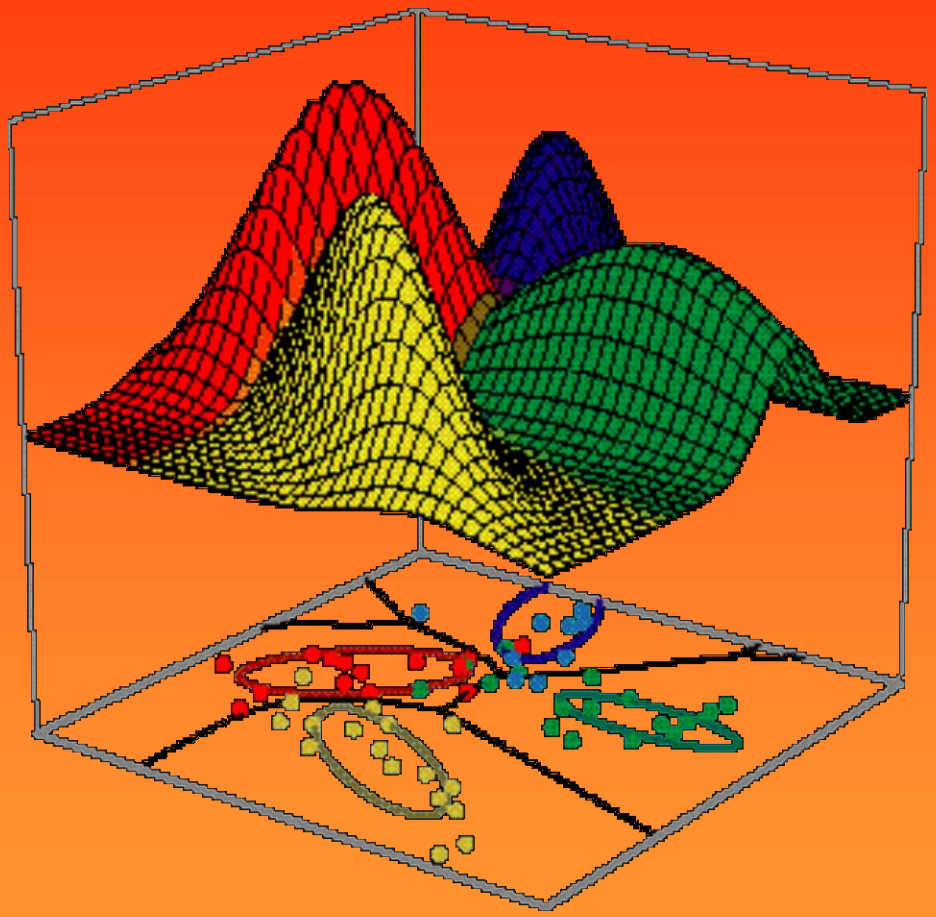


Pattern Classification

All materials in these slides were taken from *Pattern Classification (2nd ed) by* R. O. *Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000* with the permission of the authors and the publisher

9　线性可分问题及求解

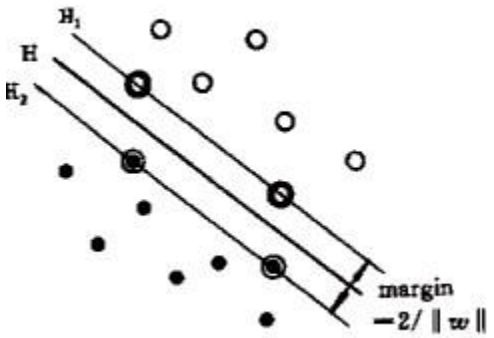
- 9.1　两类线性可分情况
- 9.2　最小化均方误差（ Minimum Squared Error Procedures . MSE）
- 9.3　多类别推广

9.1 两类线性可分的情况

- **线性可分：**

有n个样本 $x_1, x_2, \ldots x_n$ ， 分别属于两个类别 ω_1, ω_2 ，如果存在一个线性判别函数 $g(x) = a^t x$ 够完全正确的对他们分类，这些样本就是线性可分的。

其中权重向量 a 被称为分离向量或者解向量。



9.2 最小化均方误差

- 假设 $a^t x_i = b_i$

b_i为指定的类别标签（每一类别具有的特定标示符）

最小化均方误差过程

- 从而将求解问题替换为求解一组线性方程组的问题，更严格但更容易理解。

It's time for you to go on a diet.

最小化均方误差和伪逆（Pseudoinverse）

对于所有样本 $x_1, x_2, \ldots, x_n$，我们想找到一个权重向量a，使得 $a^t x_i = b_i$.

矩阵表示法为:

$$\begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1d} \\ x_{20} & x_{21} & \cdots & x_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n0} & x_{n1} & \cdots & x_{nd} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \cdots \\ a_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix} \Leftrightarrow Xa = b$$

误差向量: $e = Xa - b$

- Sum-of-squared-error 判别函数:

$$J_s(a) = \|Xa - b\|^2 = \sum_{i=1}^{n} (a^t x_i - b_i)^2$$

- 梯度：

$$\nabla J_s = \sum_{i=1}^{n} 2(a^t x_i - b_i) y_i = 2X^t(Xa - b)$$

- 令梯度等于零，可以得到 $X^t X a = X^t b$

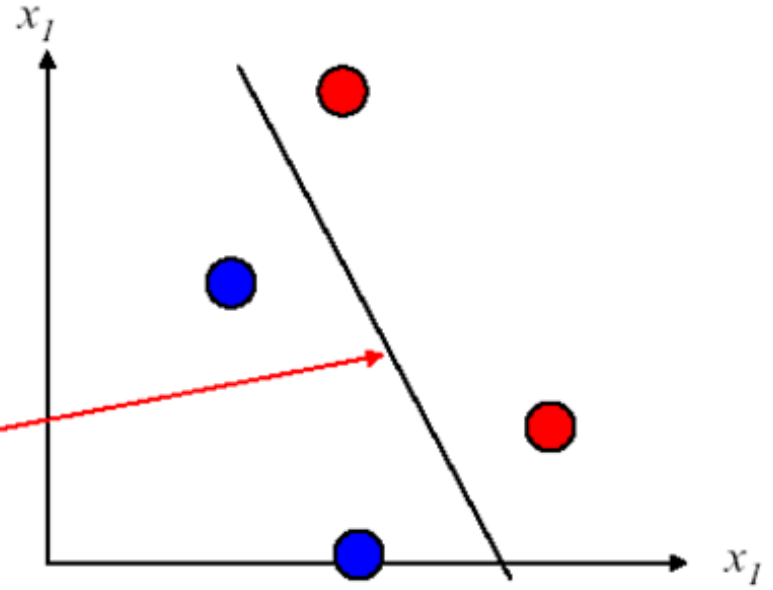
如果 $X^t X$ 是非奇异矩阵，则有：

$$a = (X^t X)^{-1} X^t b = X^+ b$$

X^+ 称为X的伪逆

Example of Linear Classifier by Pseudoinverse

- ω_1: $(1,2)^t$ and $(2,0)^t$
- ω_2: $(3,1)^t$ and $(2,3)^t$



Sample Matrix (d = 1+2, n = 4)

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 0 \\ -1 & -3 & -1 \\ -1 & -2 & -3 \end{bmatrix}$$

$$a^t \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix} = 0$$

Pseudo-inverse

$$X^* = (X^t X)^{-1} X^t = \begin{bmatrix} 5/4 & 13/12 & 3/4 & 7/12 \\ -1/2 & -1/6 & -1/2 & -1/6 \\ 0 & -1/3 & 0 & -1/3 \end{bmatrix}$$

Assuming $b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

our solution is $a = X^t b = \begin{bmatrix} 11/3 \\ -4/3 \\ -2/3 \end{bmatrix}$

如何对新样本进行分类（测试样本）？

$a.x > 0$ → First class

$a.x < 0$ → Second class

X ：新样本

面向多类的最小化均方误差（MSE）

$$B = \begin{bmatrix} B_1 \\ B_2 \\ \cdots \\ B_c \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

$$XA = B$$

$$A = X^{+}B$$

$$= inv(X'X)X'B$$

其他形式？

- 最小化均方误差方法的变形

$$for\ x_i \in \omega_1,\ a^t x_i > 0.$$

$$for\ x_i \in \omega_2,\ a^t x_i < 0.$$

将所有属于类别 ω_2 的样本乘以-1被称为标准化，标准化之后只有一个不等式 $a^t x_i > 0.$

- 找线性判别函数的问题可转变成最小化准则函数（损失函数）的问题

最小化准则函数的几个方法（优化方法）：

- Gradient Descent （梯度下降法）
- Newton's method & Quasi-Newton Method （牛顿法和拟牛顿法）
- Conjugate Gradient (共轭梯度法)

Gradient Descent Procedures

最小化准则函数：$J(\mathrm{a})$

a为解向量，那么梯度下降过程为：

$$\mathrm{a}(\mathrm{k}+1) = \mathrm{a}(\mathrm{k}) - \eta\,(\mathrm{k})\nabla J(\mathrm{a}(\mathrm{k}))$$

η 为一个标量，被称为学习率或学习步长，可以为固定值，也可随训练情况而改变。

- $\nabla J(\mathrm{a}(\mathrm{k}))$ 为 $J(\mathrm{a}(\mathrm{k}))$ 的梯度

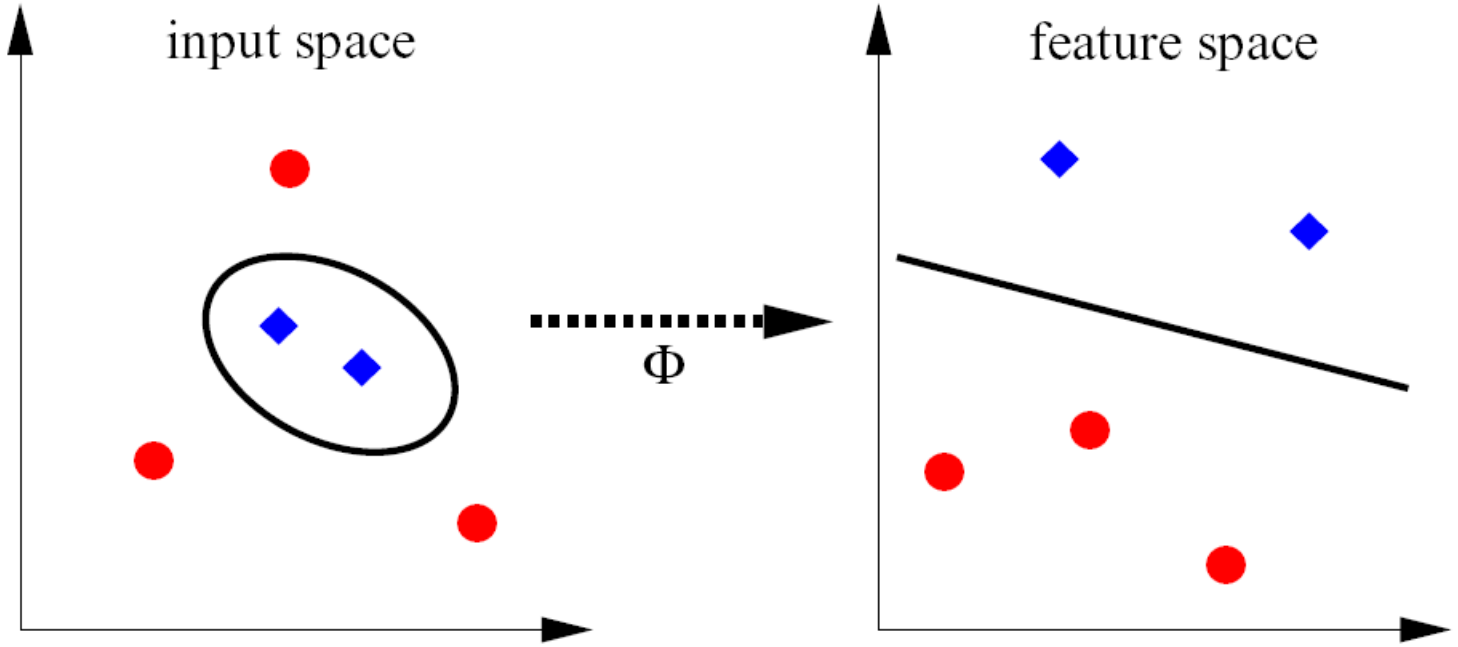
- 适用性：
  - 特别适用于没有解析解的问题

非线性最小化均方误差过程：

Just for your reference

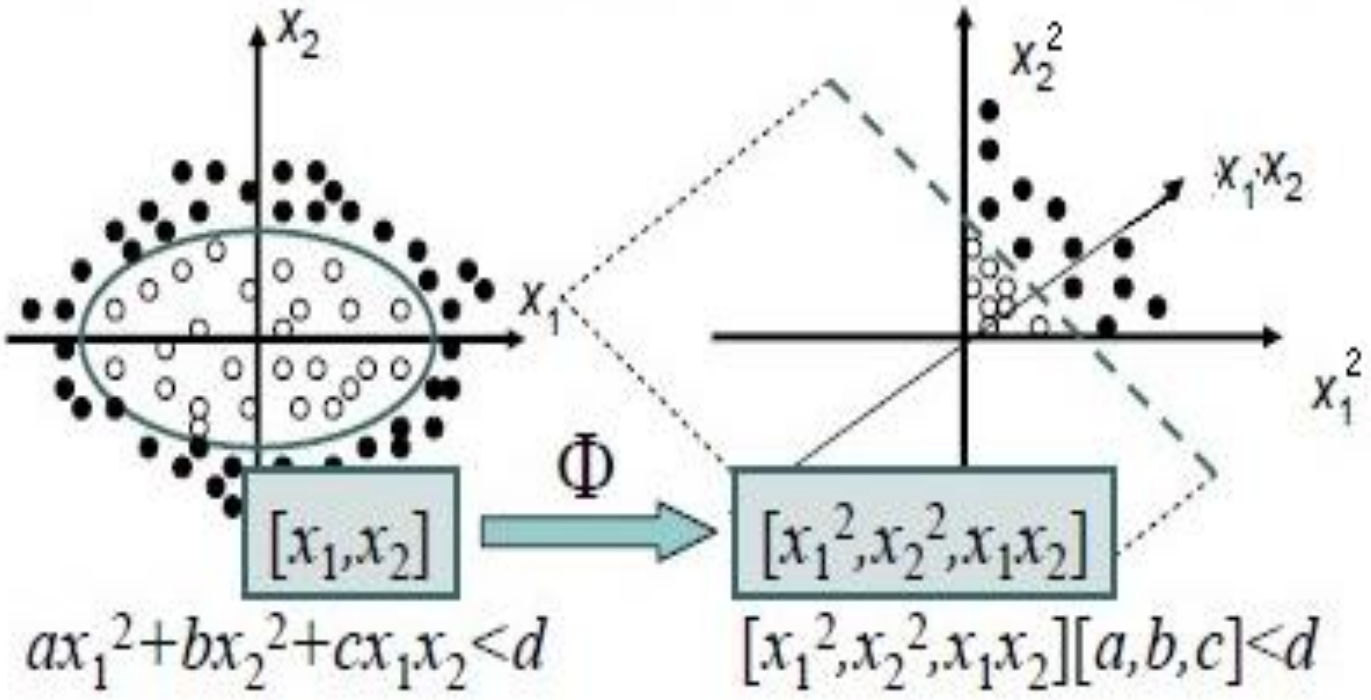
非线性最小化均方误差算法==

非线性变换+最小化均方误差算法



非线性变换的可能优势：将非线性可分问题转化为线性可分问题！

非线性最小化均方误差


x_2
x_1
$[x_1,x_2]$
Φ
x_2^2
x_1x_2
x_1^2
$[x_1^2,x_2^2,x_1x_2]$
$ax_1^2+bx_2^2+cx_1x_2<d$
$[x_1^2,x_2^2,x_1x_2][a,b,c]<d$

非线性最小化均方误差

在原始空间中最小均方误差：

$$\begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1d} \\ x_{20} & x_{21} & \cdots & x_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n0} & x_{n1} & \cdots & x_{nd} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \cdots \\ a_d \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \cdots \\ -1 \end{pmatrix} \Leftrightarrow Xa = B$$

新空间中最小化均方误差：

$$Z\beta = B \qquad Z = [Z_1 \ldots Z_n] \qquad Z_i = \varphi(X_i)$$

非线性最小化均方误差过程：KMSE

Because of

$$\beta = \sum_{j=1,\ldots,n} \gamma_j \varphi(\mathrm{X}_j)$$

$$\varphi(\mathrm{X}_i)^T \varphi(\mathrm{X}_j) = k(\mathrm{X}_i, \mathrm{X}_j)$$

we have

$$K\gamma = B$$

$$K = \begin{pmatrix} k(\mathrm{X}_1, X_1) & k(X_1, X_2) & \ldots & k(\mathrm{X}_1, X_n) \\ k(\mathrm{X}_2, X_1) & k(X_2, \mathrm{X}_2) & \ldots & k(X_2, X_n) \\ \ldots & \ldots & \ldots & \ldots \\ k(X_n, X_1) & k(X_n, \mathrm{X}_2) & \ldots & k(X_n, X_n) \end{pmatrix}$$

非线性最小化均方误差过程：KMSE

- 核函数:

- (1)

$$k(\mathrm{X}_i, \mathrm{X}_j) = \exp(-\frac{||\mathrm{X}_i - \mathrm{X}_j||^2}{\sigma})$$

(2) $$k(\mathrm{X}_i, \mathrm{X}_j) = (\mathrm{X}_i^T \mathrm{X}_j + c)^d$$

非线性最小化均方误差过程：KMSE

- 训练阶段:
- 得到 $$\gamma = K^{-1}B$$
- 测试阶段

$$b = \sum_{i=1}^{n} \gamma_i k(\mathrm{X}_i, \mathrm{X})$$

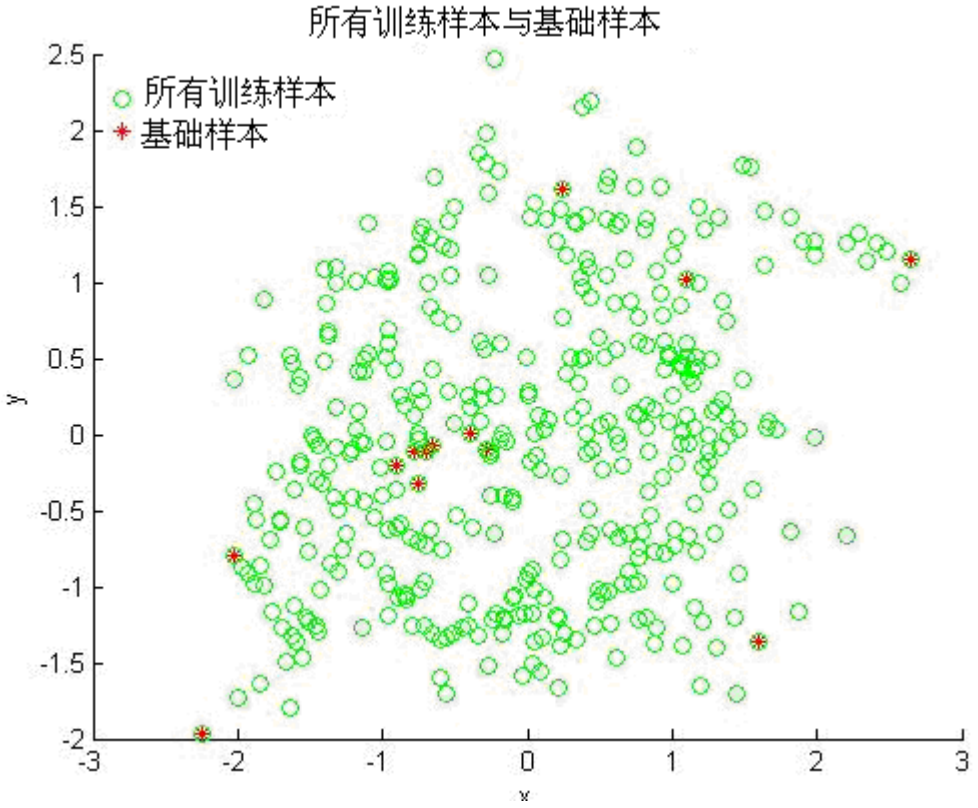
如果 b 更接近 1，就将测试样本分类为第一类；否者将其归为第二类.

KMSE的缺点与改进

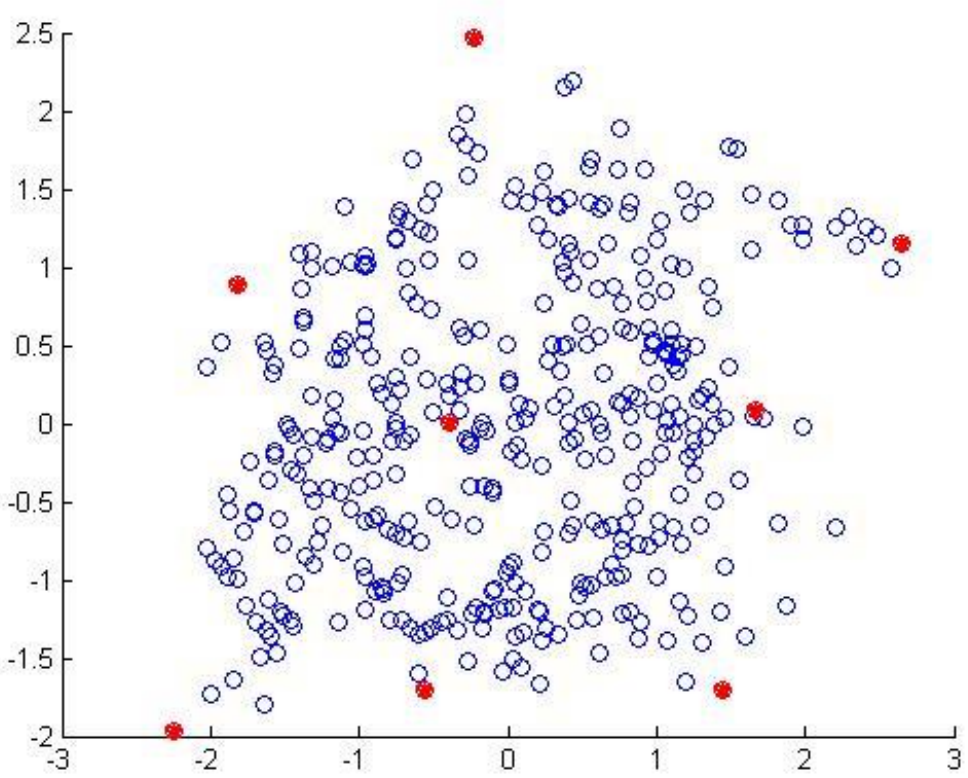
- 计算复杂度随着训练样本的增多快速升高!
- 如果 $\beta = \sum_{j=1,\ldots,s} \gamma'_j \varphi(\mathrm{X}_j), s << n$

$b = \sum_{i=1}^{s} \gamma'_i k(\mathrm{X}_i, \mathrm{X})$ 计算复杂度会大大减小

KMSE的缺点与改进



KMSE的缺点与改进



与线性判别分析的关系

- 选择合适的向量 b，MSE 判别函数 $a^t x$ 和线性判别将会等同

- 我们使用线性判别函数而不是广义线性判别函数

假定

- 有n个 d维样本集合 $x_1,\ldots x_n$，n_1 是标记为 w_1的子集 D_1 的个数，n_2 是标记为w_2的子集D_2的个数
- 如果通过添加阈值分量 $x_0=1$ 来构成一个增广模式向量，则从x_i得到一个样本y_i
- 如果样本属于w_2, 全部的模式向量（特征向量）乘以 -1
- 在没有损失的情况下，假设前 n_1 个样本标记为w_1，后个 n_2 个标记为w_2

- MSE 方法 Xa=b 等价于Fisher's Linear Discriminant

- 条件: 样本数目趋于无穷.

矩阵 Y 可以划分为:

$$Y = \begin{bmatrix} 1_1 & X_1 \\ -1_2 & -X_2 \end{bmatrix},$$

1_i 是n_i 个元素为1的列向量，X_i 是一个 n_i-by-d 矩阵，它的行为样本的标签 w_i.

$$a = \begin{bmatrix} \omega_0 \\ w \end{bmatrix}$$

$$b = \begin{bmatrix} \frac{n}{n_1} 1_1 \\ \frac{n}{n_2} 1_2 \end{bmatrix}$$

分块矩阵形式为：

$$\begin{bmatrix} 1_1^t & -1^t{}_2 \\ X^t{}_1 & -X^t{}_2 \end{bmatrix} \begin{bmatrix} 1_1 & X_1 \\ -1_2 & -X_2 \end{bmatrix} \begin{bmatrix} \omega_0 \\ w \end{bmatrix} = \begin{bmatrix} 1_1^t & -1^t{}_2 \\ X^t{}_1 & -X^t{}_2 \end{bmatrix} \begin{bmatrix} \frac{n}{n_1} 1_1 \\ \frac{n}{n_2} 1_2 \end{bmatrix}$$

样本均值

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x \qquad i = 1, 2$$

And the pooled sample scatter matrix

$$Sw = \sum_{i=1}^{2} \sum_{x \in D_i} (x - m_i)(x - m_i)^t \qquad (51)$$

相乘可以得到

$$\begin{bmatrix} n & (n_1 m_1 + n_2 m_2)^t \\ (n_1 m_1 + n_2 m_2) & S_w + n_1 m_1 m_1^t + n_2 m_2 m_2^t \end{bmatrix} \begin{bmatrix} \omega_0 \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ n(m_1 - m_2) \end{bmatrix}$$

这可以看作是一对方程，第一个方程可以用w_0 表示：

$$\omega_0 = -m^t w$$

m 是所有样本的均值.

用第二个方程代替可以得到：

$$\left[\frac{1}{n} S_w + \frac{n_1 n_2}{n^2} (m_1 - m_2)(m_1 - m_2)^t \right] w = m_1 - m_2$$

对任意的W，$(m_1-m_2)(m_1-m_2)^t w$ 和 m_1-m_2 同方向，则有：

$$\frac{n_1 n_2}{n^2}(m_1-m_2)(m_1-m_2)^t w=(1-a)(m_1-m_2),$$

a 是一个标量

可由 $$\left[\frac{1}{n}S_w+\frac{n_1 n_2}{n^2}(m_1-m_2)(m_1-m_2)^t\right]w=m_1-m_2$$

得到 $$w=S_w^{-1}(m_1-m_2)$$

除了一个不太重要的常量因子, 和 Fisher's linear discriminant. 是一样的

此外,可以得到权重阈值 w_0 ， 和判别准则：

Decide ω_1 if $w^t(x-m)>0$; otherwise decide ω_2 .