第三讲

模式识别

(Pattern Recognition)

武汉大学计算机学院

Email: zhanglefei@whu.edu.cn

第2章 线性判别分析(Cont.)

2.1 判别函数
2.2 线性判别函数
2.3 线性判别函数的性质
2.4 广义线性判别函数
2.5 线性分类器设计
2.6 分段线性分类器设计(*:不要求)
2.7 非线性分类器设计(*:不要求)

2.5 线性分类器设计

内容提要

一.梯度下降法—迭代法

二.感知器法(Perceptron)

三.最小平方误差准则—非迭代法

(MSE法: Mean Square Error, 均方误差法)

四.韦—霍氏法(Widrow-Hoff)—迭代法

(LMS法:Least Mean Squared,最小均方算法)

五.何—卡氏法(用于判断迭代过程中是否线性可分)

(Ho-Kashyap算法)

六.Fisher分类准则

七.线性判别分析编程举例

2.5 线性分类器设计

上一讲我们讨论了线性判别函数形式为：g(x)=$\boldsymbol{W}^{T}\boldsymbol{X}$

其中$\boldsymbol{X}$= $(X1, X2,\ldots, Xn,1)^{T}$ n维特征向量(增广特征向量)

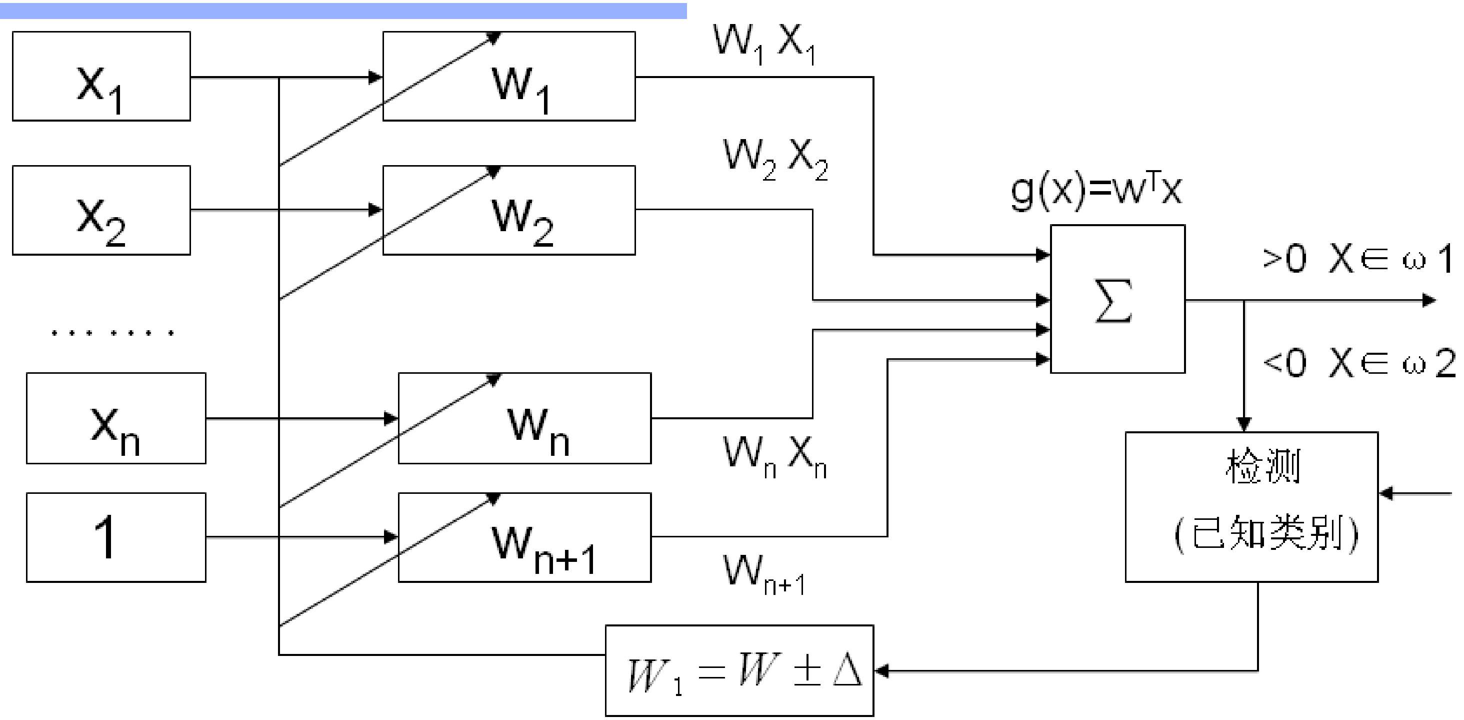
$\boldsymbol{W}$= $(W1, W2,\ldots,Wn , Wn+1)^{T}$ n维权向量(增广权向量)

$$\text{分类准则}\begin{cases} g(x)>0, x\in\omega_1 \\ g(x)<0, x\in\omega_2 \end{cases}$$

通常通过特征提取(抽取)可以获得n维特征向量，因此n维权向量是要求解的。

求解权向量的过程就是分类器的训练过程，使用已知类别的有限的学习样本来获得分类器的权向量被称为有监督的分类。

利用已知类别学习样本来获得权向量的训练过程如下：



已知$X_1 \in \omega_1$，通过检测调整权向量，最终使$X_1 \in \omega_1$；
已知$X_2 \in \omega_2$，通过检测调整权向量，最终使$X_2 \in \omega_2$。
这样就可以通过有限的样本去决定权向量。

利用方程组来求解权向量

对二类判别函数$g(x) = W_1X_1 + W_2X_2 + W_3$

已知训练集：X_a, X_b, X_c, X_d且

当$(X_a, X_b) \in \omega_1$时 $g(x) > 0$

当$(X_c, X_d) \in \omega_2$时 $g(x) < 0$

设$X_a = (X_{1a}, X_{2a})^T$ $X_b = (X_{1b}, X_{2b})^T$

$X_c = (X_{1c}, X_{2c})^T$ $X_d = (X_{1d}, X_{2d})^T$

判别函数可联立成：

$X_{1a}W_1 + X_{2a}W_2 + W_3 > 0$ ①

$X_{1b}W_1 + X_{2b}W_2 + W_3 > 0$ ②

$X_{1c}W_1 + X_{2c}W_2 + W_3 < 0$ ③

$X_{1d}W_1 + X_{2d}W_2 + W_3 < 0$ ④

求出W_1, W_2, W_3

将③ ④式正规化，得：

$-X_{1c}W_1 - X_{2c}W_2 - W_3 > 0$

$-X_{1d}W_1 - X_{2d}W_2 - W_3 > 0$

因此$g(x) = W^T X > 0$， 其中$W = (W_1, W_2, W_3)^T$

$$X = \begin{bmatrix} X_{1a} & X_{2a} & 1 \\ X_{1b} & X_{2b} & 1 \\ -X_{1c} & -X_{2c} & -1 \\ -X_{1d} & -X_{2d} & -1 \end{bmatrix}$$ 为各模式增广矩阵

为N*(n+1)矩阵(此处N=4, n=2)

N为样本数，n为特征数

训练过程就是对已知类别的样本集求解权向量**W**，这是一个线性联立不等式方程组求解的过程。

求解时：

①只有对线性可分的问题，$g(x) = \mathbf{W}^T\mathbf{X}$才有解；

②联立方程的解是非单值，在不同条件下，有不同的解，所以就产生了求最优解问题；

③求解**W**的过程就是训练的过程。训练方法的共同点是，先给出准则函数(Criterion function)，再寻找使准则函数趋于极值的优化算法，不同的算法有不同的准则函数。算法可以分为迭代法和非迭代法。

一. 梯度下降法—迭代法

要对不等式方程组$W^TX>0$求解，首先定义准则函数(目标函数)$J(W)$，再求$J(W)$的极值使W优化。因此求解权向量的问题就转化为对一目标函数求极值的问题。解决此类问题的方法是梯度下降法(1847年，法国数学家Cauchy提出)。

方法就是从起始值W_1开始，算出W_1处目标函数的梯度向量$\nabla J(W_1)$，则下一步的W值为：

$$W_2=W_1-\rho_1\nabla J(W_1)$$

W_1为起始权向量　　　ρ_1为迭代步长

$J(W_1)$为目标函数

$\nabla J(W_1)$为W_1处的目标函数的梯度向量

在第k步的时候

$W_{k+1} = W_k - \rho_k \nabla J(W_k)$　ρ_k为正比例因子(learning rate)

这就是梯度下降法的迭代公式。这样一步步迭代就可以收敛于解向量，ρ_k取值很重要

ρ_k太大，迭代太快，引起振荡，甚至发散。

ρ_k太小，迭代太慢。

应该选最佳ρ_k。

选最佳ρ_k

目标函数J(W)二阶Taylor级数展开式为：

$$J(W) \approx J(W_k) + \nabla J^T (W - W_k) + 1/2 (W - W_k)^T D (W - W_k) \quad (1)$$

其中D为当$W = W_k$时J(W)的二阶偏导数矩阵

将$W = W_{k+1} = W_k - \rho_k \nabla J(W_k)$代入(1)式得：

$$J(W_{k+1}) \approx J(W_k) - \rho_k \|\nabla J\|^2 + 1/2 \rho_k^2 \nabla J^T D \nabla J \quad (2)$$

其中$\nabla J = \nabla J(W_k)$

对ρ_k求导数，并令导数为零有

最佳步长为$\rho_k = \|\nabla J\|^2 / (\nabla J^T D \nabla J)$

这就是最佳ρ_k的计算公式，但因二阶偏导数矩阵D的计算量太大，因此此公式很少用。

若令$W=W_{k+1}$，(1)式为

$$J(W_{k+1})=J(W_k)+\nabla J^T(W_{k+1}-W_k)+1/2(W_{k+1}-W_k)^T D(W_{k+1}-W_k)$$

对W_{k+1}求导，并令导数为零可得：

最佳迭代公式：$W_{k+1}= W_k- D^{-1}\nabla J$ —牛顿法的迭代公式

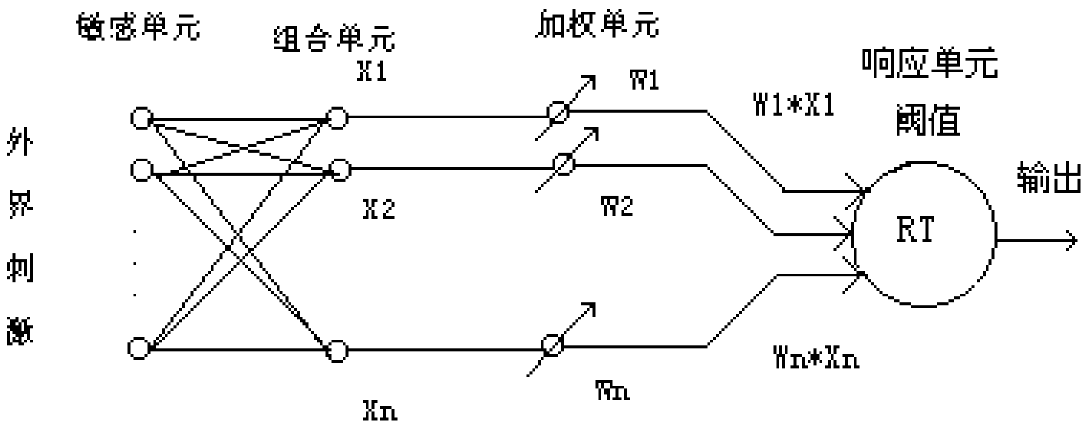
D^{-1}是D的逆阵

讨论：牛顿法是基于二阶Taylor级数且比梯度下降法收敛更快，但是D的计算量大并且要计算D^{-1}。当D为奇异矩阵时，无法用牛顿法。

二. 感知器法（Perceptron）

感知器是由美国Rosenblatt于1957年提出的有监督学习(有导师学习)神经网络模型，其学习算法于1958年提出。

感知器原理结构：



nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last – a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO
CONSERVATION
SONGBIRDS À LA CARTE
Illegal harvest of millions of Mediterranean birds
PAGE 452
RESEARCH ETHICS
SAFEGUARD TRANSPARENCY
Don't let openness backfire on individuals
PAGE 459
POPULAR SCIENCE
WHEN GENES GOT 'SELFISH'
Dawkins's calling card forty years on
PAGE 462
NATURE.COM/NATURE
28 January 2016 £10
Vol. 529, No. 7587

通过对W的调整，可实现判别函数$g(x) = W^T X > R_T$

其中R_T为响应阈值

定义感知准则函数：只考虑错分样本

定义：$J(W)=\sum_{X\in S_e}\left(-W^T X\right)$ 其中S_e为被W错误分类样本集合

当分类发生错误时就有$W^T X < 0$，或$-W^T X > 0$，所以J(W) 总是正值。错误分类愈少， J(W)就愈小。

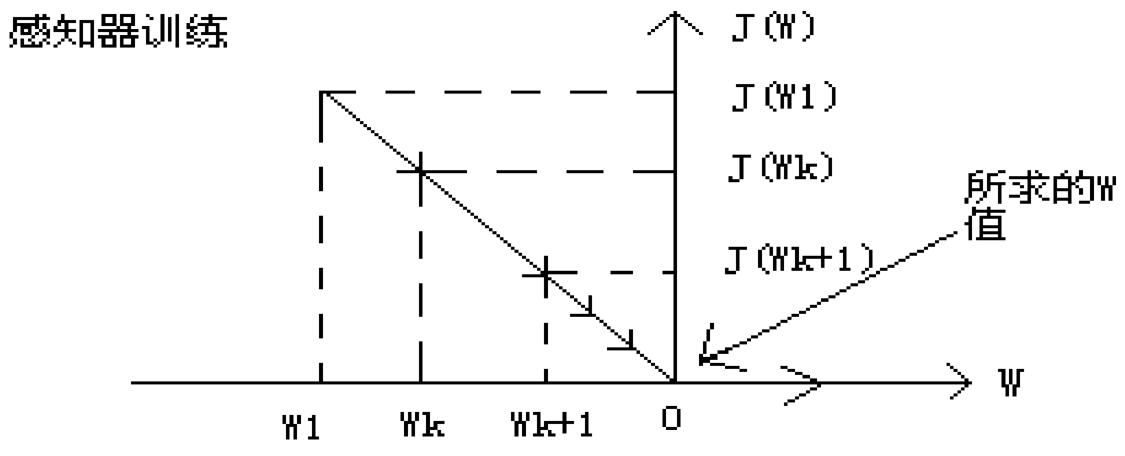
理想情况为J(W)=0即求最小值的问题。

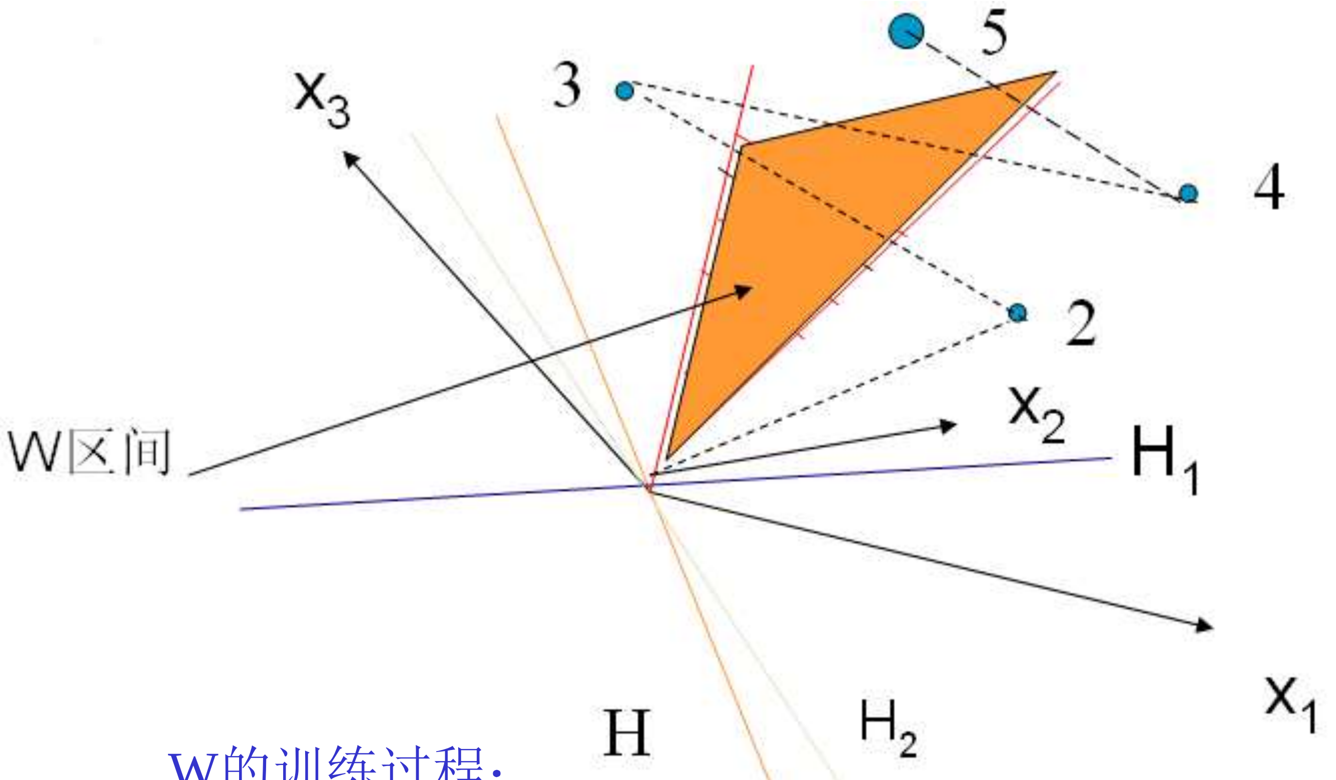
求最小值对W求梯度 $\nabla J=\frac{\partial J(W)}{\partial W}=\sum_{X\in S_e}(-X)$

代入迭代公式中 $W_{k+1}=W_k-\rho_k \nabla J$

即感知器迭代公式 $W_{k+1}=W_k+\rho_k\sum_{X\in S_e}X$

由J(W)经第k+1次迭代的时候，J(W)趋于0，收敛于所求的W值。





W的训练过程：

例如，$x_1, x_2, x_3 \in \omega_1$，作 x_1, x_3的垂直线可得解区(如图)

假设起始权向量w_1=0 $\rho_k = 1$

1. x_1, x_2, x_3三个向量相加得向量2,垂直于向量2的超平面H将x_3错分。
2. x_3与向量2相加得向量3,垂直于向量3的超平面H_1，将x_1错分。
3. 依上法得向量4,垂直于向量4做超平面，H_2将x_3错分。
4. x_3与向量4相加得向量5，向量5在解区内，垂直于向量5的超平面可以把 x_1, x_2, x_3分成一类 。

- **感知器算法：**

1.错误分类，修正w_k

若$w_k^T x \le 0$并且$x \in \omega_1$　　$w_{k+1} = w_k + \rho_k x$

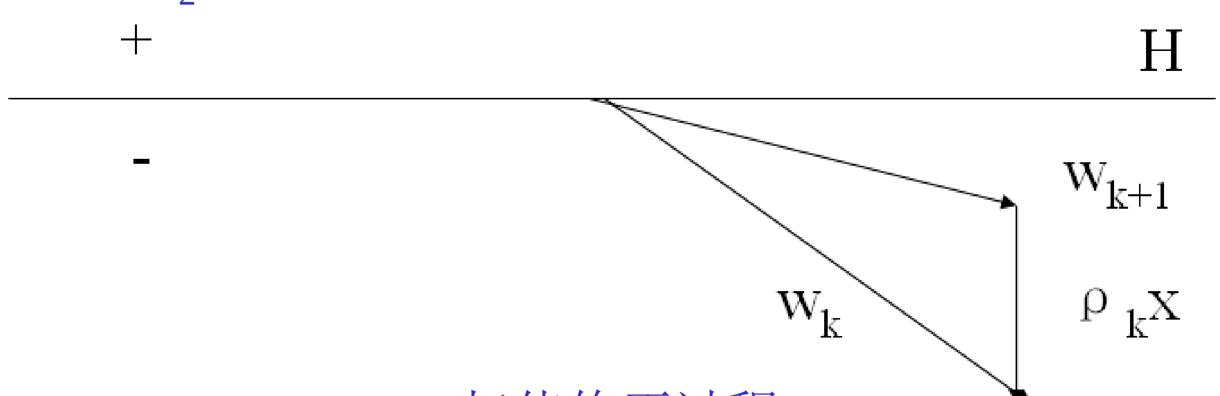
若$w_k^T x \ge 0$并且$x \in \omega_2$　　$w_{k+1} = w_k - \rho_k x$

2.正确分类，w_k不修正

若$w_k^T x > 0$并且$x \in \omega_1$

若$w_k^T x < 0$并且$x \in \omega_2$

$w_{k+1} = w_k$



权值修正过程

- ρ_k选择准则

① 固定增量原则 ρ_k固定非负数

② 绝对修正规则 $\rho_k > \dfrac{|w^T x|}{x^T x}$

③ 部分修正规则 $\rho_k = \lambda \dfrac{|w^T x|}{x^T x}$ $\quad 0 < \lambda \leq 2$

例：有两类样本

$\omega_1=(x_1,x_2)=\{(1,0,1)^T,(0,1,1)^T\}$

$\omega_2=(x_3,x_4)=\{(1,1,0)^T,(0,1,0)^T\}$

解：先求四个样本的增广模式

$x_1=(1,0,1,1)^T \qquad x_2=(0,1,1,1)^T$

$x_3=(1,1,0,1)^T \qquad x_4=(0,1,0,1)^T$

假设初始权向量 $w_1=(1,1,1,1)^T \qquad \rho_k=1$

第一次迭代：

$w_1^Tx_1=(1,1,1,1)\ (1,0,1,1)^T=3>0$ 所以不修正w_1

$w_1^Tx_2=(1,1,1,1)\ (0,1,1,1)^T=3>0$ 所以不修正w_1

$w_1^Tx_3=(1,1,1,1)\ (1,1,0,1)^T=3>0$ 所以修正w_1

$w_2=w_1-x_3=(0,0,1,0)$

$w_2^Tx_4=(0,0,1,0)^T\ (0,1,0,1)=0$ 所以修正w_2

$w_3=w_2-x_4=(0,-1,1,-1)$

第一次迭代后,权向量w3=(0,-1,1,-1)，再进行第2，3，…次迭代，如下表所示：

训练样本	$w_k^T x$	修正式	修正后的权值w_{k+1}	迭代次数
x_1 1 0 1 1 x_2 0 1 1 1 x_3 1 1 0 1 x_4 0 1 0 1	+ + + 0	w_1 w_1 w_1-x_3 w_2-x_4	1 1 1 1 1 1 1 1 0 0 1 0 0 –1 1 -1	1
x_1 1 0 1 1 x_2 0 1 1 1 x_3 1 1 0 1 x_4 0 1 0 1	0 + 0 -	w_3+x_1 w_4 w_4-x_3 w_5	1 –1 2 0 1 –1 2 0 0 –2 2 –1 0 –2 2 -1	2
x_1 1 0 1 1 x_2 0 1 1 1 x_3 1 1 0 1 x_4 0 1 0 1	+ - - -	w_5 w_5+x_2 w_6 w_6	0 –2 2 –1 0 –1 3 0 0 –1 3 0 0 –1 3 0	3
x_1 1 0 1 1 x_2 0 1 1 1 x_3 1 1 0 1 x_4 0 1 0 1	+ + - -	w_6 w_6 w_6 w_6	0 –1 3 0 0 –1 3 0 0 –1 3 0 0 –1 3 0	4

直到在一个迭代过程中权向量相同，训练结束。

$w_6=w=(0,-1,3,0)$， 判别函数$g(x)= -x_2+3x_3$

- 感知器算法只对线性可分样本有收敛的解，对非线性可分样本集会造成训练过程的振荡，这是它的缺点。

线性不可分样本集的分类解(取近似解)

对于线性可分的样本集，可用感知器法解出正确分类的权向量。当样本集线性不可分时，用上述感知器法求权值时算法不收敛。如果我们把循环的权向量取平均值作为待求的权向量，或就取其中之一为权向量，一般可以解到较满意的近似结果。

例：现有样本(见右图)

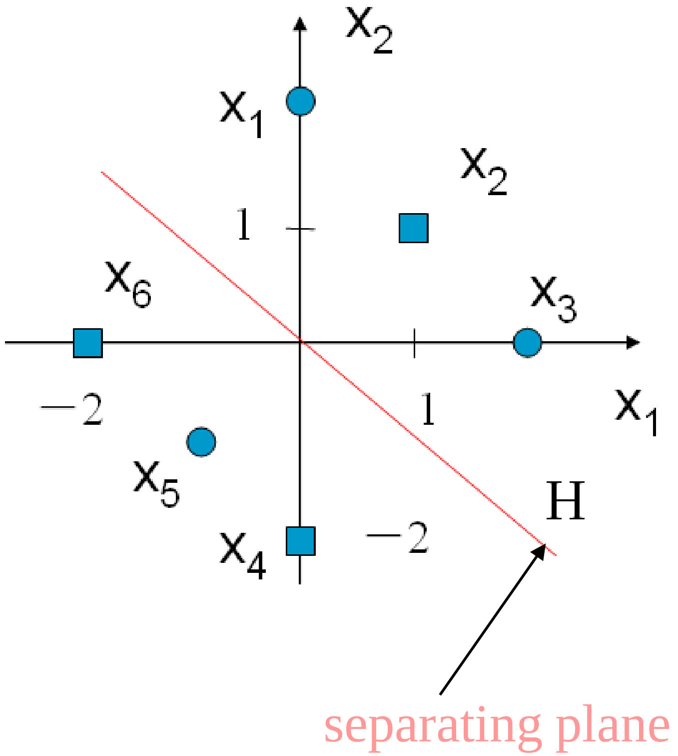
ω1：$X1=(0,2)^T$, $X3=(2,0)^T$

$X5=(-1,-1)^T$

ω2：$X2=(1,1)^T$, $X4=(0,-2)^T$

$X6=(-2,0)^T$

试求权向量的近似解。



解：此为线性不可分问题，利用感知器法求权向量。

权向量产生循环(-1, 2, 0), (0, 2, 2), (-1, 1, 1), (-1, 1, 1) (-1, 1, 1), (0, 0, 0), (-1, 2, 0)

因此算法不收敛，我们可以取循环中任一权值，例如取$W=(0,2,2)^T$

则判别函数为： $g(x)= 2x_1+2x_2$

判别面方程为： $g(x)= 2x_1+2x_2＝0$ 所以$x_1+x_2＝0$

由图看出判别面H把二类分开，但其中x_2错分到ω_1类，而x_1错分到ω_2类，但大部分分类还是正确的。

➢三. 最小平方误差准则——非迭代法

(MSE法: Mean Square Error, 均方误差法)

前面研究了线性不等式方程组 $g(x)=W^TX>0$ 的解法。它们共同点是企图找一个权向量W，使错分样本最小。

现将不等式组变成右边的形式：$W^TX_i=b_i>0$

$$令X=\left(X_1,X_2,\ .....\ X_m\right)^T$$

$$b=\left(b_1,b_2,\ .....\ b_m\right)^T 为给定的任意正常数$$

$$X=\begin{bmatrix} X_1 \\ X_2 \\ ..... \\ X_m \end{bmatrix}=\begin{bmatrix} X_{11} & X_{12} & .... & X_{1n} \\ X_{21} & \cdots & \cdots & X_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}$$

每个样本有n个特征

(通常m>n，即行比列多)

则有联立方程XW=b 。这是矛盾方程组，方程数大于未知数，通常没有精确解。

可寻找一个权向量W，使得某个关于XW和b的函数最小化。我们定义一个误差向量：

$e=XW-b\neq 0$，把平方误差作为目标函数

$$J(W)=\|e\|^2=\|XW-b\|^2=\sum_{i=1}^{m}\left(W^T X_i - b_i\right)^2$$ MSE准则函数

W的优化就是使J(W)最小。求J(W)的梯度并为0。

$$\nabla J(W)=\sum_{i=1}^{m}2\left(W^T X_i - b_i\right) X_i = 2X^T(XW-b)=0$$

解上方程得 $X^TXW=X^Tb$

这样把求解XW=b的问题，转化为对$X^TXW=X^Tb$求解，这

一有名方程的最大好处是由于X^TX为方阵且通常是非奇异的，因此可以得到W的唯一解。

$$W=\left(X^TX\right)^{-1}X^T\mathbf{b}=X^+\mathbf{b}$$

其中$X^+=\left(X^TX\right)^{-1}X^T$为$X$的伪逆(称其为规范矩阵)---"矩阵论"中伪逆的推论

只要计算出**X⁺**就可以得到**W**

MSE的解是由**b**决定的。若取**b**为一个任意固定的值，可以通过最小化平方误差准则函数得到一个在可分和不可分情况下都很有用的判别函数。如**b**可以取：

$$\mathbf{b}=\begin{bmatrix} m/m_1 \\ \cdots\cdots\cdots \\ m/m_1 \\ m/m_2 \\ \cdots\cdots \\ m/m_2 \end{bmatrix}$$

其中m/m_1有m_1个，m/m_2有m_2个

(m=m1+m2)

选择其他的**b**当然会得到不同的判决边界。

最小平方误差法同后面要介绍的Fisher法是一致的。

四.韦—霍氏法(Widrow-Hoff)—迭代法

(LMS法:Least Mean Squared,最小均方算法)

上节得到MSE法的W解为：$W=X^{+}b$

伪逆 $X^{+}=\left(X^{T}X\right)^{-1}X^{T}$ 计算量大

在计算X^{+}时，

1.要求$X^{T}X$矩阵为非奇异

2.由于计算量太大而引入了较大误差

因此考虑用迭代法来求解：

求J(W)的梯度(See slide 26)

$\nabla J(W)=2X^{T}(XW-b)$ 代入迭代公式

W_1任意设定

$W_{k+1}=W_k-\rho_k X^{T}(XW_k-b)$

因此下降算法不论X^TX是否奇异，总能产生一个解。若训练样本无限的重复出现，则简化为：

W_1任意

$W_{k+1}=W_k+\rho_k(b_k-W_k^TX^k)X^k$

取$\rho_K=\dfrac{\rho_1}{k}$

ρ_k随迭代次数k的增大而减少，以保证算法收敛于满意的W值。

五.何—卡氏法(用于判断迭代过程中是否线性可分)(Ho-Kashyap算法)

若训练样本线性可分时，感知器法可求出界面，但对不可分问题由于不收敛只能取平均。最小平方误差法不论样本是否线性可分都能给出一加权向量，但不能保证此向量就是分界向量，这里介绍一种方法可以检测迭代过程中是否线性可分。

- 因最小平方误差法的J(W)的解为

$$W = \left(X^T X\right)^{-1} X^T b = X^+ b$$

对b前后两次迭代后，$b_{k+1} = b_k + \delta b_k$

- 由于**XW=b b>0**

因此，b的增量可表示为$\delta b_k = c[XW_k - b_k + |XW_k - b_k|]$ (1)

- c为校正系数

当$(XW_k - b_k) \leq 0$ 时 $\delta b_k = 0$

当$(XW_k - b_k) > 0$ 时 $\delta b_k = 2c[XW_k - b_k]$

- 引入误差向量e_k

$e_k=XW_k-b_k$判断是否线性可分

因此，(1)式可以写成：$\delta b_k = c[e_k + |e_k|]$

- 所以J(W)的解为

$$W_{k+1} = X^+ b_{K+1} = X^+[b_K + \delta b_k] = X^+ b_K + X^+ \delta b_k = W_k + c X^+[e_K + |e_k|]$$

初始条件$W_1=X^+b_1$并且$b_1>0$

- 迭代时检测

若$e_k \ge 0$时，XW >b，系统线性可分，迭代收敛；

若$e_k<0$时，XW <b，系统线性不可分，迭代不收敛。

下面举例说明e_k的作用

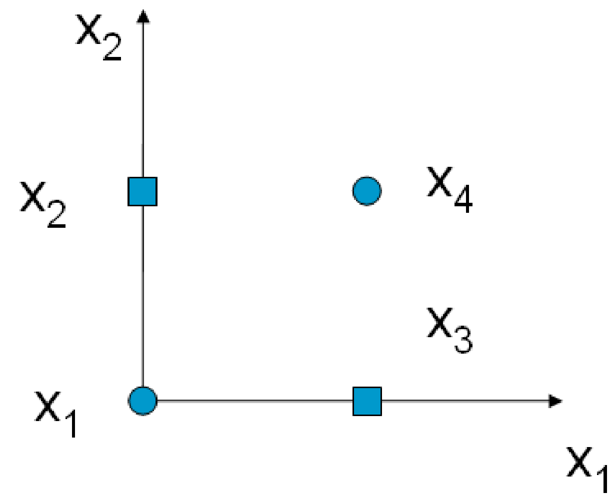
- 例：

  $\omega_1=\{(0,0)^T,(0,1)^T\}$　　$\omega_2=\{(1,0)^T,(1,1)^T\}$

- 解：正规化

  对ω_2取负，有

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



X的规范矩阵为

$$X^+=\left(X^TX\right)^{-1}X^T=\frac{1}{2}\begin{bmatrix} -1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ 3/2 & 1/2 & -1/2 & 1/2 \end{bmatrix}$$

MATLAB Pseudo inverse： X^+=pinv(X)

取$b_1=(1,1,1,1)^T$ $c=1$ $W_1=X^+b_1=(-2,0,1)^T$

$$XW_1=\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & -1 \\ -1 & -1 & -1 \end{bmatrix}\begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}=\left(1,1,1,1\right)^T>0$$

所以W_1为所求解 $e_1=XW_1-b_1=0$ 系统线性可分

若四个样本变成：

$\omega_1=\{(0,0)^T,(1,1)^T\}$　　$\omega_2=\{(0,1)^T,(1,0)^T\}$

解：

$$X=\begin{bmatrix}0 & 0 & 1\\ 1 & 1 & 1\\ 0 & -1 & -1\\ -1 & 0 & -1\end{bmatrix}\qquad X^{+}=\frac{1}{2}\begin{bmatrix}-1 & 1 & 1 & -1\\ -1 & 1 & -1 & 1\\ 3/2 & -1/2 & -1/2 & -1/2\end{bmatrix}$$

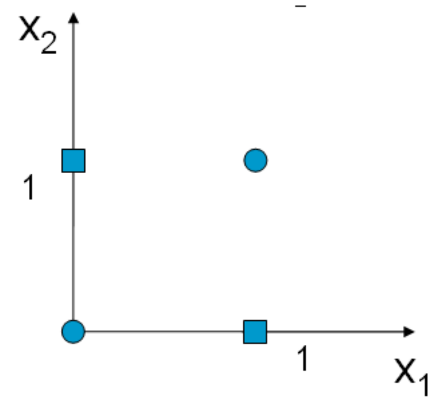
取$b_1=(1,1,1,1)^T$　$c=1$

$W_1=X^{+}b_1=(0,0,0)^T$

$e_1=XW_1-b_1=(-1,-1,-1,-1)^T<0$

系统线性不可分

c为校正系数，取$0<c\le 1$



在算法进行过程中，应在每一次迭代时，检测e_k的值。只要出现$e_k<0$，迭代应立即停止。

六. Fisher分类准则

Fisher(1890-1962, 英国数学家，生物学家，现代统计学奠基人之一。证明了孟德尔的遗传律符合达尔文的进化论。

Fisher判别是一种应用极为广泛的线性分类方法(1936)。

- Fisher判别的基本思想：

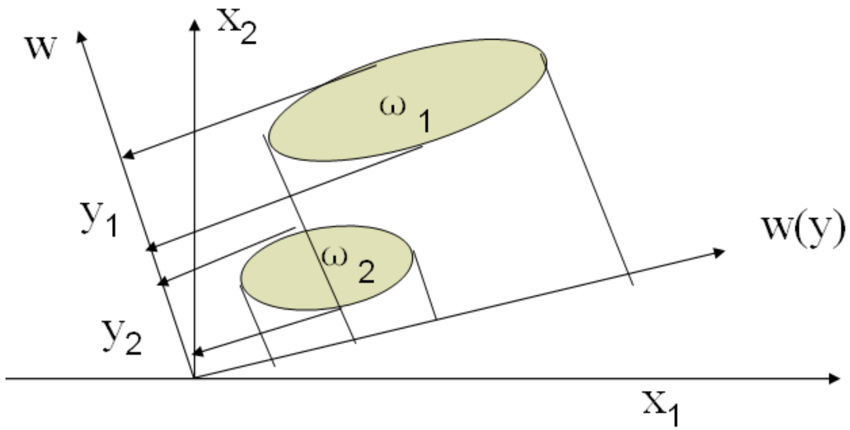
  希望投影后的一维数据满足：

  - 两类之间的距离尽可能远；
  - 每一类自身尽可能紧凑。

- 准则的描述：

  - 用投影后数据的统计性质—均值和离散度的函数作为判别优劣的标准。

下面讨论通过映射投影来降低维数的方法。



X空间 $X=W^TX-W0>0$ $X\in\omega_1$

$X=W^TX-W0<0$ $X\in\omega_2$

映射Y空间 $Y=W^TX-W0>0$ $X\in\omega_1$

$Y=W^TX-W0<0$ $X\in\omega_2$

把X空间各点投影到Y空间得一直线上，维数由2维降为1维。若适当选择W的方向，可以使二类分开。下面我们从数学上寻找最好的投影方向，即寻找最好的变换向量W的问题。

在X空间的均值向量: $\overline{X_i} = \frac{1}{N_i}\sum_{x\in\omega_i} X$

N_i是ω_i的样本数 (i=1,2)

在Y空间的投影均值: $\overline{Y_i} = \frac{1}{N_i}\sum_{Y\in\omega_i} Y = \frac{1}{N_i}\sum_{X\in\omega_i} W^T X = W^T \overline{X_i}$

$$\therefore \overline{Y_1} = W^T\,\overline{X_1} \qquad \overline{Y_2} = W^T\,\overline{X_2}$$

投影样本之间的分离性用投影样本均值之差表示

$\boxed{|\overline{Y_1}-\overline{Y_2}| = |W^T(\overline{X_1}-\overline{X_2})|}$ 类间分离性越大越好

投影样本类内离散度:

$$\sigma_i^2 = \sum_{Y\in\omega_i}\left(Y-\overline{Y_i}\right)^2 = \sum_{X\in\omega_i}\left(W^T X - W^T\overline{X_i}\right)^2 = W^T S_i W$$

其中 $\boxed{S_i = \sum_{X\in\omega_i}\left(X-\overline{X_i}\right)\left(X-\overline{X_i}\right)^T\text{—样本类内离散度矩阵}}$

$$\sigma_1^2 = W^T S_1 W \qquad \sigma_2^2 = W^T S_2 W$$

注: $s_i=cov(X_i)*(N_i-1)$

$$S_1=\sum_{X\in\omega_1}\left(X-\overline{X_1}\right)\left(X-\overline{X_1}\right)^T \qquad S_2=\sum_{X\in\omega_2}\left(X-\overline{X_2}\right)\left(X-\overline{X_2}\right)^T$$

投影样本总的离散度可用$(\sigma_1{}^2+\sigma_2{}^2)$来表示，要求投影样本的总离散度越小越好。

*Fisher*准则函数的构造 $\boxed{J(W)=\dfrac{\left|\overline{Y_1}-\overline{Y_2}\right|^2}{\left(\sigma_1{}^2+\sigma_2{}^2\right)}}$

类间距

总类内离散度

$J(W)$的分子$\left(\overline{Y_1}-\overline{Y_2}\right)^2=\left(W^T\bar{X}_1-W^T\bar{X}_2\right)^2=W_TS_bW$

其中 $\boxed{S_b=\left(\overline{X_1}-\overline{X_2}\right)\left(\overline{X_1}-\overline{X_2}\right)^T-\text{样本类间离散度矩阵}}$

$J(W)$的分母$\sigma_1{}^2+\sigma_2{}^2=W^TS_1W+W^TS_2W=W^TS_wW$

$\boxed{S_w=S_1+S_2--S_w\text{为总的类内离散度矩阵}}$

所以Fisher准则函数为$J(W)=\dfrac{W^T S_b W}{W^T S_w W}$

其中 S_w为类内离散度矩阵, S_b为类间离散度矩阵

对$J(W)$求极大值, 得 $\boxed{W^* = S_w^{-1}\left(\overline{X}_1 - \overline{X}_2\right)\ \text{教材P57，推导过程此略}}$

上式就是n维X空间向一维Y空间的最好投影方向，它实际上是多维空间向一维空间的一种映射(W^*就是使模式样本的投影在类间最分散、类内最集中的最优解)。

$$\boxed{S_t = S_w + S_b \text{ -- } S_t\text{为总体离散度矩阵}}$$

这样就把一个n维的问题转化为一维的问题。
现在一维空间设计 Fisher分类器：

$$Y = W^{*T}X > W_0 \Rightarrow X \in \omega_1$$

$$Y = W^{*T}X < W_0 \Rightarrow X \in \omega_2$$

W_0阈值的选择

$$1. W_0 = \frac{\overline{Y}_1 + \overline{Y}_2}{2}$$

$$2. W_0 = \frac{N_1\overline{Y}_1 + N_2\overline{Y}_2}{N_1 + N_2} = \frac{N_1 W^T \overline{X}_1 + N_2 W^T \overline{X}_2}{N_1 + N_2}$$

$$3.W_0=\overline{Y}_1+(\overline{Y}_2-\overline{Y}_1)\frac{\sum_{k=1}^{N_1}(Y_{k1}-\overline{Y}_1)^2}{\sum_{k=1}^{N_1}(Y_{k1}-\overline{Y}_1)^2+\sum_{k=1}^{N_2}(Y_{k2}-\overline{Y}_2)^2}$$

Y_{ki}表示第i类中第k个样本的投影值

N_1为ω_1样本数　　N_2为ω_2样本数

当W_0选定后，对任一样本X，只要判断$Y=W^{*T}X>W_0$，则

$X\in\omega_1$；　$Y=W^{*T}X<W_0$，则$X\in\omega_2$；分类问题就解决了。

两类问题Fisher算法实现步骤小结：

(1)求两类样本均值向量 $\bar{\mathbf{x}}_1$和$\bar{\mathbf{x}}_2$；

(2)求两类样本类内离散度矩阵S_i；

(3)求总的类间散度矩阵$S_w = S_1 + S_2$;

(4)求向量W*，$W^* = S_w^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$；

(5)对两类已知样本，求出它们在W*上的投影点y_i:

$$y_1 = W^{*T} X_1, y_2 = W^{*T} X_2$$

(6)求各类样本的均值 $\bar{Y}_i, \bar{Y}_i = \frac{1}{N_i} \sum_{Y \in \omega_i} Y$；

(7)选取阈值W_0，如取 $W_0 = \frac{\bar{Y}_1 + \bar{Y}_2}{2}$；

(8)对未知样本X，计算它在W*上的投影点y: $y = \mathbf{W}^{*T} \mathbf{X}$;

(9)根据判别规则对未知样本X进行分类：$Y = W^{*T} X > W_0$，则$X \in \omega_1$；　　$Y = W^{*T} X < W_0$，则$X \in \omega_2$。

广义Fisher准则(: 了解，用到时再具体学)

基于两类问题的Fisher分类准则，能很容易地扩展为多类问题的Fisher准则，又称广义Fisher准则。在d维空间中，对于M类问题，一定存在(M-1)个线性判别函数。因此广义Fisher准则所涉及的投影问题是：在d维空间中，将M类样本集合投影在(M-1)维空间上(在此假设，$M \le d$，即类别数小于特征空间的维数)。

多类问题的(M-1)个判别函数为

$$y_i = \mathbf{w}_i^T \mathbf{x},\ i = 1, 2, \ldots, M-1$$

或者写为矩阵形式

$$\mathbf{y} = \mathrm{W}^T \mathbf{x}$$

其中, $\mathrm{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{M-1}]$, $\mathbf{y} = [y_1, y_2, \ldots, y_{M-1}]^T$

因此，广义Fisher准则所要解决的问题是：寻找M类样本集合**x**在W张成的(M-1)维空间的最佳投影向量集合W*。

Fisher判别分析举例:

1. 问题

两种蠓Af和Apf已由生物学家根据它们的触角和翼长加以区分，两个矩阵中分别给出了6只Apf 和9只Af蠓的触角长(对应于矩阵的第1列)和翼长(对应于矩阵的第2列)的数据(See next slide)。根据触角长和翼长这两个特征来识别一个标本是Af还是Apf是重要的。

(1)试给出该问题的Fisher分类器;

(2)有三个待识别的标本，它们分别是$(1.24, 1.80)^T$, $(1.28, 1.84)^T$, $(1.40, 2.04)^T$，试问这三个标本属于哪一种蠓。

2.假设

(1)两种群Apf 和Af的两个特征的期望值、标准差、相关系数与由矩阵数据给出的样本统计量一致；

(2)两种群Apf 和Af的两个特征服从二元正态分布；

(3)所给样本数据无误差。

蠓Apf

$$\begin{bmatrix} 1.14 & 1.78 \\ 1.18 & 1.96 \\ 1.20 & 1.86 \\ 1.26 & 2.00 \\ 1.30 & 2.00 \\ 1.28 & 1.96 \end{bmatrix}$$

蠓Af

$$\begin{bmatrix} 1.24 & 1.72 \\ 1.36 & 1.74 \\ 1.38 & 1.64 \\ 1.38 & 1.82 \\ 1.38 & 1.90 \\ 1.40 & 1.70 \\ 1.48 & 1.82 \\ 1.54 & 2.08 \\ 1.56 & 1.78 \end{bmatrix}$$

3.应用Fisher进行判别分析

(1) 求样本均值向量

$$\bar{x}_1^{(1)} = \frac{1}{6}(1.14 + 1.18 + ... + 1.28) = 1.227$$

$$\bar{x}_1^{(2)} = \frac{1}{6}(1.78 + 1.96 + ... + 1.96) = 1.927$$

$$\bar{x}_2^{(1)} = \frac{1}{9}(1.24 + 1.36 + ... + 1.56) = 1.413$$

$$\bar{x}_2^{(2)} = \frac{1}{9}(1.72 + 1.74 + ... + 1.78) = 1.800$$

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} \bar{x}_1^{(1)} \\ \bar{x}_1^{(2)} \end{bmatrix} = \begin{bmatrix} 1.227 \\ 1.927 \end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} \bar{x}_2^{(1)} \\ \bar{x}_2^{(2)} \end{bmatrix} = \begin{bmatrix} 1.413 \\ 1.800 \end{bmatrix}$$

(2)求两类样本类内离散度矩阵Si

$$S_1=\begin{pmatrix}1.14-1.227\\1.78-1.927\end{pmatrix}\begin{pmatrix}1.14-1.227 & 1.78-1.927\end{pmatrix}+\dots$$

$$+\begin{pmatrix}1.28-1.227\\1.96-1.927\end{pmatrix}\begin{pmatrix}1.28-1.227 & 1.96-1.927\end{pmatrix}$$

$$=\begin{pmatrix}0.0197 & 0.0225\\0.0225 & 0.0389\end{pmatrix}$$

$$S_2=\begin{pmatrix}1.24-1.413\\1.72-1.804\end{pmatrix}\begin{pmatrix}1.24-1.413 & 1.72-1.804\end{pmatrix}+\dots$$

$$+\begin{pmatrix}1.56-1.413\\1.78-1.804\end{pmatrix}\begin{pmatrix}1.56-1.413 & 1.78-1.804\end{pmatrix}$$

$$=\begin{pmatrix}0.0784 & 0.0536\\0.0536 & 0.1352\end{pmatrix}$$

(3)求总的类间散度矩阵Sw

$$S_w = S_1 + S_2 = \begin{pmatrix} 0.0981 & 0.0761 \\ 0.0761 & 0.1741 \end{pmatrix}$$

(4)求向量W*

$$S_w^{-1} = \begin{pmatrix} 15.4209 & -6.7422 \\ -6.7422 & 8.6905 \end{pmatrix}$$

$$W^* = S_w^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \begin{pmatrix} -3.7326 & 2.3593 \end{pmatrix}^T$$

(5) 求两类已知样本X1和X2在W*上的投影点yi

$$y_1 = W^{*T} X_1, y_2 = W^{*T} X_2$$

$$y1 = [-0.0555 \quad 0.2199 \quad -0.0907 \quad 0.0156 \quad -0.1337 \quad -0.1534]^T$$

$$y2 = [-0.5703 \quad -0.9711 \quad -1.2816 \quad -0.8570 \quad -0.6682 \quad -1.2147 \quad -1.2302 \quad -0.8407 \quad -1.6232]^T$$

(6)求各类样本的均值 $\overline{Y}_i$

$$\overline{Y}_i = \frac{1}{N_i} \sum_{Y \in \omega_i} Y$$

$\overline{y}_1$=-0.0330

$\overline{y}_2$=-1.0286

(7)选取阈值W0

本课件介绍了三种取法，这里取：

$$W_0 = \frac{\overline{Y}_1 + \overline{Y}_2}{2} = \text{-}0.5308$$

(8)对待测的标本X，计算它在W*上的投影点y

公式：y=$\mathbf{W}^{*T}\mathbf{X}$

$$X=\begin{pmatrix}1.24 & 1.80\\ 1.28 & 1.84\\ 1.40 & 2.04\end{pmatrix}$$

$$\Rightarrow$$

$$Y=\left(-0.3816,\ -0.4365,\ -0.4126\right)^T$$

(9)根据判别规则对标本X进行分类

$Y=W^{*T}X>W0$，则$X\in\omega1$；

$Y=W^{*T}X<W0$，则$X\in\omega2$ 。

根据(7)，阈值$W_0 = -0.5308$

而：

$$Y = \left(-0.3816,\ -0.4365,\ -0.4126\right)^T$$

可以看出，$y_i > W_0$

因此，这三个标本都属于Apf蠓(ω_1类)

- 本章小结：判断分析是利用原有的模式分类信息，得到判别函数(判别函数是这种分类的函数关系式，可以是与分类信息相关的若干个特征或指标的线性关系式)，然后利用该函数去判断未知的模式或样本属于哪一类。因此，这是一个学习和预测的过程。

七.线性判别分析编程举例

1.距离判别法编程

有欧式距离法、马氏距离法等。但欧氏距离方法较为粗糙。MATLAB中提供了求马氏距离的函数。

设有m个已知模式或样本，每个模式有n个特征或指标，可知第i个特征共测得m个数据(一般m≥n)，用列向表示如下：

$$\mathbf{x_i} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{bmatrix}$$

于是，我们得到m×n阶的数据矩阵X：

$$X = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n)$$

X矩阵中每一行是一个模式(样本)数据，m×n阶矩阵X的n×n阶协方差就是COV(X)。

(1)求n维向量$\mathbf{r}=(r_1,r_2,\ldots,r_n)$到m×n阶矩阵$X$的马氏距离。

定义$\mathbf{r}-\bar{X}=(r_1-\bar{\mathbf{x}}_1,r_2-\bar{\mathbf{x}}_2,r_3-\bar{\mathbf{x}}_3,\ldots,r_n-\bar{\mathbf{x}}_n)$，则$n$维向量

$\mathbf{r}=(r_1,r_2,\ldots,r_n)$到m×n阶矩阵的马氏距离计算公式如下：

$$\mathrm{mahal}(\mathbf{r},X)=(\mathbf{r}-\bar{X})\,\mathrm{cov}(X)^{-1}(\mathbf{r}-\bar{X})^{\mathrm{T}}$$

其中：$\bar{\mathbf{x}}_i$表示第i个特征$\mathbf{x}_i$的算术平均值。

说明：在MATLAB中，mahal马氏距离函数中省略了开平方根符号。

(2)$m \times n$阶矩阵X到自身的马氏距离，相当于矩阵X的每一行(即一个模式)到X的马氏距离，其计算函数为mahal(X,X)。进行这种计算可以求出矩阵X的异常值。

例：现获取了8×3数据矩阵(8个样本数据，每个样本有3个特征)，试用马氏距离法检查出异常的样本。

$$X=\begin{bmatrix} 0.052 & 0.084 & 0.021 \\ 0.037 & 0.0071 & 0.022 \\ 0.041 & 0.055 & 0.11 \\ 0.11 & 0.021 & 0.0073 \\ 0.030 & 0.112 & 0.072 \\ 0.16 & 0.056 & 0.021 \\ 0.074 & 0.083 & 0.105 \\ 0.19 & 0.02 & 1 \end{bmatrix}$$

在MATLAB命令行窗口中输入矩阵X,并用马氏距离函数计算，或直接生成.M程序文件，然后在MATLAB中运行。程序如下：

```
%Filename:lecture3_1.m
%用马氏距离法判别获取的样本是否正常
X= [0.052,0.084,0.021;0.037,0.0071,0.022;
   0.041,0.055,0.11;0.11,0.021,0.0073;
   0.030,0.112,0.072;0.16,0.056,0.021;
   0.074,0.083,0.105;0.19,0.02,1]
d=mahal(X,X)
```

执行结果:

d=

0.6998
3.9050
0.8310
1.8819
2.7672
4.2424
0.5991
6.0735

假设我们设定阈值$d_{Threshold}=2$，凡是$d<=d_{Threshold}$判别为正常样本，否则为异常样本。则从上面结果数据可以看出，矩阵X中第二行、第五行、第六行、第八行数据异常(四个样本异常)。

2.Fisher判断分析编程

根据前面介绍的Fisher算法实现步骤，结合本课件介绍的例题，MATLAB程序清单如下：

```
%Filename：lecture3_2.m
%Fisher分类器程序
%X1,X2:两类样本矩阵
X1=[1.14 1.78
    1.18 1.96
    1.20 1.86
    1.26 2.00
    1.30 2.00
    1.28 1.96];
```

```
X2=[1.24 1.72
    1.36 1.74
    1.38 1.64
    1.38 1.82
    1.38 1.90
    1.40 1.70
    1.48 1.82
    1.54 2.08
    1.56 1.78];
N1=6;
N2=9; %N1：Apf类样本数，N2=Af类样本数
```

```
clc   %清除命令显示窗口
%两类样本均值向量
m1=(mean(X1))'  %第1类样本的均值向量
m2=(mean(X2))'  %第2类样本的均值向量
%求两类样本类内离散度矩阵
s1=cov(X1)*(N1-1)
s2=cov(X2)*(N2-1)
sw=s1+s2   %求总类间离散度矩阵
w=inv(sw)*(m1-m2)   %求W*
```

%求已知类别在W*上的投影

```
y1=w'*X1'
y2=w'*X2'
```

%求各类别样品在投影空间上的均值

```
mean1=mean(y1')
mean2=mean(y2')
```

%求阈值W0

```
W0=(mean1+mean2)/2
```

```
%待识别的三个标本
X=[1.24 1.80
   1.28 1.84
   1.40 2.04]
%Fisher判别函数
Y=w'*X'
%分类决策
for i=1:3
  if Y[i]>W0
    class=1
  else
    class=2
  end
end
```

根据本讲的Fisher算法实现流程和例题，同学们不难读懂本程序。

3.利用MATLAB classify分类函数做判断分析

在MATLAB软件包中，将已经分类的m个样本(长度为n,n也就是样本的特征数目)作为一个行向量，得到一个矩阵training,每行都属于一个分类类别，分类类别构成一个整数列向量T(共有m行)，待分类的k个样本(长度为n)作为行向量，得到一个矩阵sample,然后利用classify函数进行线性判别分析。其格式为：

[[class,err]=]classify(sample,training,target[,type]);

其中sample与training必须具有相同的列数n；target与training必须具有相同的行数m，target是一整数向量。classify的功能是对sample的每一行进行分类判别，然后划分到training指定的模式类中。

[]为可选项。其中class返回分类表；err返回误差比例信息；sample是待分类的模式数据矩阵；training是已知的分类模式矩阵；target是分类列向量。type有三种选择：

type=linear(缺省): 表示做线性判别分析;

type=quadratic: 表示做二次非线性判别分析;

type=mahalanobis: 表示采用马氏距离做判别分析。

例：植物分类的判别分析，数据如下：

表1 ω1类植物

植物	特征1	特征2	特征3	特征4	特征5	特征6
A1	135.20	36.40	10.47	44.16	36.40	3.94
A2	145.68	32.83	17.79	27.29	39.09	3.47
A3	159.37	33.38	18.37	11.81	25.29	5.22
A4	144.98	29.12	11.67	42.60	27.30	5.74
A5	169.92	32.75	12.72	47.12	34.35	5.00
A6	115.84	30.76	12.20	33.61	33.77	3.85

表2 ω2类植物

植物	特征1	特征2	特征3	特征4	特征5	特征6
B1	116.22	29.57	13.24	13.76	21.75	6.04
B2	153.11	23.09	15.62	23.54	18.18	6.39
B3	144.92	21.26	16.92	19.52	21.75	6.73
B4	140.54	21.59	17.64	19.19	15.97	4.94
B5	140.65	28.26	12.35	18.53	20.95	6.23
B6	164.02	24.74	13.63	22.20	18.06	6.04
B7	139.08	18.47	14.68	13.41	20.66	3.85
B8	137.80	20.74	11.07	17.74	16.49	4.39
B9	121.67	21.53	12.58	14.49	12.18	4.57
B10	123.24	38.00	13.72	4.64	17.77	5.75

表3 ω3类植物

植物	特征1	特征2	特征3	特征4	特征5	特征6
C1	95.21	22.83	9.30	22.44	22.81	2.80
C2	104.78	25.11	6.46	9.89	18.17	3.25
C3	128.41	27.63	8.94	12.58	23.99	3.27
C4	101.18	23.26	8.46	20.20	20.50	4.30
C5	124.27	19.81	8.89	14.22	15.53	3.03
C6	106.02	20.56	10.94	10.11	18.00	3.29
C7	95.65	16.82	5.70	6.03	12.36	4.49
C8	107.12	16.45	8.98	5.40	8.78	5.93
C9	113.74	24.11	6.46	9.61	22.92	2.53

表4 待识别的植物

植物	特征1	特征2	特征3	特征4	特征5	特征6
D1	190.33	43.77	9.73	60.54	49.01	9.04
D2	221.11	38.64	12.53	115.65	50.82	5.89
D3	182.55	20.52	18.32	42.40	36.97	11.68

试将**D1-D3**这三棵植物进行分类。

方法如下：

(1)将25个以分类植物的6个特征数据组成一个25×6的矩阵，取名为training；

(2)将6个第1类ω1(对应6个1)，10个第2类ω2(对应10个2)，9个第3类ω3(对应9个3),组成一个25×1的整数列向量，取名为target；

(3)将待识别的三个模式样本的6个特征数据组成一个3×6的矩阵，取名为sample；

(4)MATLAB程序清单

```
%Filename:lecture3_3.m
%利用MATLAB classify函数进行植物分类
A=[135.20  145.68  159.37  144.98  169.92  115.84
   36.40   32.83   33.38   29.12   32.75   30.76
   10.47   17.79   18.37   11.67   12.72   12.20
   44.16   27.29   11.81   42.60   47.12   33.61
   36.40   39.09   25.29   27.30   34.35   33.77
   3.94    3.47    5.22    5.74    5.00    3.85]';
```

```
B=[116.22  153.11  144.92  140.54  140.65  164.02  139.08  137.80  121.67  123.24
      29.57  23.09   21.26   21.59   28.26   24.74   18.47   20.74   21.53   38.00
      13.24  15.62   16.92   17.64   12.35   13.63   14.68   11.07   12.58   13.72
      13.76  23.54   19.52   19.19   18.53   22.20   13.41   17.74   14.49   4.64
      21.75  18.18   21.75   15.97   20.95   18.06   20.66   16.49   12.18   17.77
        6.04   6.39    6.73    4.94    6.23    6.04    3.85    4.39    4.57    5.75]';
C=[95.21  104.78  128.41  101.18   124.27  106.02  95.65  107.12  113.74
    22.83  25.11   27.63   23.26    19.81   20.56   16.82   16.45   24.11
     9.30   6.46    8.94    8.46     8.89    10.94   5.70     8.98   6.46
    22.44  9.89   12.58   20.20    14.22    10.11   6.03     5.40   9.61
    22.81 18.17   23.99   20.50    15.53    18.00   12.36    8.78   22.92
     2.80   3.25   3.27    4.30     3.03     3.29     4.49    5.93    2.53]';
```

```
training=[A;B;C]   %25x6样本矩阵，用于训练
target=[1;1;1;1;1;1;2;2;2;2;2;2;2;2;2;2;3;3;3;3;3;3;3;3;3];
sample=[190.33 221.11 182.55
            43.77  38.64  20.52
             9.73  12.53  18.32
            60.54 115.65  42.40
            49.01  50.82  36.97
             9.04   5.89  11.68]';
class=classify(sample,training,target)
```

(5)程序运行结果

class =

1

1

1

从结果得知，三个待识别的植物都属于第一类植物。

课后上机作业：两类问题，已知四个训练样本

$\omega 1=\{(0,0)^T,(0,1)^T\}$

$\omega 2=\{(1,0)^T,(1,1)^T\}$

使用感知器固定增量法求判别函数。

设$w1=(1,1,1,1)^T$　　$\rho_k=1$

试编写程序上机运行(使用MATLAB、 VC++、DELPHI三种语言任意一种编写均可)，写出判别函数，并给出相关图表。