf{e}_i \right\|^2 \\
 &= \sum_{k=1}^n \sum_{i=d'+1}^d a_{ik}^2 = \sum_{i=d'+1}^d \sum_{k=1}^n \left[\mathbf{e}_i^t (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m}) \mathbf{e}_i^t \right] \\
 &= \sum_{i=d'+1}^d \mathbf{e}_i^t \left[\underbrace{\sum_{k=1}^n (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})}_{\text{散布矩阵}} \mathbf{e}_i^t \right] = \sum_{i=d'+1}^d \mathbf{e}_i^t \mathbf{S} \mathbf{e}_i^t
 \end{aligned}$$

最小化 $J(\mathbf{e})$ ，约束条件为： $\|\mathbf{e}\|=1$ 使用拉格朗日乘数法：

$$J'(\mathbf{e}) = \sum_{i=d'+1}^d \left[\mathbf{e}_i^T \mathbf{S} \mathbf{e}_i - \lambda_i (\mathbf{e}_i^T \mathbf{e}_i - 1) \right]$$

$$J'(\mathbf{e}) = \sum_{i=d'+1}^d \left[\mathbf{e}_i^T \mathbf{S} \mathbf{e}_i - \lambda_i (\mathbf{e}_i^T \mathbf{e}_i - 1) \right]$$

$$\frac{\partial J'(\mathbf{e})}{\partial \mathbf{e}_i} = 2\mathbf{S} \mathbf{e}_i - 2\lambda_i \mathbf{e}_i = 0 \quad \longrightarrow \quad \mathbf{S} \mathbf{e}_i = \lambda_i \mathbf{e}_i$$

λ_i 为 \mathbf{S} 的特征值， \mathbf{e}_i 为 \mathbf{S} 的特征矢量。

$$J(\mathbf{e}) = \sum_{i=d'+1}^d \mathbf{e}_i^T \mathbf{S} \mathbf{e}_i = \sum_{i=d'+1}^d \lambda_i \mathbf{e}_i^T \mathbf{e}_i = \sum_{i=d'+1}^d \lambda_i$$

要使 $J(\mathbf{e})$ 最小，只需将 \mathbf{S} 的特征值由大到小排序，选择最大的前 d' 个特征值对应的特征向量构成一个新的 d' 维坐标系，将样本向新的坐标系的各个轴上投影，计算出新的特征矢量

$$(x_1, \dots, x_d)^T \rightarrow (a_1, \dots, a_{d'})^T \quad \text{其中} \quad a_i = \mathbf{e}_i^T (\mathbf{x} - \mathbf{m})$$

PCA算法

1. 利用训练样本集合计算样本的均值 μ 和散布矩阵 \mathbf{S} ;
2. 计算 \mathbf{S} 的特征值, 并由大到小排序;
3. 选择前 d' 个特征值对应的特征矢量作成变换矩阵 $\mathbf{E}=[\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{d'}]$;
4. 训练和识别时, 每一个输入的 d 维特征矢量 \mathbf{x} 可以转换为 d' 维的新特征矢量 \mathbf{y} :

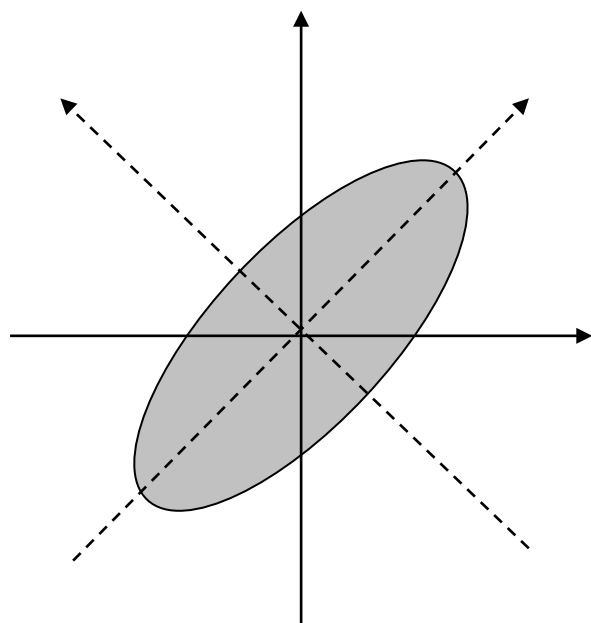
$$\mathbf{y} = \mathbf{E}^t (\mathbf{x} - \mu)$$

5. 通过 \mathbf{y} 近似重构 \mathbf{x} 为:

$$\hat{\mathbf{x}} = \mathbf{E}\mathbf{y} + \mu$$

如果 $d'=d$ 、 \mathbf{S} 满秩, \mathbf{E} 是一个正交阵, 有 $\mathbf{E}^{-1}=\mathbf{E}^T$

PCA的讨论



特征矢量正交：因为S是实对称阵，所以其特征值为实数、特征矢量正交；

变换后特征不相关：将数据向新的坐标轴投影之后，特征之间是不相关的；

冗余特征：特征值描述了变换后各维特征的重要性，特征值为0的各维特征为冗余特征。

降维误差评估：累加特征值，计算比例

$$d' = \arg \min_{1 \leq k \leq d} \left[\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^d \lambda_i} \geq \theta \right]$$

```
function [E, mu] = PCA( X, ratio )
```

```
d = size(X,2);
```

```
%计算均值和协方差矩阵
```

```
mu = mean(X);
```

```
Sigma = cov(X);
```

```
%求特征值和特征矢量
```

```
[V,L] = eigs( Sigma, d ); Lamda = diag(L);
```

```
%计算累加特征值
```

```
AccLamda = cumsum(Lamda); t = AccLamda(d) * ratio;
```

```
%计算保留特征数
```

```
dd = find( (AccLamda(2:d)>=t) & (AccLamda(1:d-1)<t) ) + 1;
```

```
E = V(:,1:dd);
```

X: 样本矩阵 ($n \times d$ 矩阵),
ratio: 特征值累加和占总和的比例
E: 基矢量矩阵 ($d \times d'$ 矩阵, 每列一个矢量)
mu: 均值矢量 (行矢量)

Y--降维之后的样本矩阵 ($n \times d'$ 矩阵)

```
function Y = PCADR(X, E, mu )
```

```
n = size(X,1);
```

```
Y = (X-repmat(mu,n,1))*E; %repmat: 平铺矩阵, 将mu平铺成一个矩阵
```

例：现有下列训练样本，请用PCA算法将2维特征降为1维，并画出训练样本和投影主轴以及投影后的样本点。

$$(10,1)^t, (9,0)^t, (10,-1)^t, (11,0)^t, (0,9)^t, (1,10)^t, (0,11)^t, (-1,10)^t$$

均值： $\mathbf{m} = (5,5)^t$ 协方差矩阵： $S = \begin{bmatrix} 25.5 & -25 \\ -25 & 25.5 \end{bmatrix}$

解： 1. 计算协方差矩阵的特征值和特征向量：

$$S\mathbf{e} = \lambda\mathbf{e} \quad |S - \lambda E| = 0$$

2. 将特征值，并由大到小排序；

3. 选择前 d' 个特征值对应的特征矢量作成变换矩阵

$$\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{d'}];$$

4. \mathbf{x} 转换为 d' 维的新特征矢量 $\mathbf{y} = \mathbf{E}^T(\mathbf{x} - \mathbf{m})$

例：现有下列训练样本，请用PCA算法将2维特征降为1维，并画出训练样本和投影主轴以及投影后的样本点。

$$(10,1)^t, (9,0)^t, (10,-1)^t, (11,0)^t, (0,9)^t, (1,10)^t, (0,11)^t, (-1,10)^t$$

$$\text{均值: } \mathbf{m} = (5,5)^t \quad \text{协方差矩阵: } S = \begin{bmatrix} 25.5 & -25 \\ -25 & 25.5 \end{bmatrix}$$

解：1 计算协方差矩阵的特征值和特征向量：

$$S\mathbf{e} = \lambda\mathbf{e}$$

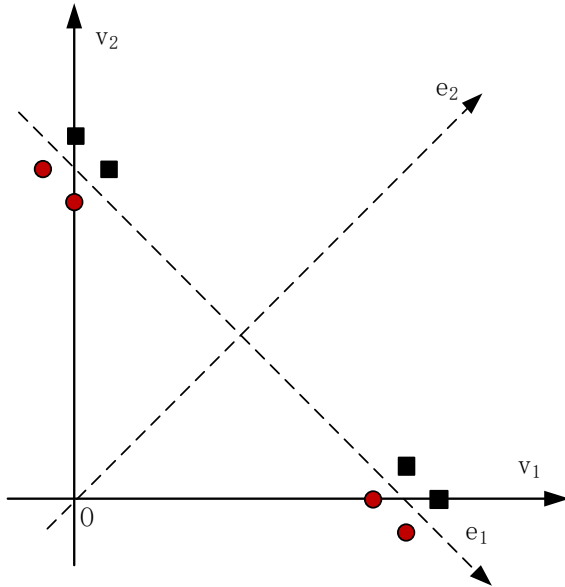
$$|S - \lambda E| = 0$$

$$\lambda_1 = 0.5 \quad \mathbf{e}_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

$$\begin{bmatrix} 25.5 & -25 \\ -25 & 25.5 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \lambda \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

$$\begin{vmatrix} 25.5 - \lambda & -25 \\ -25 & 25.5 - \lambda \end{vmatrix} = 0$$

$$\lambda_2 = 50.5 \quad \mathbf{e}_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$



2) 将特征值，并由大到小排序；

$$\lambda_1 = 50.5$$

$$\lambda_2 = 0.5$$

$$\mathbf{E} = [\mathbf{e}_1] = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \quad \mathbf{e}_2 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

3) 选择前 d' 个特征矢量构成变换矩阵 $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{d'}]$;

$$\mathbf{E} = [\mathbf{e}_1] = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

4) \mathbf{x} 转换为 d' 维的新特征矢量

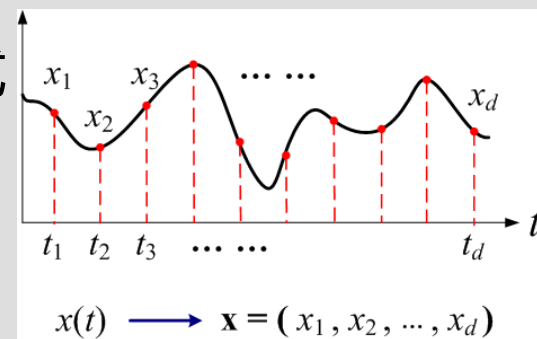
$$\mathbf{y} = \mathbf{E}^t(\mathbf{x} - \mathbf{m}) : \mathbf{m} = (5, 5)^t$$

$$(10, 1)^t, (9, 0)^t, (10, -1)^t, (11, 0)^t, (0, 9)^t, (1, 10)^t, (0, 11)^t, (-1, 10)^t$$

从信号的角度理解PCA——离散K_L变换

x的表示: 对平稳随机过程 $x(t)$ 进行采样, 用向量形式表示为 $\mathbf{x} = (x_1, x_2, \dots, x_d, \dots)$, 用完备正交系 $\mathbf{e}_j, j=1,2,\dots$ 展开为

$$\mathbf{x} = \sum_{j=1}^{\infty} c_j \mathbf{e}_j$$



x的重构误差: 用 \mathbf{x} 的前 d 项重构 $\hat{\mathbf{x}} = \sum_{j=1}^d c_j \mathbf{e}_j$, 其中 $c_j = \mathbf{e}_j^T \mathbf{x}$, 误差的期望为

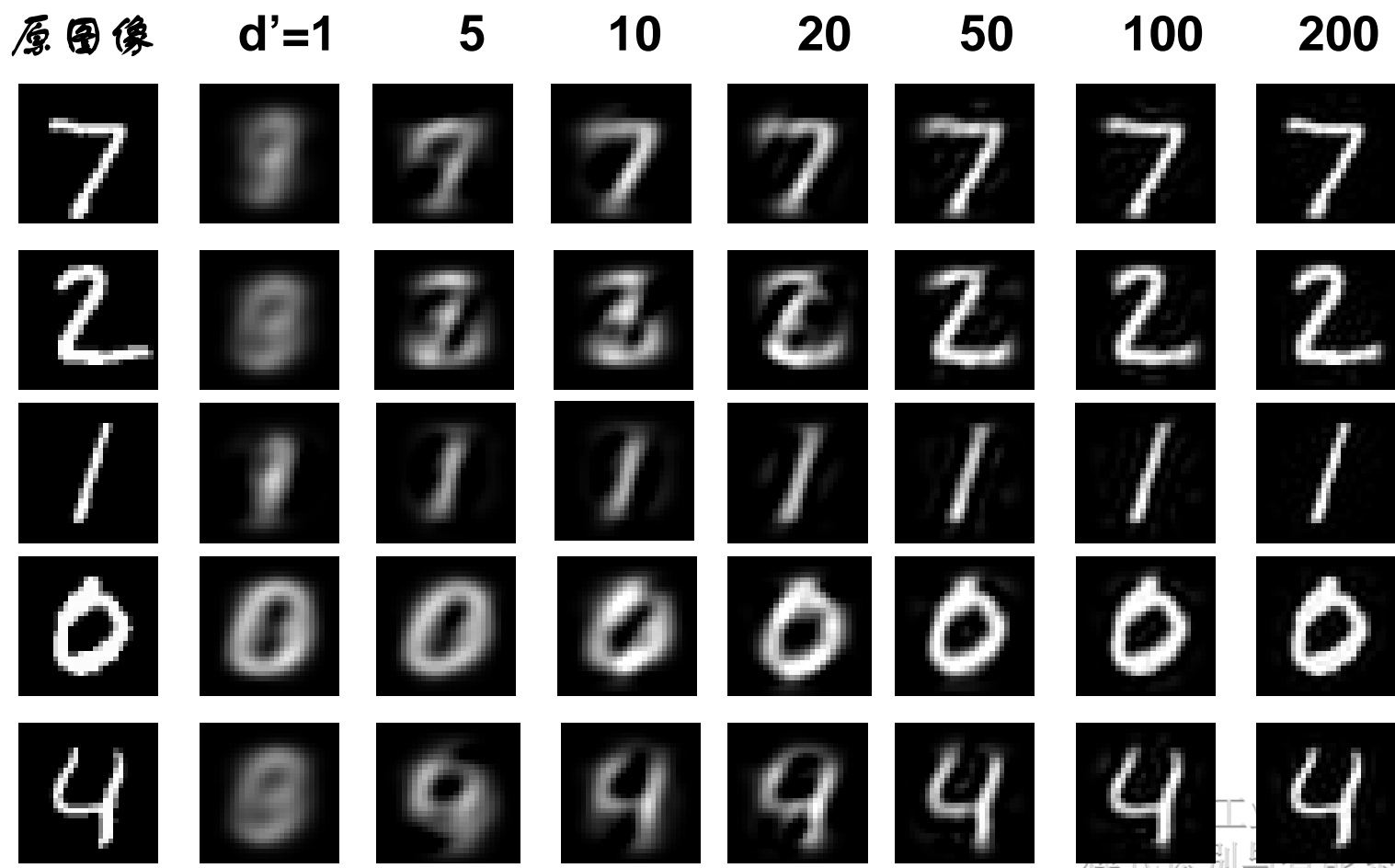
$$\xi = E \left[\sum_{j=d+1}^{\infty} c_j^2 \right] = E \left[\sum_{j=d+1}^{\infty} \mathbf{e}_j^T \mathbf{x} \mathbf{x}^T \mathbf{e}_j \right] = \sum_{j=d+1}^{\infty} \mathbf{e}_j^T E[\mathbf{x} \mathbf{x}^T] \mathbf{e}_j \equiv \sum_{j=d+1}^{\infty} \mathbf{e}_j^T \mathbf{S} \mathbf{e}_j$$

误差最小化: 在 $\|\mathbf{e}_j\|=1$ 约束下的均方误差最小化, 应用Lagrange乘子法

$$g(\mathbf{e}) = \mathbf{e}_j^T \mathbf{S} \mathbf{e}_j - \sum_{j=d+1}^{\infty} \lambda_j (\mathbf{e}_j^T \mathbf{e}_j - 1) \quad \frac{\partial g(\mathbf{e})}{\partial \mathbf{e}_j} = 0 \Rightarrow \begin{cases} \mathbf{S} \mathbf{e}_j - \lambda_j \mathbf{e}_j = 0 \\ \xi = \sum_{j=d+1}^{\infty} \lambda_j \end{cases}$$

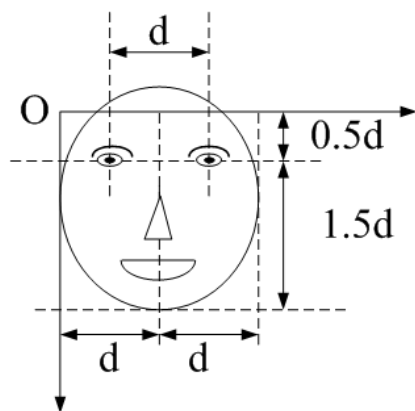
PCA重构

$$\hat{\mathbf{x}} = \mathbf{E}\mathbf{y} + \boldsymbol{\mu}$$

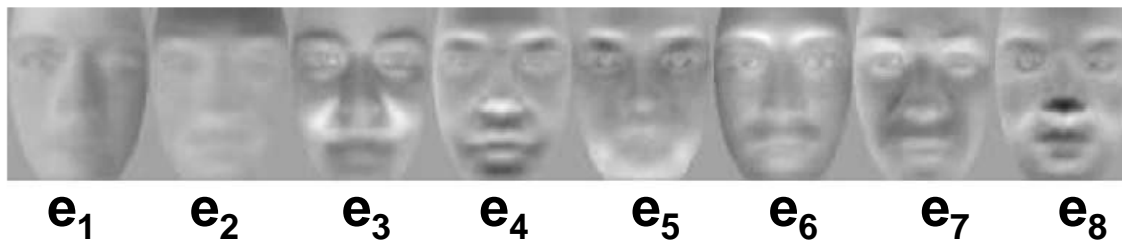


应用举例：基于PCA的人脸识别

归一化：以两眼距离为基准，进行图像校正，归一化为 $N \times N$ 维图像 \mathbf{x}



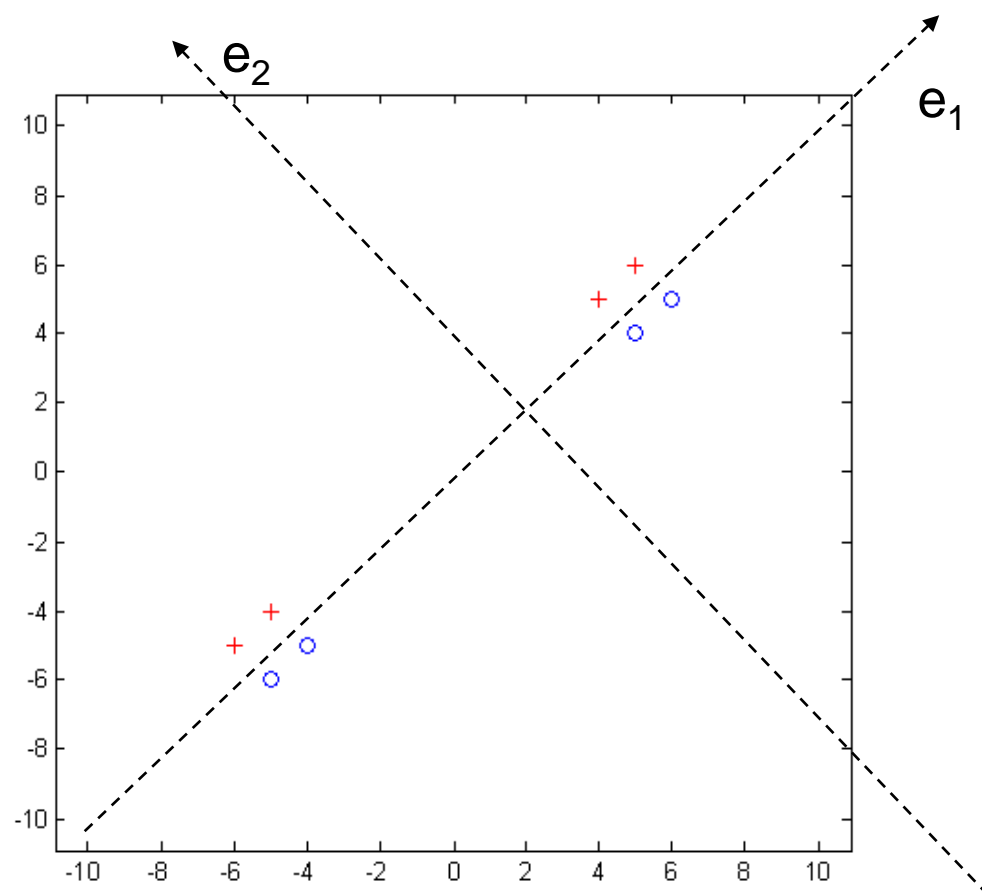
PCA：计算样本 \mathbf{x} 总体散布矩阵，进行PCA，压缩为 d 维特征 \mathbf{y} ，重构图像为 $\hat{\mathbf{x}}$



分类：可采用多种方法进行，如计算信噪比阈值、SVM等

$$R_{SN} = 10 \times \lg \left(\frac{\|\mathbf{x}\|^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|^2} \right)$$

5.3.2 基于Fisher准则的可分性分析 (FDA, Fisher Discriminant Analysis)



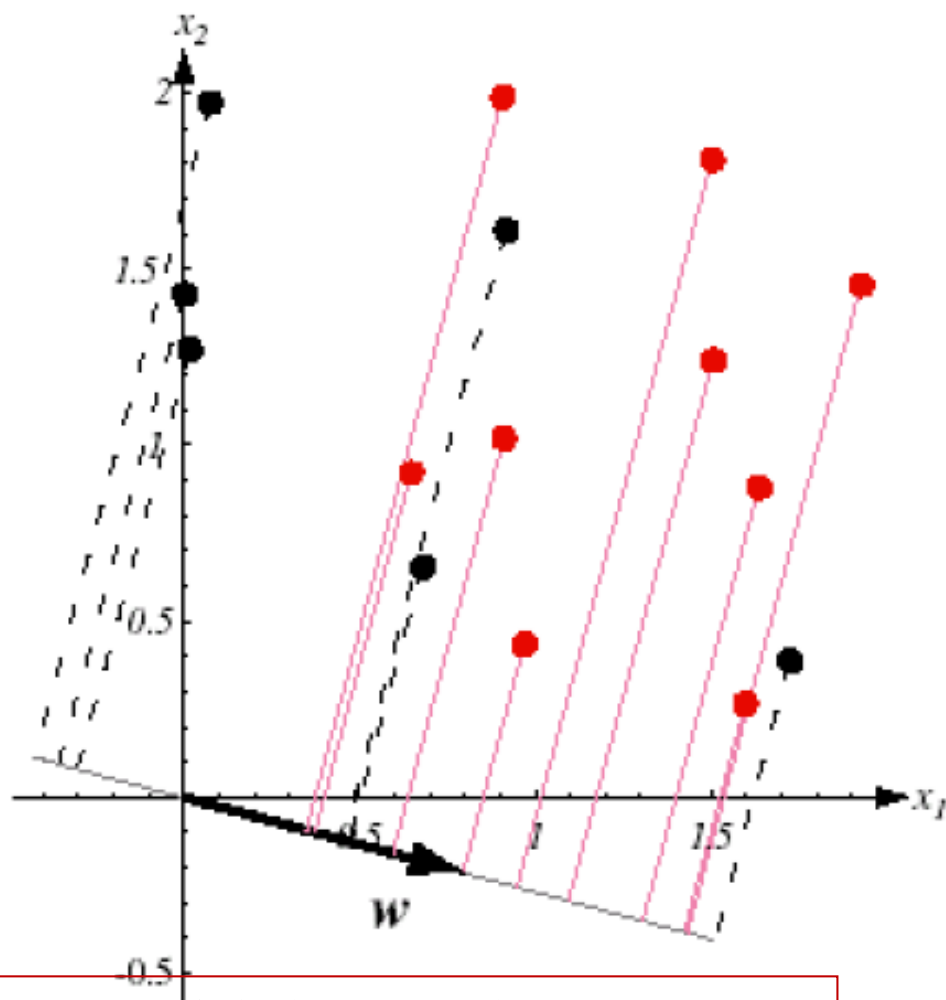
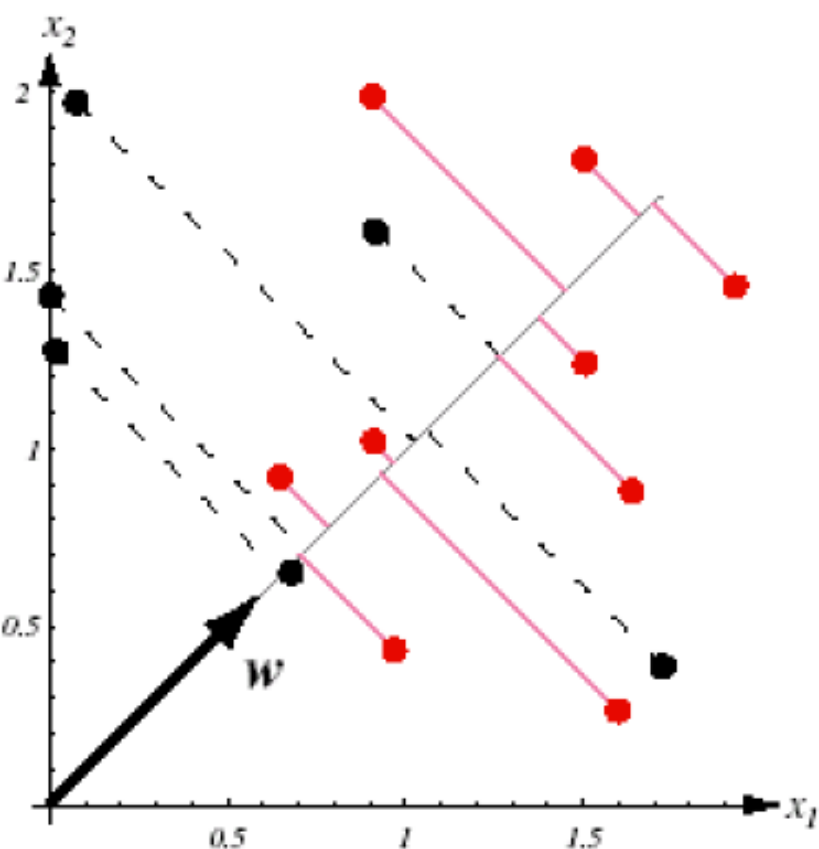
FDA与PCA

PCA的可分性：PCA整体对待所有的样本，寻找“均方误差最小”下的最优线性映射，不考虑样本的类别属性，它所忽略的投影方向有可能恰恰包含了重要的可分性信息；

FDA的可分性：FDA是在可分性最大意义下的最优线性映射，充分保留了样本的类别可分性信息；

FDA别名：FDA还被称为多重判别分析：MDA, Multiple Discriminant Analysis 或LDA(Linear Discriminant Analysis)。

Fisher 线性判别准则



如何选择直线方向 W ，使样本可分性最好？

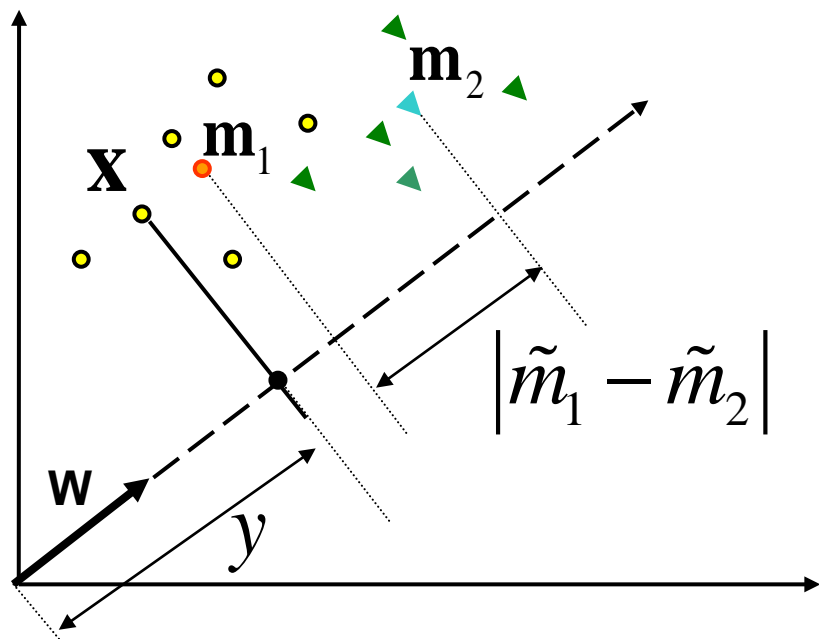
二分类问题 可分性准则函数的构造

样本点 \mathbf{x} 在 \mathbf{w} 方向上的投影 $y = \mathbf{w}^t \mathbf{x}$

两类样本均值 $\mathbf{m}_1, \mathbf{m}_2$ 投影之差: $|\tilde{m}_1 - \tilde{m}_2| = |\mathbf{w}^t (\mathbf{m}_1 - \mathbf{m}_2)|$

第 i 类投影的内类散布: $\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{m}_i)^2$

总体内类散布: $\tilde{s}_1^2 + \tilde{s}_2^2$



可分性准则函数:
(Fisher线性判别准则)

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

准则函数的广义瑞利商形式

总类内散布矩阵: $\mathbf{S}_w = \sum_{i=1}^2 \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$

类间散布矩阵: $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$

$$\tilde{s}_i^2 = \sum_{\mathbf{x} \in D_i} (\mathbf{w}^t \mathbf{x} - \mathbf{w}^t \mathbf{m}_i)^2 = \sum_{\mathbf{x} \in D_i} \mathbf{w}^t (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \mathbf{w} = \mathbf{w}^t \mathbf{S}_i \mathbf{w}$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t \mathbf{S}_1 \mathbf{w} + \mathbf{w}^t \mathbf{S}_2 \mathbf{w} = \mathbf{w}^t \mathbf{S}_w \mathbf{w}$$

$$|\tilde{m}_1 - \tilde{m}_2| = (\mathbf{w}^t \mathbf{m}_1 - \mathbf{w}^t \mathbf{m}_2)^2 = \mathbf{w}^t (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t \mathbf{w} = \mathbf{w}^t \mathbf{S}_B \mathbf{w}$$

Fisher线性判别准则: $J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_w \mathbf{w}}$ 广义瑞利商

广义特征向量

\mathbf{A} 为 n 阶实对称矩阵, \mathbf{B} 为 n 阶实对称正定矩阵
满足 $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ 的数 λ 为 \mathbf{A} 相对于 \mathbf{B} 的特征值
与 λ 相对应的非零解 \mathbf{x} 称为属于 λ 的特征向量

广义瑞利商特性

非零向量 \mathbf{x}_0 是 $R(\mathbf{x})$ 的驻点的充要条件是

\mathbf{x}_0 为 $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ 的属于特征值 λ 的特征向量

$$R(\mathbf{x}) = \frac{\mathbf{x}^t \mathbf{A} \mathbf{x}}{\mathbf{x}^t \mathbf{B} \mathbf{x}}$$

$$\mathbf{x}^t \mathbf{B} \mathbf{x} R(\mathbf{x}) = \mathbf{x}^t \mathbf{A} \mathbf{x}$$

$$2\mathbf{B} \mathbf{x} R(\mathbf{x}) + \mathbf{x}^t \mathbf{B} \mathbf{x} \frac{dR}{d\mathbf{x}} = 2\mathbf{A} \mathbf{x}$$

$$\frac{dR}{d\mathbf{x}} = \frac{2(\mathbf{A} \mathbf{x} - R(\mathbf{x}) \mathbf{B} \mathbf{x})}{\mathbf{x}^t \mathbf{B} \mathbf{x}} = 0$$

$$\mathbf{A} \mathbf{x}_0 = R(\mathbf{x}_0) \mathbf{B} \mathbf{x}_0$$

\mathbf{x}_0 为 $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ 的属于特征值 $\lambda = R(\mathbf{x}_0)$ 的特征向量

非零向量 \mathbf{x}_0 是 $R(\mathbf{x})$ 的驻点的充要条件是
 \mathbf{x}_0 为 $\mathbf{Ax} = \lambda\mathbf{Bx}$ 的属于特征值 λ 的特征向量

Fisher线性判别准则:
$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_w \mathbf{w}}$$

类间散布矩阵:
$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$$

总类内散布矩阵:
$$\mathbf{S}_w = \sum_{i=1}^2 \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

最大化 $J(\mathbf{w})$ 必须满足:
$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}$$

λ 为: \mathbf{S}_B 相对于 \mathbf{S}_w 的特征值, \mathbf{w} 为对应的特征向量

当 \mathbf{S}_w 非奇异时
$$\mathbf{S}_w^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}$$

c 个类别向 d' 个方向投影——FDA算法

□将 d 维特征降为 d' 维特征，使降维后的样本具有最大可分性

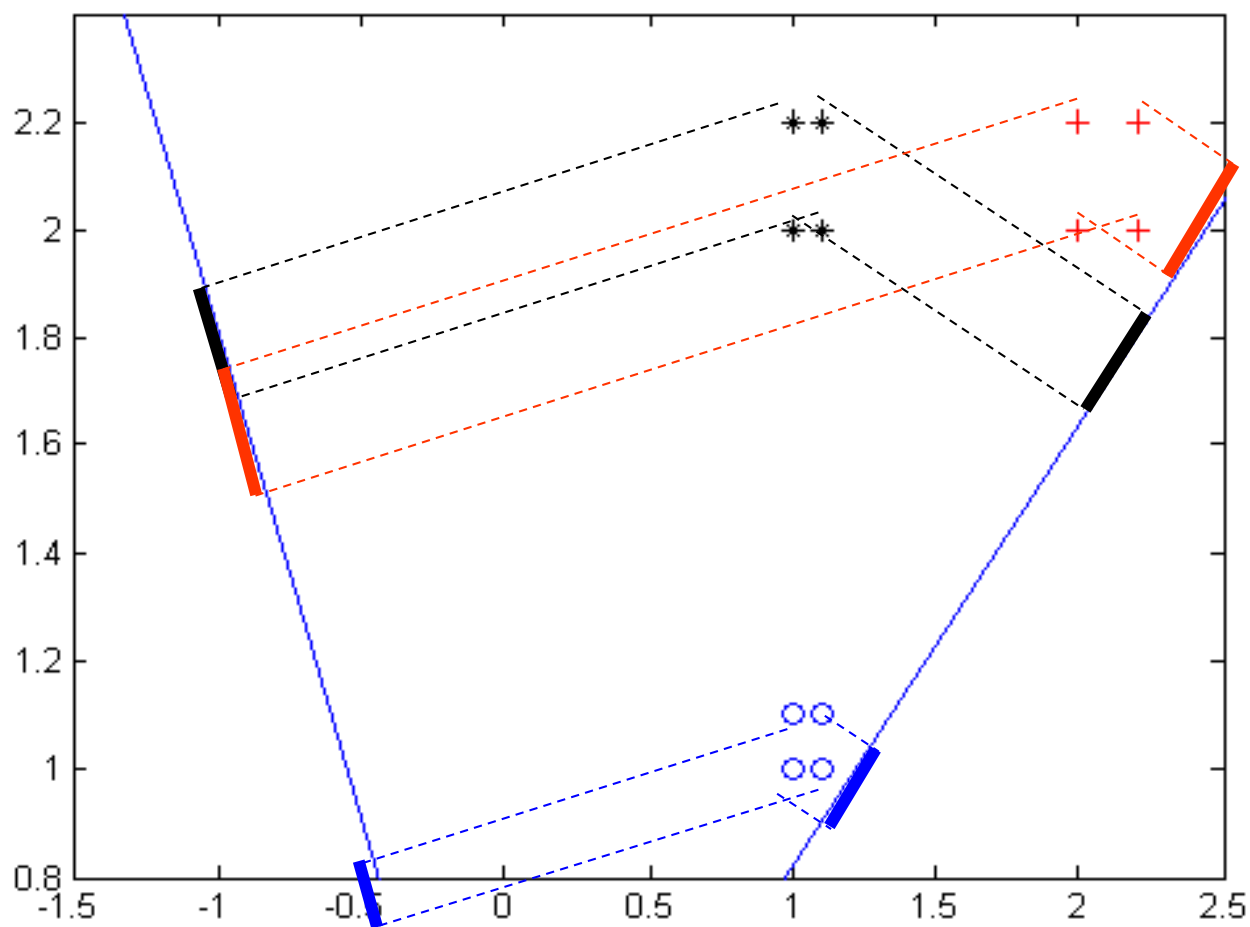
$$\mathbf{S}_w = \sum_{i=1}^c \mathbf{S}_i \quad \mathbf{S}_b = \sum_{i=1}^c n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^t$$

1. 计算类内散布矩阵 \mathbf{S}_w 和类间散度矩阵 \mathbf{S}_b ;
2. 计算 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的特征值和特征矢量;
3. 选择前 d' 个特征矢量作为列矢量成变换矩阵

$$\mathbf{E}=[\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{c-1}];$$

4. 每一个输入的 d 维特征矢量 \mathbf{x} 可以转换为 d' 维的新特征矢量 $\mathbf{x}' = \mathbf{E}^t \mathbf{x}$ 。

3类问题FDA



FDA的讨论

□经FDA变换后，新的坐标系不正交；

➤ $S_w^{-1}S_b$ 不是对称阵，特征矢量不具正交性，变换后，特征之间仍具有一定相关性

□只有当样本数足够多时，才能够保证类内散度矩阵 S_w 为非奇异矩阵。

□新坐标维数最多为 $c-1$ ， c 为类别数；

➤ $S_w^{-1}S_b$ 至多存在 $c-1$ 个大于0的特征值

FDA——作为线性分类器学习方法

- 找到使得两个类别样本可分性最强的投影方向 \mathbf{w} ;
- 将所有的训练样本和待识别样本变换为1维特征 $\mathbf{w}^t \mathbf{x}$
- 设定一个合适的阈值 $-w_0$ 进行分类

$$\mathbf{w}^t \mathbf{x} \begin{cases} \geq -w_0, & \mathbf{x} \in \omega_1 \\ < -w_0, & \mathbf{x} \in \omega_2 \end{cases}$$

$$g(\mathbf{x}) = \mathbf{w}^t \mathbf{x} + \underline{w_0} \begin{cases} \geq 0, & \mathbf{x} \in \omega_1 \\ < 0, & \mathbf{x} \in \omega_2 \end{cases}$$

阈值如何确定?——贝叶斯决策

成分分析的其它方法

统计学方法：

独立成分分析(ICA, Independent Component Analysis)

多维尺度变换(MDS, Multidimensional Scaling)

典型相关分析(CCA, Canonical Correlation Analysis)

偏最小二乘(PLS, Partial Least Square)

核方法：

引入核函数，将线性方法推广为非线性方法；

流形学习 (Manifold Learning)：

“非线性流形”在局部用“线性流形”近似，如Isomap
和Locally Linear Embedding (LLE)

多尺度变换 (MDS: Multidimensional Scaling)

MDS的目标: 给定样本之间距离（相似度）的条件下，在低维空间对样本进行表示，力图保证样本之间的距离不变。

MDS的优化: 给定距离矩阵 $D = (d_{ij})_{n \times n}$

$$\min_{\mathbf{x}_1, \dots, \mathbf{x}_n} \sum_{i < j} \left(\|\mathbf{x}_i - \mathbf{x}_j\| - d_{ij} \right)^2$$

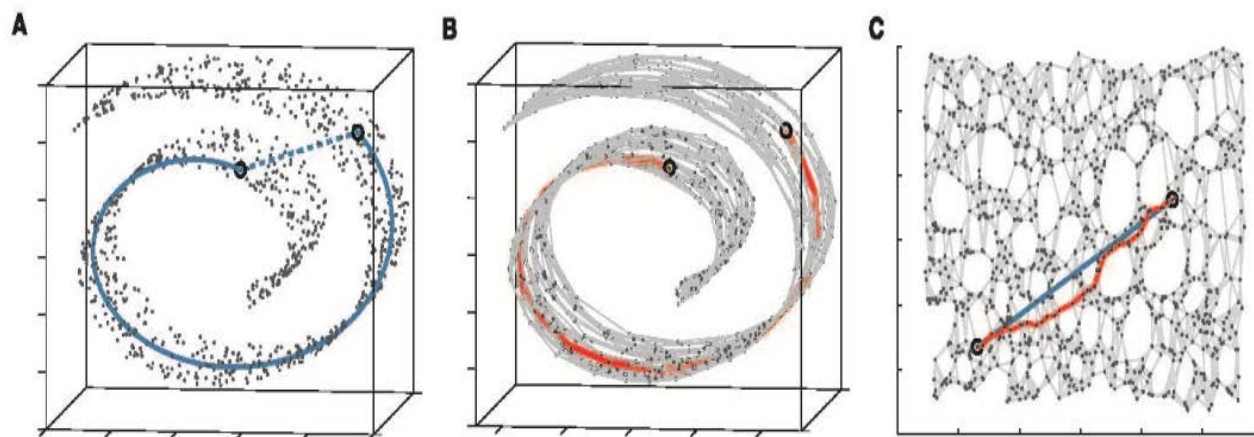
MDS求解: 方法很多，一种是直接采用梯度法优化 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 。

ISOMAP

Science 290 : 2319-2323, 22 December 2000

<http://waldron.stanford.edu/~isomap/>

测地距离：



ISOMAP的思想：根据两个样本的测地距离构造相似图，采用MDS完成向低维空间的映射。

ISOMAP算法

□ ISOMAP算法:

1. 构造样本的 ε 近邻或 k 近邻图，连接近邻样本节点，计算相连节点之间的（欧氏）距离 $d_G(i, j)$ ，不相连节点之间 $d_G(i, j) = \infty$;

2. 计算测地距离：（Dijkstra算法）

$$d_G(i, j) = \min_{k=1, \dots, n} (d_G(i, j), d_G(i, k) + d_G(k, j))$$

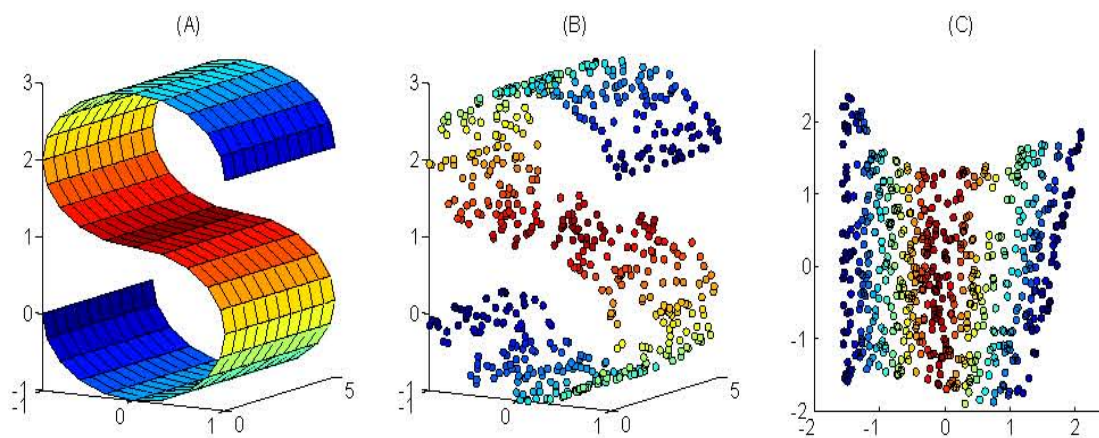
3. $D = (d_G(i, j))_{n \times n}$ ，用NMDS计算低维空间映射。

LLE: Locally Linear Embedding

Science, v. 290, Dec. 22, 2000. pp. 2323--2326.

<http://www.cs.nyu.edu/~roweis/lle/>

LLE的思想： 样本分布在一个嵌入于高维空间中的非线性流形，连续非线性流形的局部可以用一个线性流形逼近。



在局部线性流形上计算每个样本的线性表示；

保持局部线性流形的拓扑不变性，向低维空间映射样本点。