第二章　统计模式识别中的几何分类法

中科大 自动化系　郑志刚

2018.9

- § 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、线性分类器的设计
- § 2-5、广义线性判别函数
- § 2-6、非线性判别函数
- § 2-7、非线形分类器的设计
- § 2-8、最优分类超平面

§ 2-1 判别函数

- 在第1章中，说明了模式的概念，它是取自客观世界中的一次抽样试验样本的被测量值的综合。如果试验对象和测量条件相同，所有测量值具有重复性，即在多次测量中，它们的结果不变，获得的模式称为**确定性的模式**。否则，测量值是随机的，这样的模式称为**随机性的模式**，简称随机模式。下面介绍确定性模式的分类方法。

§ 2-1 判别函数（续）

- 假设对一模式X已抽取n个特征，表示为：

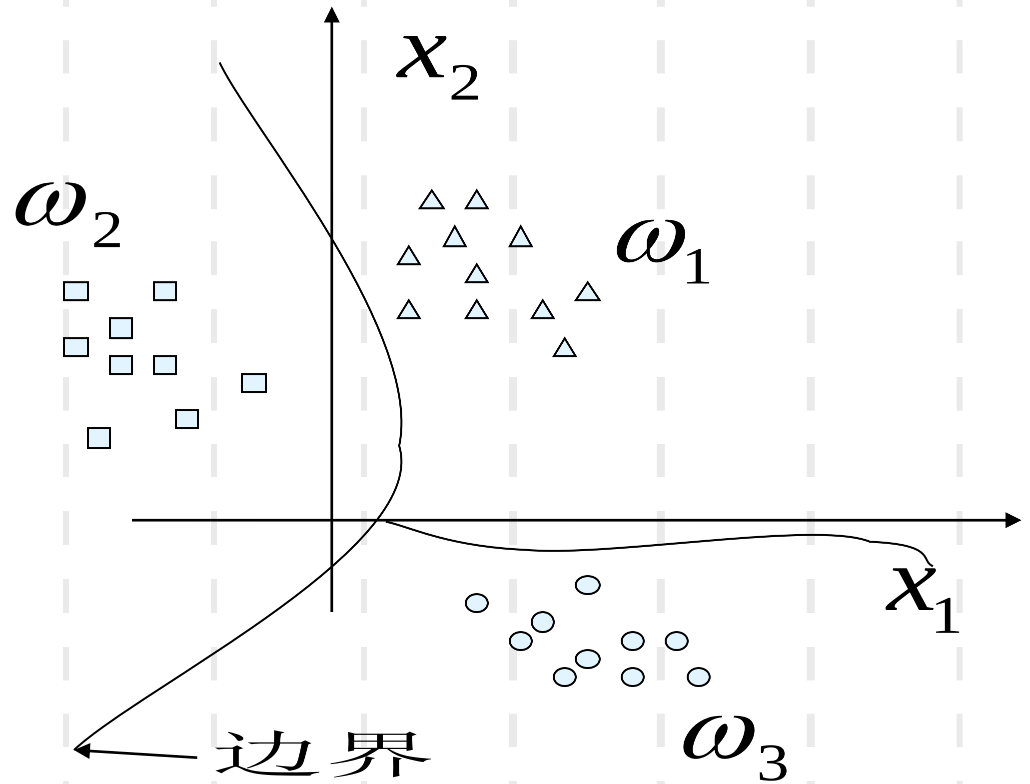
$$X = (x_1, x_2, x_3, \ldots, x_n)^T$$

X是n维空间的一个向量

- 模式识别问题就是根据模式X的n个特征来判别模式属于$\omega_1, \omega_2, \ldots, \omega_m$ 类中的那一类。

§ 2.1 判别函数（续）

- 例如下图：三类的分类问题，它们的边界线就是一个判别函数



§ 2.1 判别函数（续）

- 判别函数包含两类：
  - 一类 是线性判别函数：
    - 线性判别函数
    - 广义线性判别函数

      （所谓广义线性判别函数就是把非线性判别函数映射到另外一个空间变成线性判别函数）
    - 分段线性判别函数
  - 另一类是非线性判别函数

§ 2-2 线性判别函数

- 我们现在对两类问题和多类问题分别进行讨论。
- (一)两类问题　即:
- $$\omega_i = (\omega_1, \omega_2)^T, M = 2$$
  - 1. 二维情况：取两个特征向量

$$X = (x_1, x_2)^T, n = 2$$

- 这种情况下 判别函数:

$$g(x) = w_1 x_1 + w_2 x_2 + w_3$$

w为参数，x_1, x_2为坐标向量

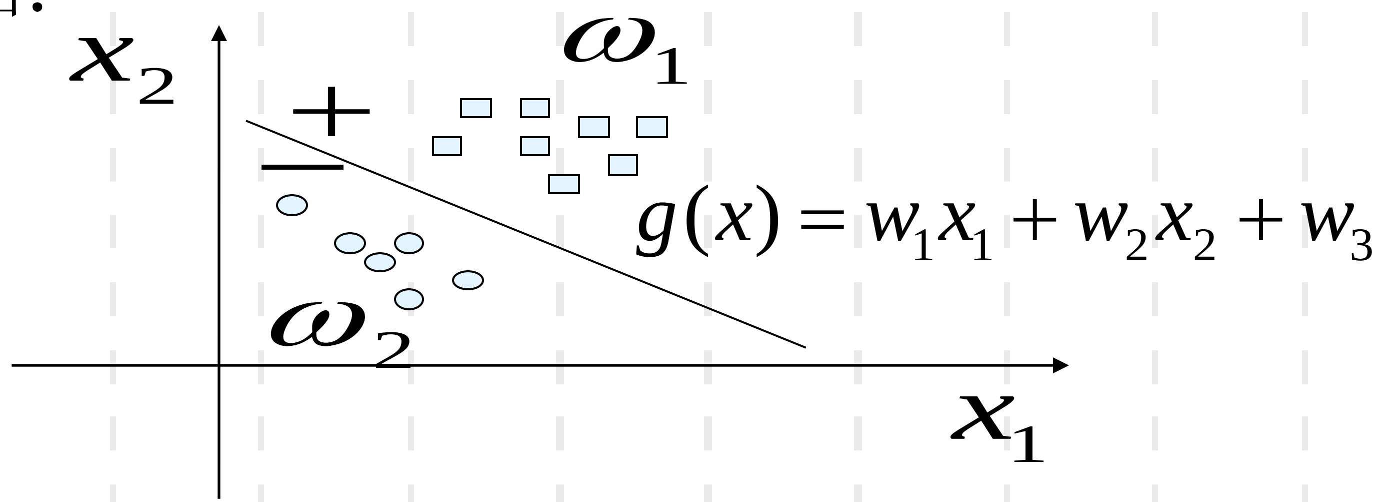
1. 二维情况

- 在两类别情况，判别函数 $g(x)$ 具有以下性质：

$$g_i(x)=\begin{cases} >0, X\in\omega_1 \\ <0, X\in\omega_2 \end{cases}$$

$$g(x)=0, X\text{不定}$$

- 这是二维情况下判别由判别边界分类.
- 情况如图：



2. n维情况

- 现抽取n个特征为：

$$X=(x_1,x_2,x_3,\ldots x_n)^T$$

- 判别函数：$g(x)=w_1x_1+w_2x_2+\ldots\ldots+w_nx_n+w_{n+1}$

$$=W_0X+w_{n+1}$$

$W_0=(w_1,w_2,\ldots,w_n)^T$ 为权向量，

$X=(x_1,x_2,\ldots,x_n)^T$ 为模式向量。

- 另外一种表示方法： $g(x)=W^TX$

$W=(w_1,w_2,\ldots,w_n,w_{n+1})^T$ 为增值权向量，

$X=(x_1,x_2,\ldots,x_n,1)^T$ 为增值模式向量。

2. n维情况

- 模式分类：

$$g(x) = W^T X \begin{cases} > 0, x \in \omega_1 \\ < 0, x \in \omega_2 \end{cases}$$

- 当 $g_1(x) = W^T X = 0$ 为判别边界。当n=2时，二维情况的判别边界为一直线。当n=3时，判别边界为一平面，n>3时，则判别边界为一超平面。

(二) 多类问题

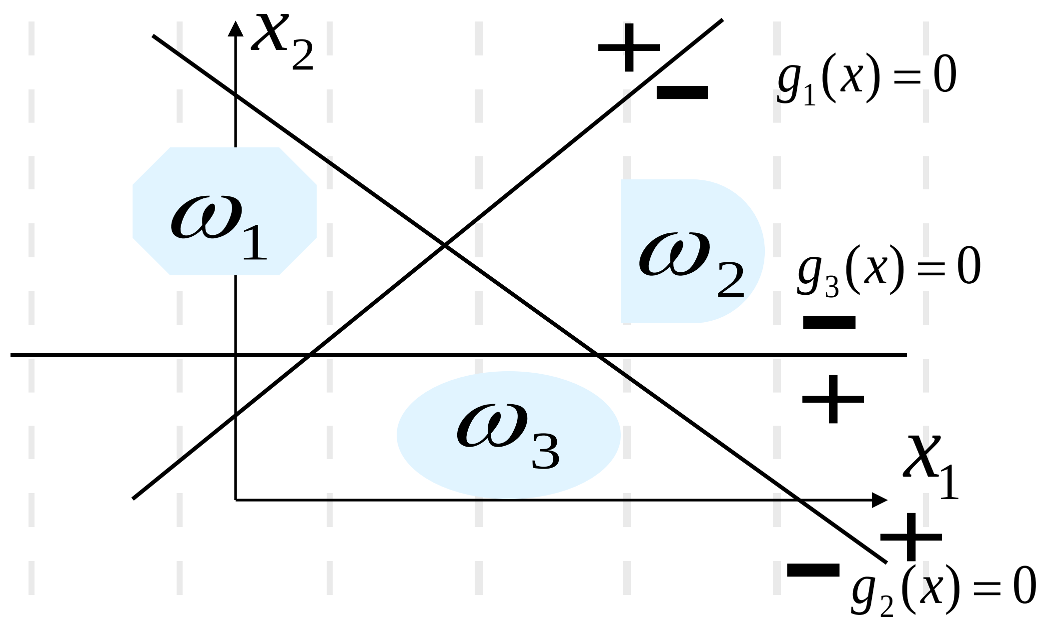
- 对于多类问题，模式有 $\omega_1, \omega_2, \ldots, \omega_m$ 个类别。可分三种情况：
- 1、**第一种情况：每一模式类与其它模式类间可用单个判别平面把一个类分开。**这种情况，M类可有M个判别函数，且具有以下性质：

$$g_i(x) = W_i^T X \begin{cases} > 0, X \in \omega_i \\ < 0, \text{其它}, i = 1,2,\ldots, M\text{。} \end{cases}$$

式中 $W_i = (w_{i1}, w_{i2}, \ldots, w_{in}, w_{in+1},)^T$ 为第 i 个判别函数的权向量。

1、第一种情况

- 右图所示，每一类别可用单个判别边界与其它类别相分开 。
- 如果一模式X属于ω_1，则由图可清楚看出:这时$g_1(x)>0$而$g_2(x)<0$， $g_3(x)<0$ 。 ω_1 类与其它类之间的边界由$g_1(x)=0$确定.



1、第一种情况（续）

- 例：已知三类$\omega_1,\omega_2,\omega_3$的判别函数分别为：

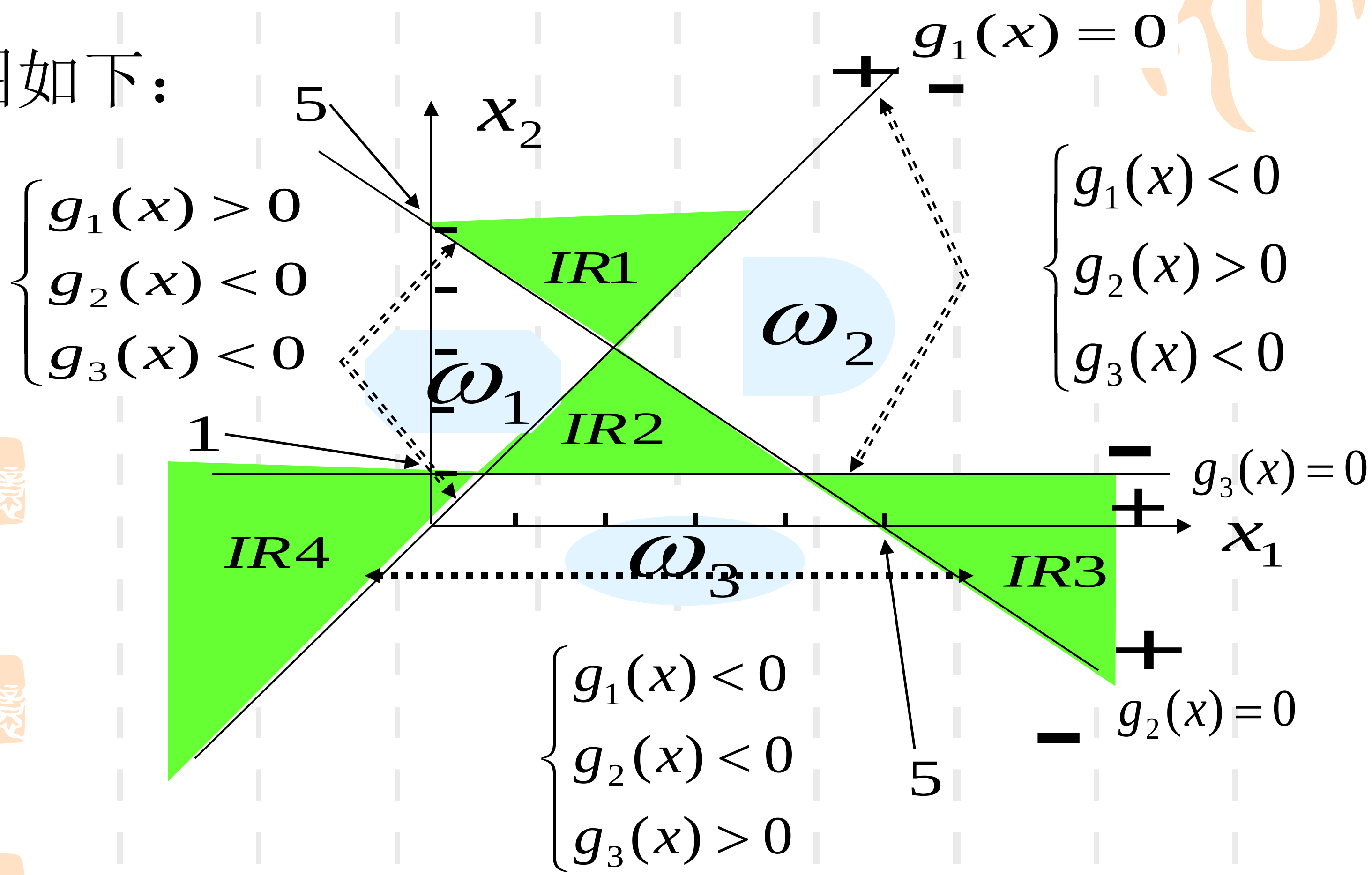
$$\begin{cases} g_1(x) = -x_1 + x_2 \\ g_2(x) = x_1 + x_2 - 5 \\ g_3(x) = -x_2 + 1 \end{cases}$$

- 因此三个判别边界为：

$$\begin{cases} g_1(x) = -x_1 + x_2 = 0 \\ g_2(x) = x_1 + x_2 - 5 = 0 \\ g_3(x) = -x_2 + 1 = 0 \end{cases}$$

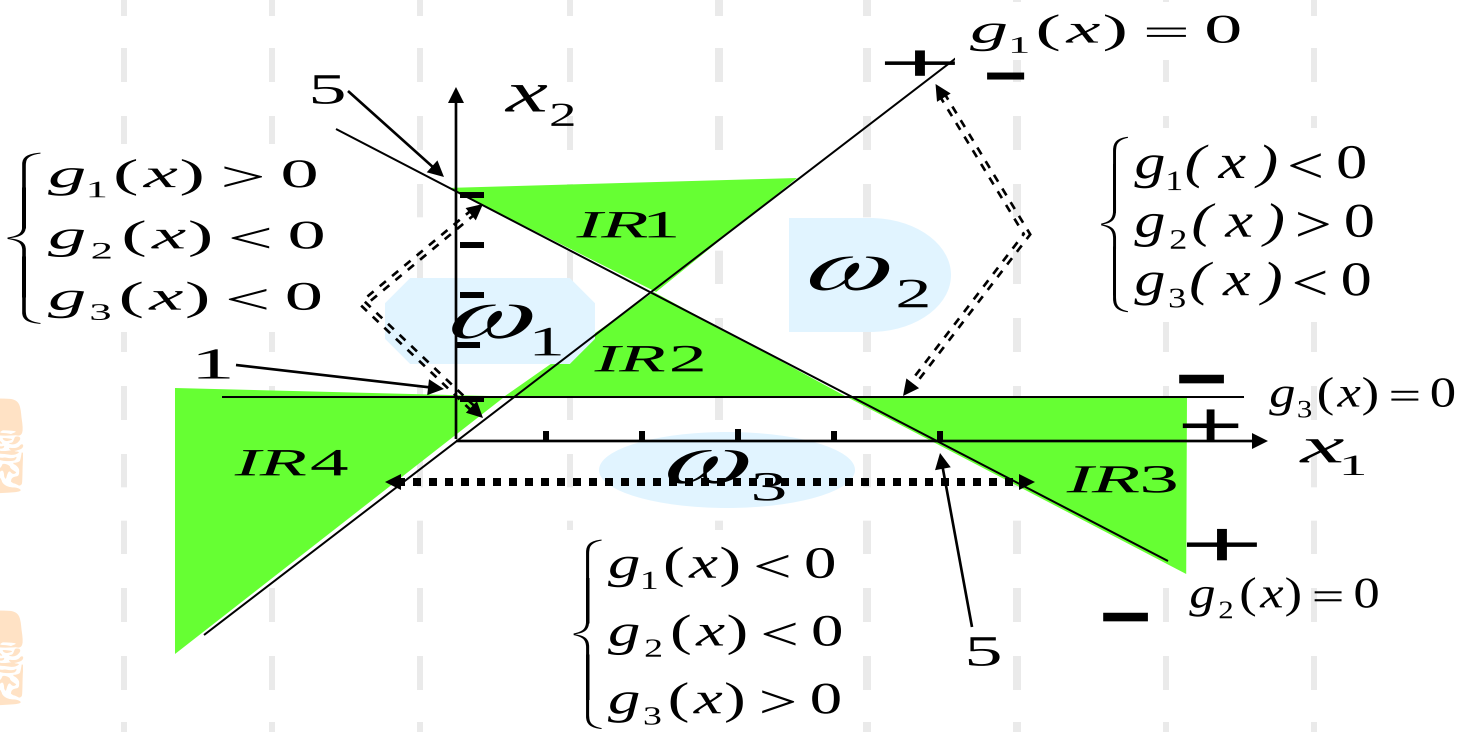
1、第一种情况（续）

❖作图如下：



1、第一种情况（续）

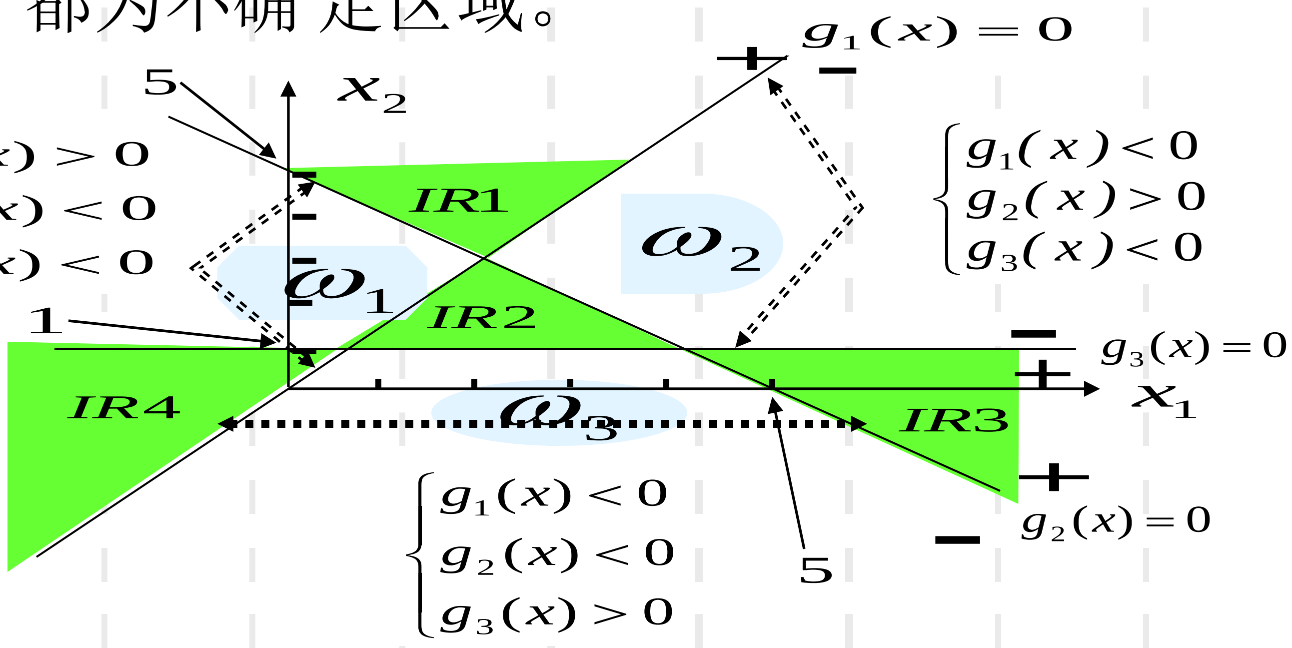
- 对于任一模式*X*如果它的 $g_1(x)>0$， $g_2(x)<0$， $g_3(x)<0$
- 则该模式属于ω_1类。相应ω_1类的区域由直线$-x_2+1=0$的正边、直线$-x_1+x_2-5=0$ 和直线$-x_1+x_2=0$的负边来确定。



1、第一种情况（续）

- 必须指出，如果某个***X***使二个以上的判别函数 $g_i(x)>0$ 。则此模式***X***就无法作出确切的判决。如图中 **IR1,IR3，IR4**区域。

- 另一种情况是**IR2**区域，判别函数都为负值。**IR1**，**IR2**，**IR3**，**IR4**。都为不确 定区域。



1、第一种情况（续）

- 问当$x=(x_1,x_2)^T=(6,5)^T$时属于那一类

  代入判别函数方程组：

$$\begin{cases} g_1(x) = -x_1 + x_2 \\ g_2(x) = x_1 + x_2 - 5 \\ g_3(x) = -x_2 + 1 \end{cases}$$

得：

$$g_1(x) = -1, g_2(x) = 6, g_3(x) = -4.$$

- 结论：　$g_1(x)<0$，　$g_2(x)>0$，　$g_3(x)<0$所以它属于ω_2类

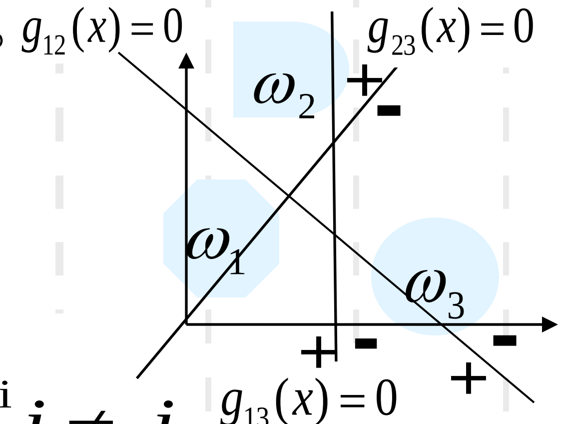
➢2、第二种情况：

➢每两两模式类间可分别用判别平面分开

❖这样 有 $M(M-1)/2$个判别平面。

❖对于两类问题，M=2，则有一个判别平面。

❖同理，三类问题则有三个判别平面。



❖ 判别函数： $g_{ij}(x)=W_{ij}^{T}X$

❖ 判别边界： $g_{ij}(x)=o$

❖ 判别条件：$g_{ij}(x)\begin{cases} >0 \rightarrow 当x\in\omega_i \\ <0 \rightarrow 当x\in\omega_j \end{cases} \quad i\neq j$

2、第二种情况（续）

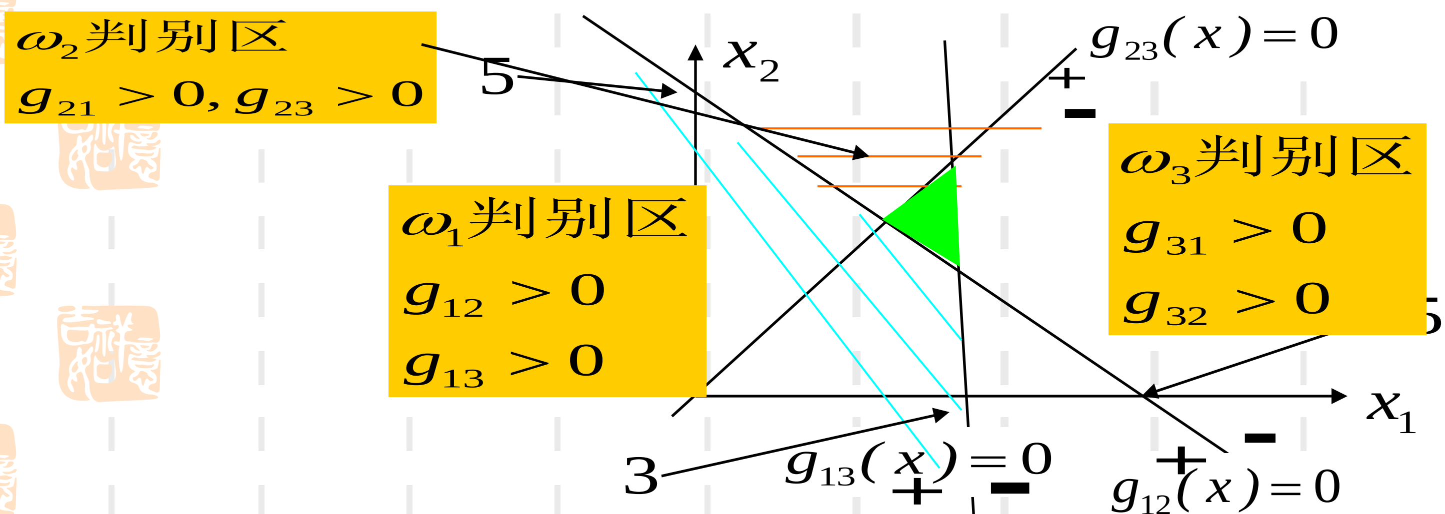
- 判别函数性质： $g_{ij}(x) = -g_{ji}(x)$
- 假设判别函数为：

$$\begin{cases} g_{12}(x) = -x_1 - x_2 + 5 \\ g_{13}(x) = -x_1 + 3 \\ g_{23}(x) = -x_1 + x_2 \end{cases}$$

- 判别边界为：

$$\begin{cases} g_{12}(x) = -x_1 - x_2 + 5 = 0 \\ g_{13}(x) = -x_1 + 3 = 0 \\ g_{23}(x) = -x_1 + x_2 = 0 \end{cases}$$

- 用方程式作图：



2、第二种情况（续）

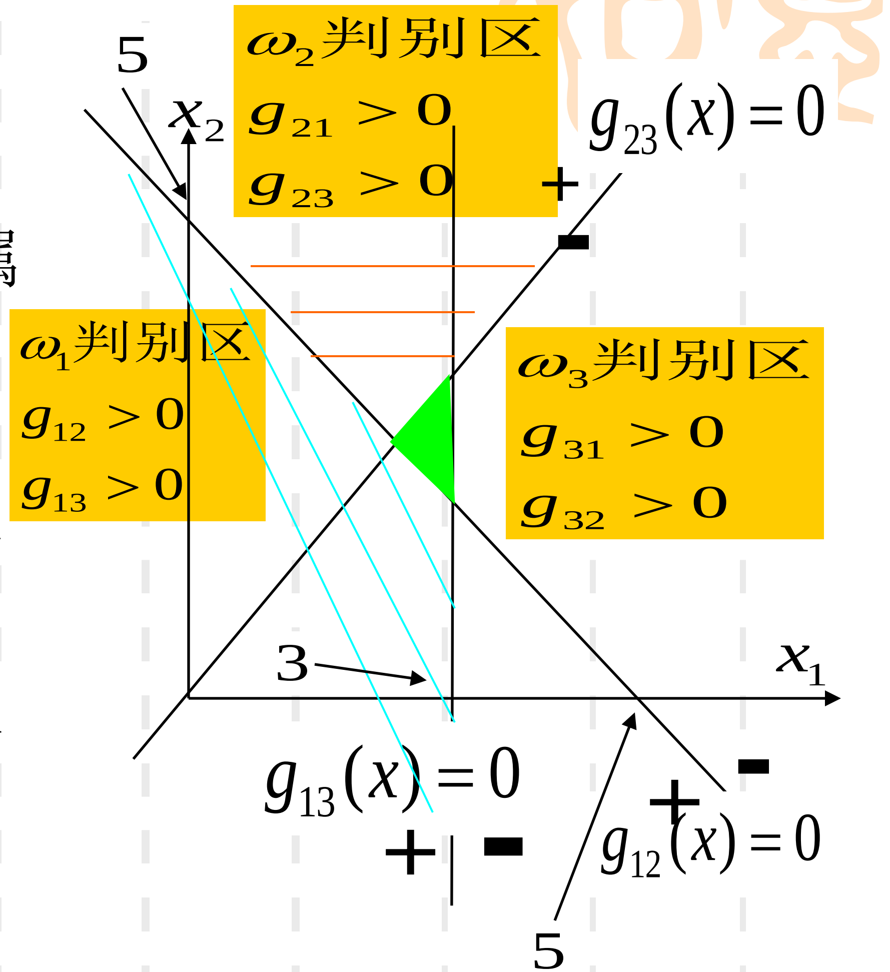
- 结论：判别区间增大，不确定区间减小，比第一种情况小的多.
- 问:未知模式$X=(x_1,x_2)^T=(4,3)^T$属于那一类
- 代入判别函数可得:

$$g_{12}(x)=-2, g_{13}(x)=-1, g_{23}(x)=-1$$

- 把下标对换可得:

$$g_{21}(x)=2, g_{31}(x)=1, g_{32}(x)=1$$

- 因为 $g_{3j}(x)>0$
- 结论：所以X 属于ω_3类



➢3、第三种情况

- 每类都有一个判别函数,存在M个判别函数
- 判别函数：$g_k(x)=W_K X \qquad K=1,2,\ldots,M$
-
- 判别规则：$g_i(x)=W_k^T X\begin{cases}\text{最大，当}x\in\omega_i \\ \text{小，其它}\end{cases}$
- 或 $g_i(x)=\max\limits_{j=1,\ldots k}\{g_j(x)\}$
- 判别边界：$g_i(x)=g_j(x)$ 或$g_i(x)-g_j(x)=0$
- 就是说，要判别模式$\boldsymbol{X}$属于那一类，先把$\boldsymbol{X}$代入**M**个判别函数中，判别函数最大的那个类别就是$\boldsymbol{X}$所属类别。类与 类之间的边界可由 $g_i(x)=g_j(x)$ 或$g_i(x)-g_j(x)=0$来确定。这就是最大值判决规则。

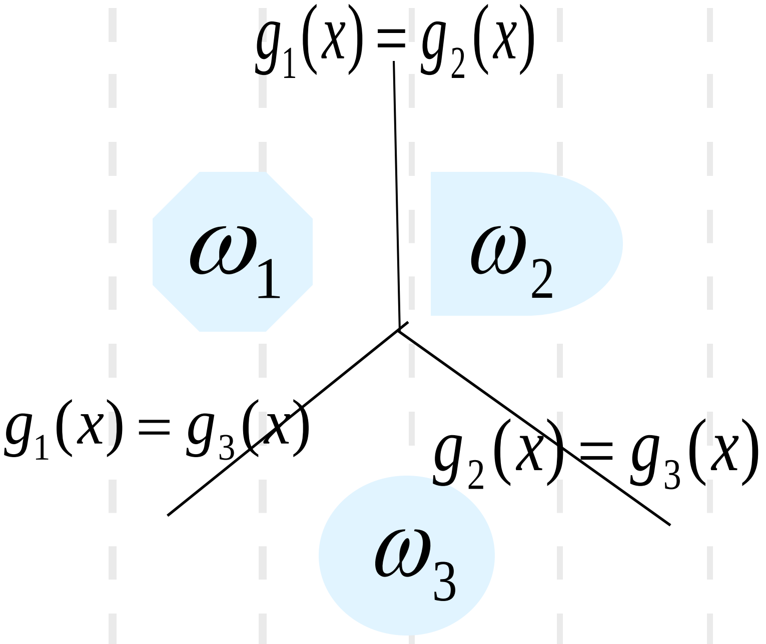
3、第三种情况（续）

- 右图所示是 M=3 的例子。
- 对于 ω$_1$ 类模式，必然满足 $g_1(x) > g_2(x)$ 和 $g_1(x) > g_3(x)$ 。
- 假设判别函数为：

$$\begin{cases} g_1(x) = -x_1 + x_2 \\ g_2(x) = x_1 + x_2 - 1 \\ g_3(x) = -x_2 \end{cases}$$

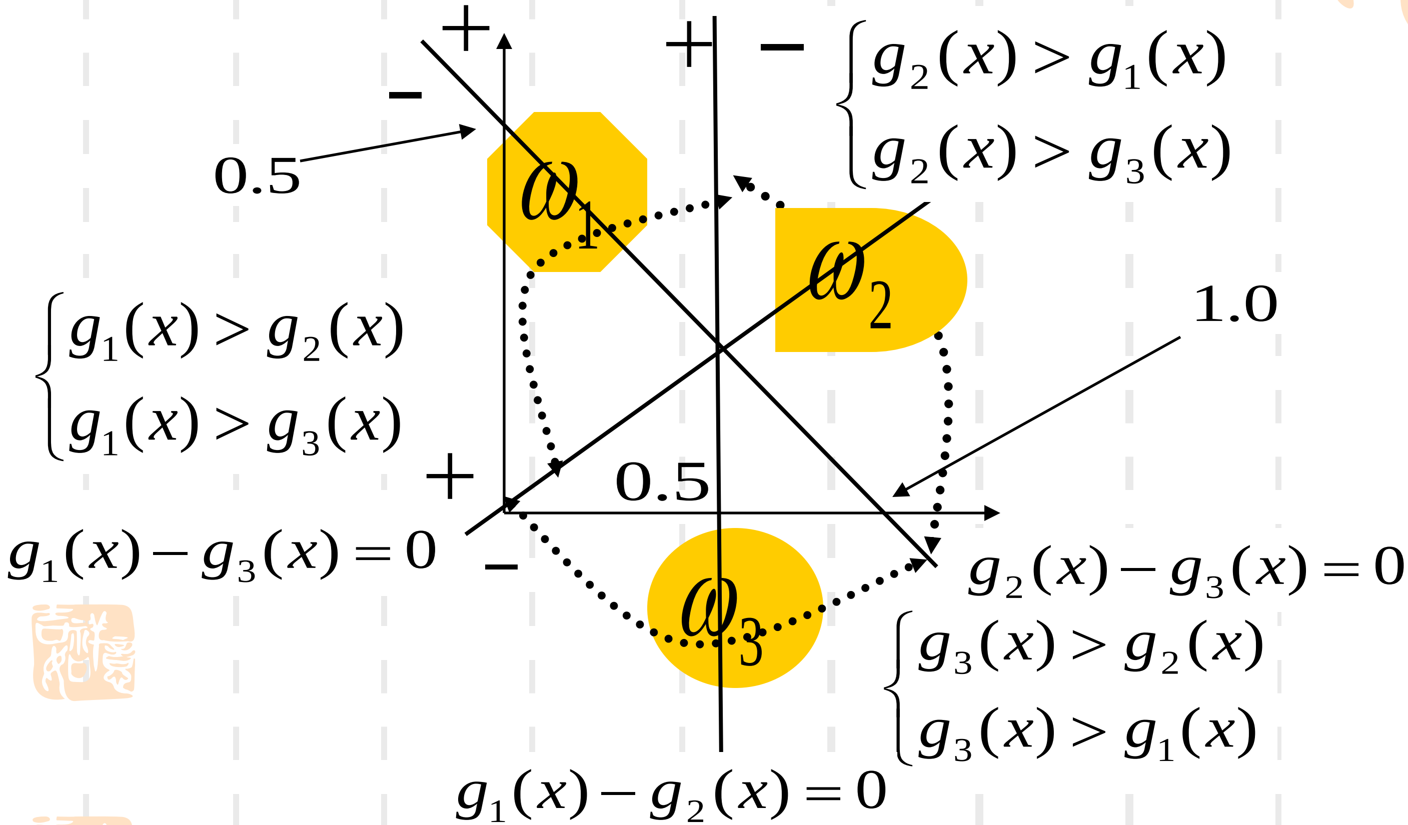
- 则判别边界为：

$$\begin{cases} g_1(x) - g_2(x) = -2x_1 + 1 = 0 \\ g_1(x) - g_3(x) = -x_1 + 2x_2 = 0 \\ g_2(x) - g_3(x) = x_1 + 2x_2 - 1 = 0 \end{cases}$$



3、第三种情况（续）

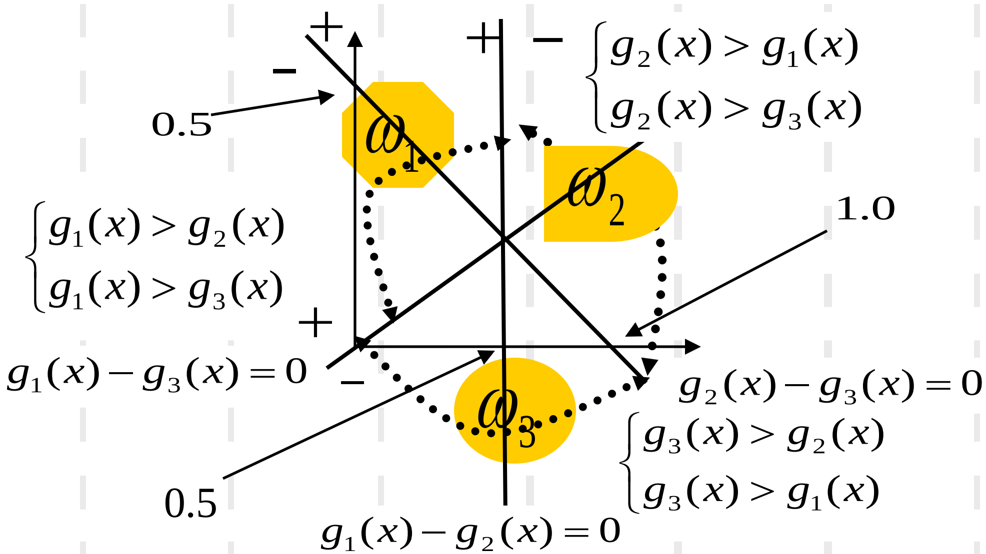
- 用上列方程组作图如下：



- 结论：不确定区间没有了，所以这种是最好情况。

3、第三种情况（续）

- 问假设未知模式$x=(x_1,x_2)^T=(1,1)^T$，则x属于那一类。
- 把它代入判别函数：$g_1(x), g_2(x), g_3(x)$.
- 得判别函数为：$g_1(x)=0, g_2(x)=1, g_3(x)=-1$
- 因为 $g_2(x)>g_3(x), g_2(x)>g_1(x)$
- 所以模式$x=(1,1)^T$属于ω_2类。

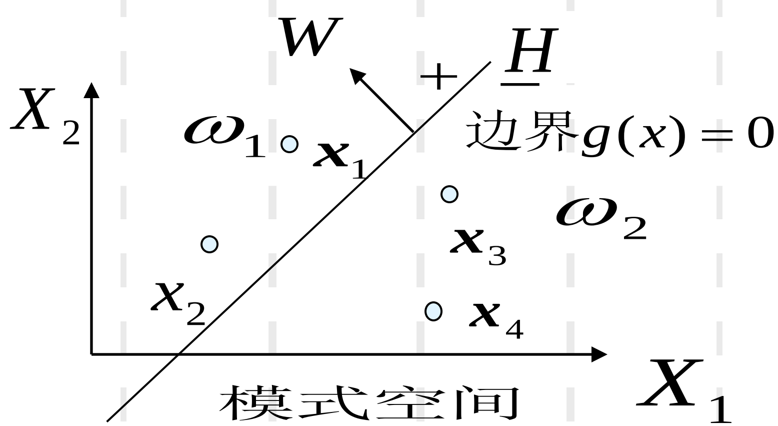


三种情况总结

- 本节把多类问题分成了三种情况进行了讨论，每一种都建立了相应的线性判决函数和有关判决规则。第一种情况，把多类问题转为$w_i/\overline{w_i}$二分法问题；第二种情况，把多类问题转为 w_i/w_j 二分法问题；第三种情况，使用最大值判决规则，对相邻的多种类型也可以转变成w_i/w_j 二分法问题。总之，多类问题的三种情况均可以转变成两类问题。而且，$w_i/\overline{w_i}$ 在i类与其它类之间确定决策面，
- w_i/w_j 在i类与j类之间确定决策面，显然后者比较容易。

§2-3、线性判别函数的性质

- 1、模式空间与加权空间 $g(x)=W_i^T X$
  - 模式空间：由 $X=(x_1,x_2,x_3,\ldots x_n)^T$ 构成的n维欧氏空间。
  - W是此空间的加权向量，它决定模式的分界面H，W与H正交。
  - 加权空间：以 $w_1,w_2,\ldots,w_{n+1}$ 为变量构成的欧氏空间
  - 模式空间与加权空间的几何表示如下图：



1、模式空间与加权空间

- 加权空间的构造：$g(x)=w_1x_1+w_2x_2+w_3$
- 设 $x_1=(x_{11},x_{12})^T$ 是加权空间分界面上的一点，代入上式得：$g(x_1)=w_1x_{11}+w_2x_{12}+w_3=0$,这是加权空间的边界
- 该式表示一个通过加权空间原点的平面，此平面就是加权空间图中的平面①,同样令$g(x_2)$ =$g(x_3)$ =$g(x_4)$=0，分别作出通过加权空间原点的平面②③④图中用阴影表示的部分是各平面的正侧。

设：$x_1,x_2\in\omega_1$

$x_3,x_4\in\omega_2$

$$\because\ g(x)\begin{matrix}>\\ \\<\end{matrix}0\quad\begin{matrix}x\in\omega_1\\ \\x\in\omega_2\end{matrix}$$

$$\left.\begin{aligned}w_1x_{11}+w_2x_{12}+w_3>0\\w_1x_{21}+w_2x_{22}+w_3>0\end{aligned}\right\}\omega_1$$

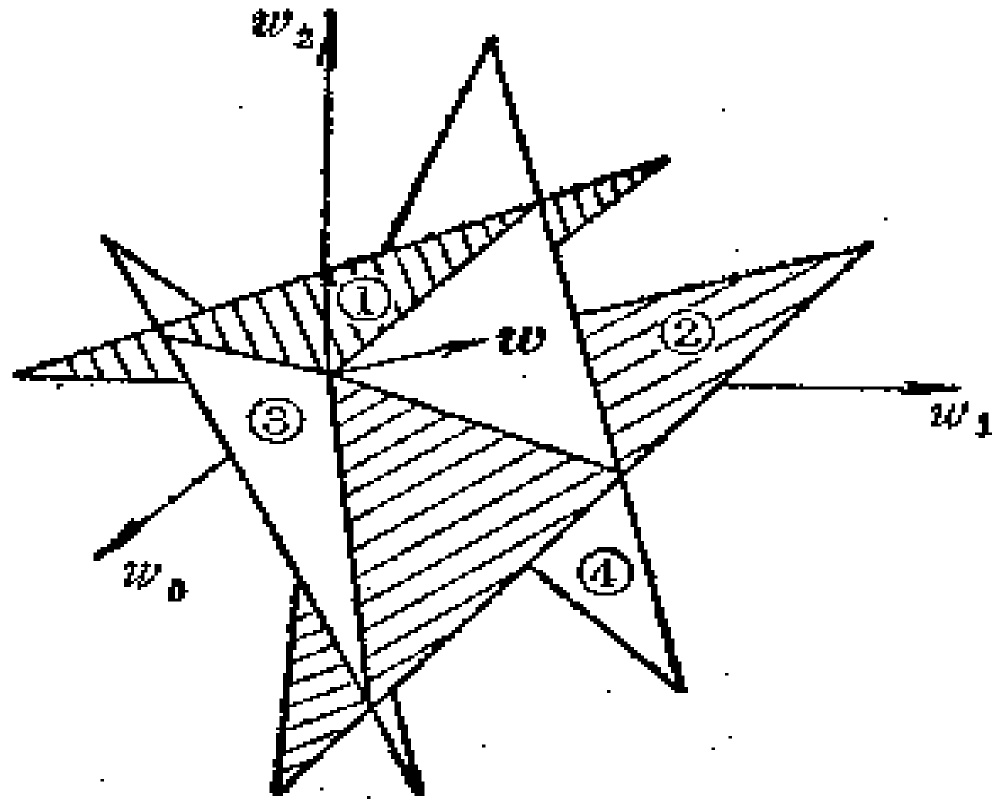
$$\left.\begin{aligned}w_1x_{31}+w_2x_{32}+w_3<0\\w_1x_{41}+w_2x_{42}+w_3<0\end{aligned}\right\}\omega_2$$

最终形成凸多面锥

1、模式空间与加权空间（续）

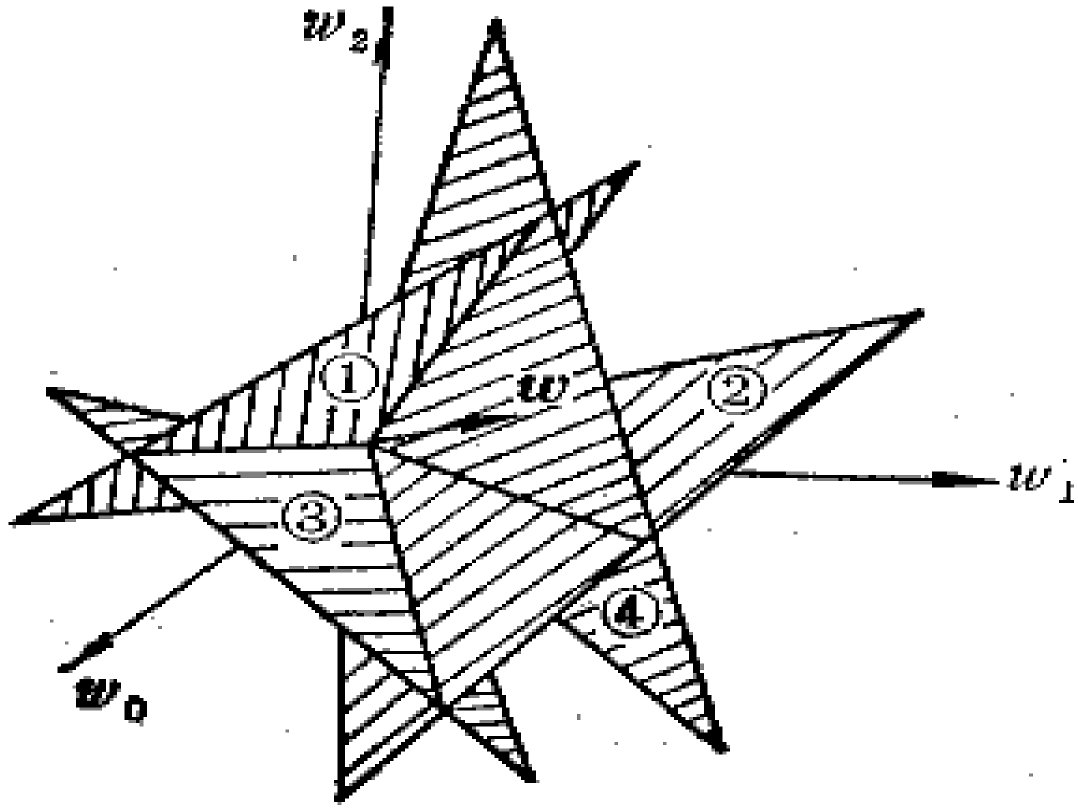
- 这是一个不等式方程组，它的解$W=(w_1,w_2,w_3)^T$ 处于由ω$_1$类所有模式决定的平面的正边和由ω$_2$类所有模式决定的平面的负边，它的解区即为凸多面锥。
- 如图所示：(b)为加权空间，（c）为正规化后的加权空间。
- 由上可以得到结论：加权空间的所有分界面都通过坐标原点。这是加权空间的性质。
- 为了更清楚，下面用二维权空间来表示解向量和解区。

加权空间判别界面



(b)

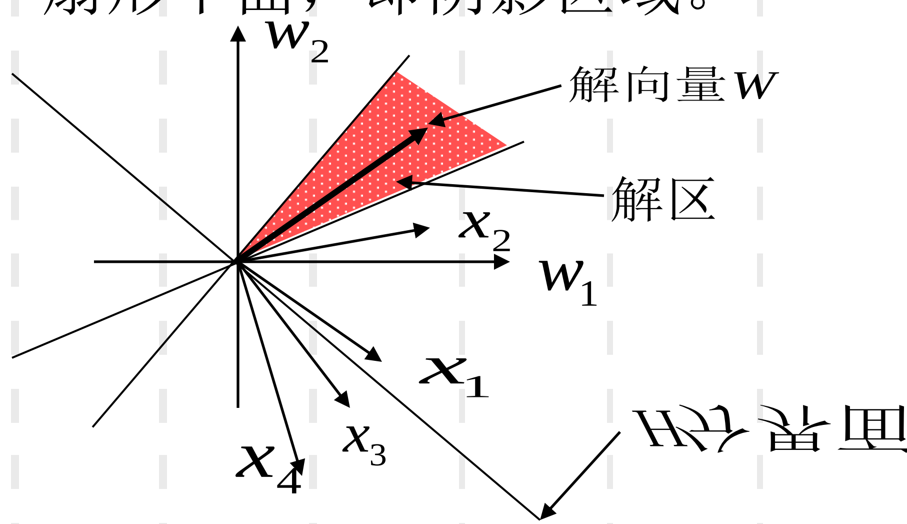
1、模式空间与加权空间（续）



(c)

2、解向量和解区

- 在三维空间里，令$w_3 = 0$ 则为二维权空间。如图：
- 给定一个模式X，就决定一条直线：$g(x) = W^T X = 0$
- 即分界面H，W与H正交，W称为解向量。
- 解向量的变动范围称为解区。
- 因$x_1,x_2 \in \omega_1$，$x_3,x_4 \in \omega_2$**由图可见**x_1,x_3**离的最近，所以分**界面H可以是x_1,x_3之间的任一直线，由垂直于这些直线的W就构成解区，解区为一扇形平面，即阴影区域。



解向量与解区

➢ 2、解向量的解区（续）

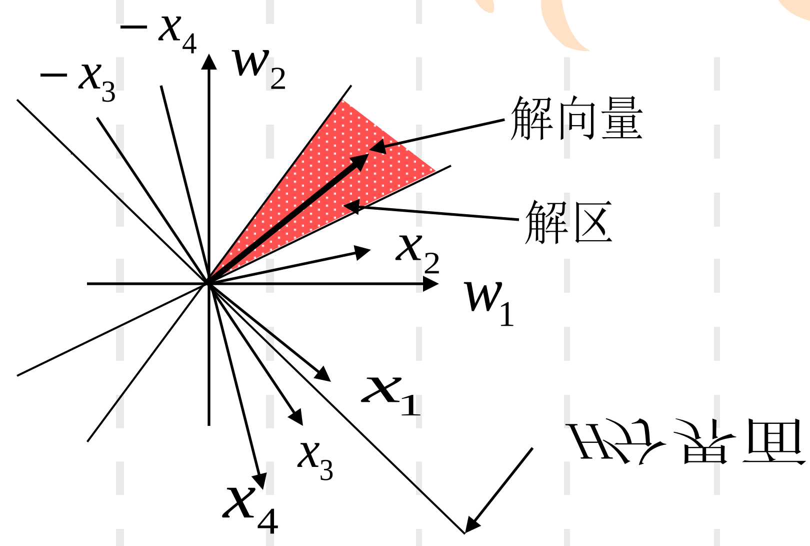
❖ 把不等式方程正规化：

$$
\begin{cases}
w_1 x_{11} + w_2 x_{12} + w_3 > 0 \\
w_1 x_{21} + w_2 x_{22} + w_3 > 0 \\
- w_1 x_{31} - w_2 x_{32} - w_3 > 0 \\
- w_1 x_{41} - w_2 x_{42} - w_3 > 0
\end{cases}
$$

❖ 正规化：

$$g(x) = W_i^T X > 0$$

$$W = (w_1, w_2, \ldots, w_n, w_{n+1},)$$



正规化

2、解向量的解区（续）

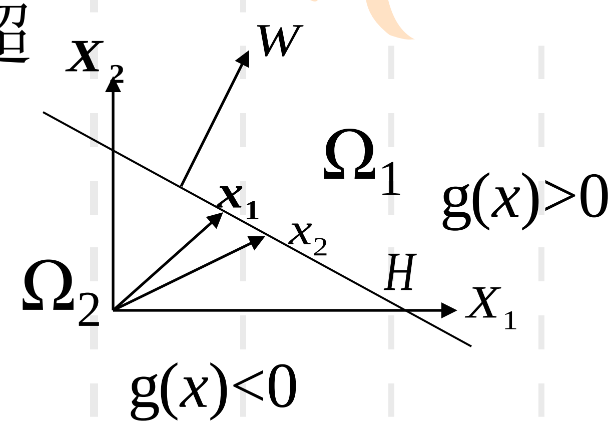
- 解向量特性：

(1) 一般不唯一。落入权空间的一个凸多面锥围成的区域内的每一个点都可作为。该区域称为解区。

(2) 每个训练样本都对解区提供一个限制。样本越多，限制越严，越可靠，越是靠近解区的中心。越可靠，意味着分类错误越小。

3、超平面的几何性质

- $g(x)=W^TX=0$决定一个决策界面，当$g(x)$为线性时，这个决策界面便是一个超平面***H***，并有以下性质：



- 性质①：***W***与***H***正交（如图所示）
  - 假设x_1,x_2是***H***上的两个向量
  - 所以$W^Tx_1+w_{n+1}=W^Tx_2+w_{n+1}=0$

$$\therefore W^T(x_1-x_2)=0,(x_1-x_2)\text{矢量一定在}H\text{上}$$

  - ***W*** 与(x_1-x_2) 垂直，即***W***与***H***正交。
  - 一般说，超平面***H***把特征空间分成两个半空间。即Ω_1,Ω_2空间，当x在Ω_1空间时$g(x)>0$,***W***指向Ω_1，为***H***的正侧，反之为***H***的负侧.

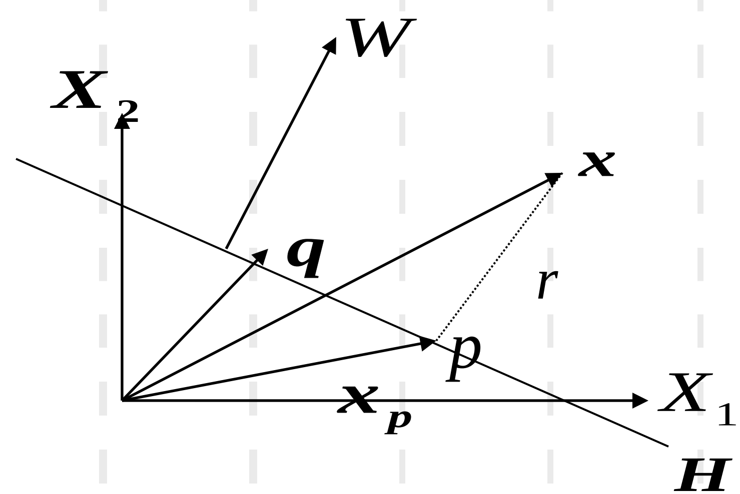
3、超平面的几何性质（续）

- 性质 ②: $\|r\| = \dfrac{g(x)}{\|W\|}$

- $\boldsymbol{x}$矢量到H的正交投影 $\|r\|$与 $g(x)$ 值成正比

$$x = x_p + r = x_p + |r| \frac{W}{\|W\|}$$

  - 其中： x_p： x在H 的投影向量，
  - r是x 到H 的垂直距离。
  - $\dfrac{W}{\|W\|}$是W方向的单位向量。



3、超平面的几何性质（续）

- 另一方面:

$$g(x) = W^T x + w_{n+1} = W^T (x_p + r) + w_{n+1}$$

$$= W^T x_p + W^T r + w_{n+1}$$

因为p在H上。$\therefore W^T x_p + w_{n+1} = 0$

$$\therefore g(x) = W^T r = W^T (\|r\| \frac{W}{\|W\|}) = \|r\| \frac{W^T W}{\|W\|} = \|r\| \|W\|$$

$$\therefore \|r\| = \frac{g(x)}{\|W\|}, \|r\| \text{是投影的绝对值} (W^T W = \|W\|^2)$$

- 这是超平面的第二个性质，矢量x到超平面的正交投影$\|r\|$正比与g(x)的函数值。

3、超平面的几何性质（续）

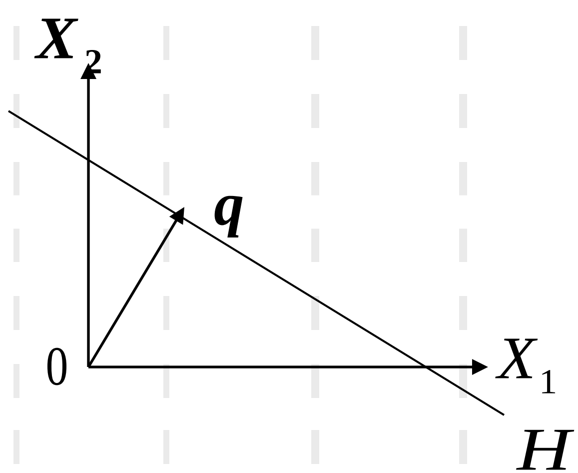
❖性质③：$\|q\| = \dfrac{W_{n+1}}{\|W\|}$,原点到$H$的距离与$W_{n+1}$成正比

因为$g(x) = W^T x + w_{n+1} = w_{n+1}$(因原点$x = 0$)

$$\therefore \|r\| = \frac{g(x)}{\|W\|} = \frac{W_{n+1}}{\|W\|} = \|q\|$$

因$x = 0$时x到H的投影为$\|r\| = \|q\|$

$$\therefore \|q\| = \frac{W_{n+1}}{\|W\|}$$



3、超平面的几何性质（续）

- 性质④：

若$W_{n+1}>0$,则H在原点正侧，若$W_{n+1}<0$,则H在原点负侧。

若$W_{n+1}=0$,则$g(x)=W^T x$，说明超平面H通过原点。

结论：

（a）超平面H的平面与W正交，方向由W决定

(b)超平面H的位置由W_{n+1}来决定。

$(c)g(x)$正比于x到H代数距离。

x在H的正侧，$g(x)>0$,否则，反之。

§2-4 线性分类器的设计

上一节我们讨论了线性判别函数形式为:$g(x)=W^TX$

其中 $X=(X_1, X_2 \ldots X_n)$　　n维特征向量

$W=(W_1, W_2 \ldots W_n, W_{n+1})$　　n维权向量

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

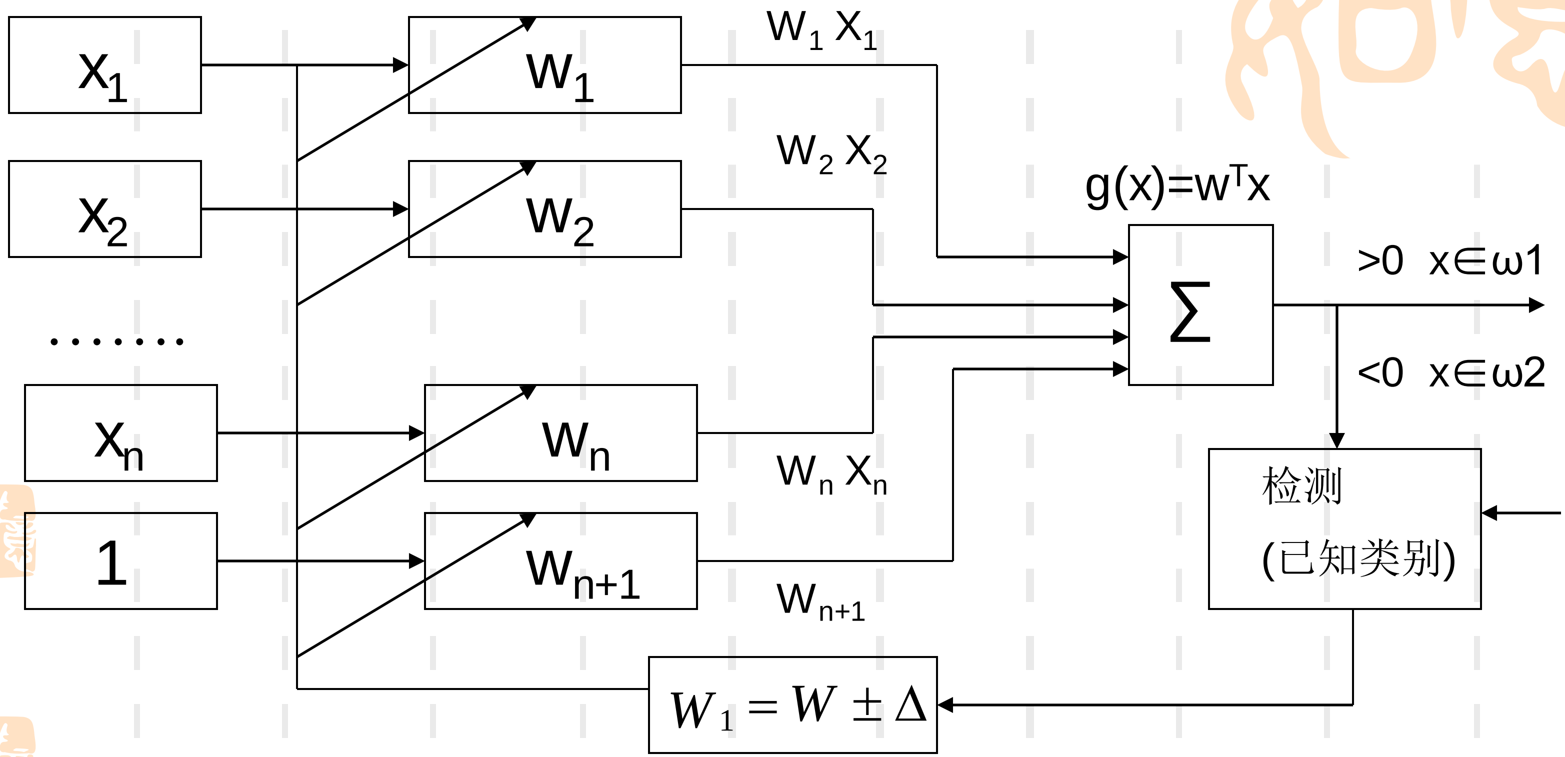
通常通过特征抽取可以获得n维特征向量，因此n维权向量是要求解的。

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。

基本概念

- 样本集：特征向量集。
- 训练样本集：用于训练分类器的样本集合(或由训练样本构成的集)。
- 测试样本集：用于检测分类器性能的样本集合
- 一般来说，要确定加权向量w，必须首先采集一些样本，并把它们变换为特征空间里的模式向量。这样的模式样本就是训练样本。为此而作的实验就是训练实验。训练实验一般是在各个类型中独立进行。如果训练样本的类别属性是预先已知的，这样的训练实验就是预分类的训练实验，又叫监督试验

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 $(X_a, X_b) \in W_1$时 $g(x) > 0$

当 $(X_c, X_d) \in W_2$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：

$$X_{1a}W_1 + X_{2a}W_2 + W_3 > 0 \quad ①$$

$$X_{1b}W_1 + X_{2b}W_2 + W_3 > 0 \quad ②$$

$$X_{1c}W_1 + X_{2c}W_2 + W_3 < 0 \quad ③$$

$$X_{1d}W_1 + X_{2d}W_2 + W_3 < 0 \quad ④$$

求出W_1, W_2, W_3

将③ ④式正规化，得

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

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

所以 g(x) $=W^TX > 0$ 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1矩阵

为N*(n+1）矩阵

N为样本数，n为特征数

确定线性分类器的主要步骤

- 第一步：采集训练样本，构成训练样本集。其中每个样本的属性是预知的，而且样本已经经过变换/选择，得到对分类最有效的d个特征。
- 第二步：确定一个准则 J ,满足：
  - $J = J(w, x)$
  - $J(w, x)$能反映分类器的性能，且对于的权值,所得分类器最优。
- 第三步：求 $J(w,x)$ 的权值，得到解向量 w^* ,同时设计求解w的最优算法。
- 一旦得到 w^* ，训练实验的任务就完成了。在下面研究各种训练算法时，假定第一步已经完成。

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) $=W^{T}X$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的问题

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，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

欲对不等式方程组$W^TX>0$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标量函数求极值的问题。解决此类问题的方法是梯度下降法。

方法就是从起始值W_1开始，算出W_1处目标函数的梯度矢量▽$J(W_1)$，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

▽$J(W_1)$为W_1处的目标函数的梯度矢量

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

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矢量。

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)二阶台劳级数展开式为

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

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：

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

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$W=W_{k+1}$上式为

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

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

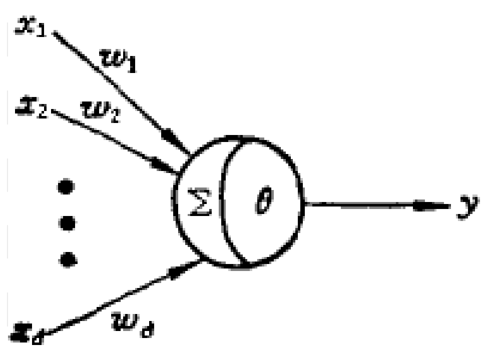
讨论：牛顿法比梯度法收敛的更快，但是D的计算量大并且要计算D^{-1}。当D为奇异时，无法用牛顿法。

二、感知器法

感知器（perceptron）是一个具有单层计算单元的人工神经网络。感知器训练算法就是由这种神经网络演变来的。

1．感知器的概念

美国学者F. Rosenblatt在1957年提出了感知器的模型，如下图。



感知器

x_i 为输入

y 为输出

θ 为阈值

感知器实质上是一种神经元模型，是一个多输入/单输出的非线性器件。用一个数学表达式表示，就是：

$$y = f\left(\sum_{i=1}^{d} w_i x_i - \theta\right)$$

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：$J(W) = \sum_{X \in X_0} \left(-W^T X\right)$ 其中x_0为错分样本

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。

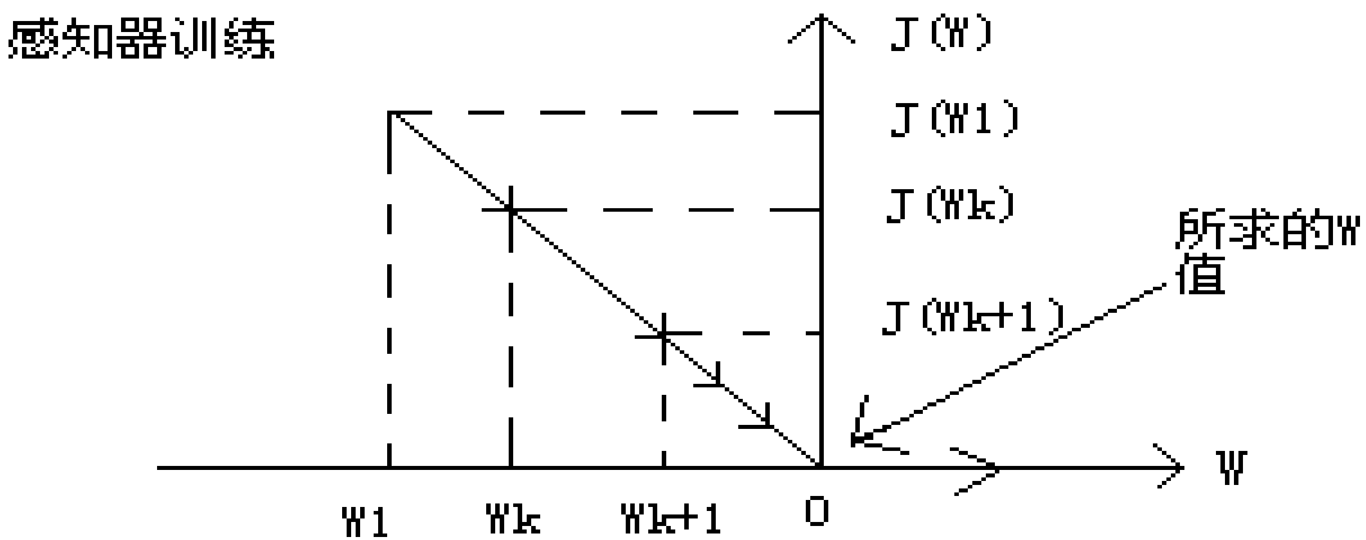
理想情况为 $J(W) = 0$ 即求最小值的问题。

求最小值对W求梯度 $\nabla J = \frac{\partial J(W)}{\partial W} = \sum_{X \in X_0}(-X)$

代入迭代公式中$W_{k+1} = W_k - \rho_k \nabla J$

即感知器迭代公式：$W_{k+1} = W_k + \rho_k \sum_{X \in X_0} X$

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

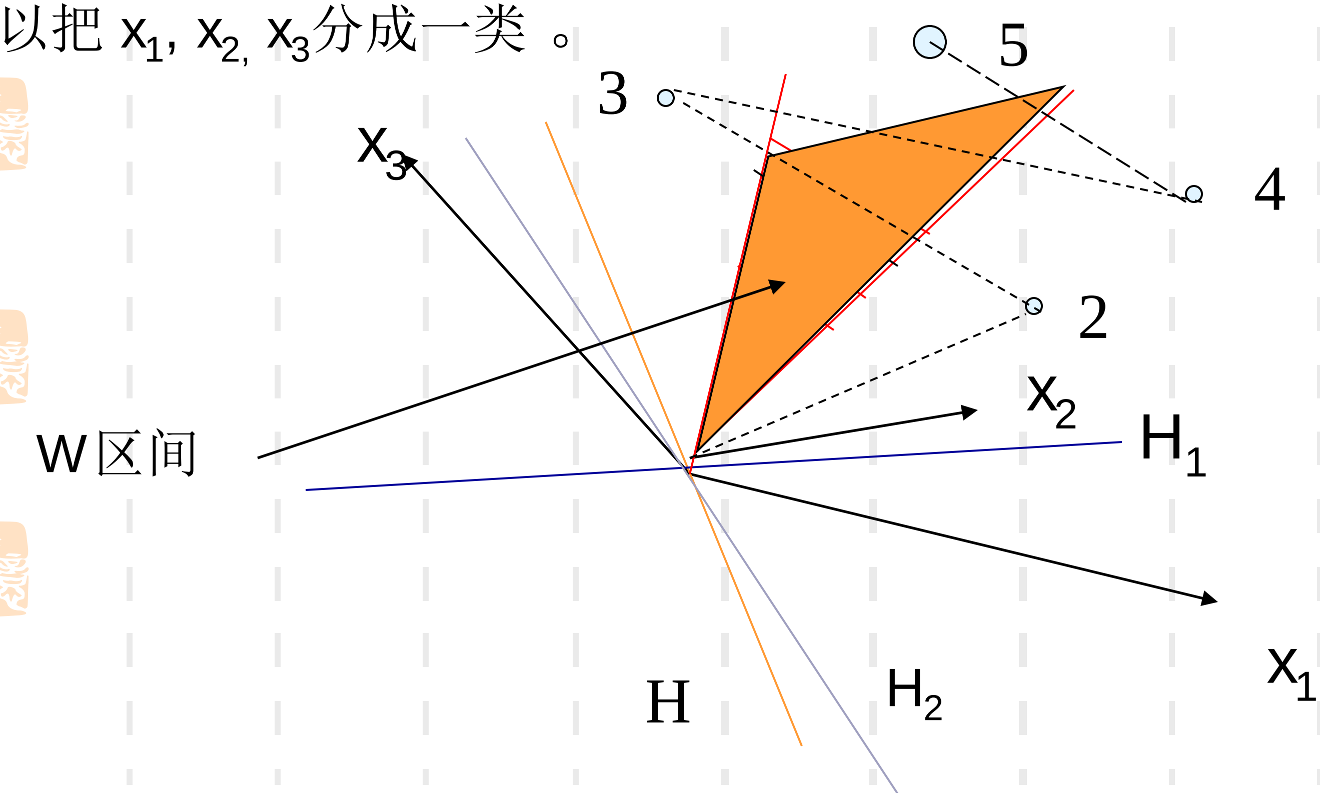


W的训练过程：

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_3三个矢量相加得矢量2,垂直于矢量2的超平面H将x_3错分.
2. x_3与矢量2相加得矢量3,垂直于矢量3的超平面H_1,将x_1错分.
3. 依上法得矢量4,垂直于矢量4做超平面, H_2将x_3错分
4. x_3与矢量4相加得矢量5,矢量5在解区内,垂直于矢量5的超平面可以把 x_1, x_2, x_3分成一类 。



2. 感知器训练算法

已知感知器输出函数：$y=\begin{cases}1\ , & \sum_{i=1}^{d} w_i x_i-\theta>0 \\ -1, & \sum_{i=1}^{d} w_i x_i-\theta\le 0\end{cases}$

令：$w_{d+1}=-\theta$，$w_0^T=(w_1,w_2,\ldots\ldots w_d)$，$g(x)=w_0^T x+w_{d+1}$（$g(x)$ 是判决函数）

又令：$y=1$，$y=-1$ 分别表示两种类型，则感知器输出函数变为：

若：$g(x)\begin{cases}>0 \\ \le 0\end{cases}$，则 $x\in\begin{cases}w_i \\ w_j\end{cases}$

［可以这样考虑：用样本进行训练时，若 $x\in w_i, g(x)>0$，则 w 不变。而 $g(x)<0$，则修改 w，直到所有样本都满足条件为止。］

针对 w_i / w_j 两类问题，利用增广型模式向量和增广型加权向量及判决规则： $w^T x\begin{cases}>0\\<0\end{cases}$， 则 $x\in\begin{cases}w_i\\w_j\end{cases}$， 说明感知器算法。

设训练样本集 $X=\{x_1, x_2, \ldots\ldots x_n\}$， 其中每个样本 $x_k, k=1,2,\ldots\ldots,n$， 分别属于类型 w_i 或类型 w_j， 且 x_k 类别属性已知。为了确定加权向量 w^*，执行：

① 给定初始值：置 $k=0$，分别给每个权向量赋任意值，可选常数 $c>0$。

② 输入训练样本 x_k， $x_k\in\{x_1, x_2, \ldots\ldots, x_n\}$

③ 计算判决函数值： $g(x_k)=[w(k)]^T x_k$

④ 修正权向量 $w(k)$，修正规则如下：

若 $x_k \in w_i$ 和 $g(x_k) \le 0$，则 $w(k+1) = w(k) + cx_k$

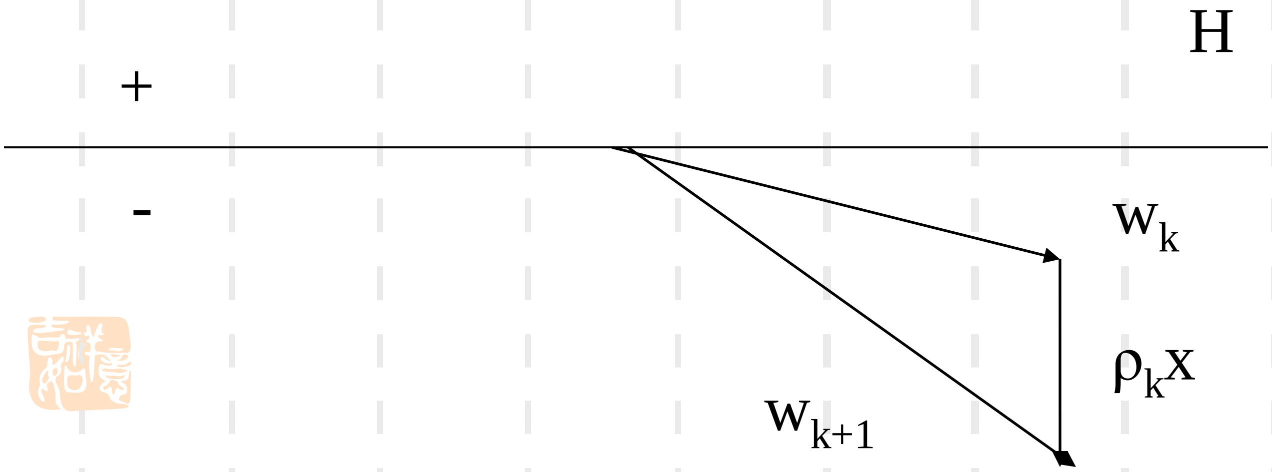
若 $x_k \in w_j$ 和 $g(x_k) > 0$，则 $w(k+1) = w(k) - cx_k$

如果类型 w_j 的训练样本 x_k 的各分量均乘以（－1），则修正规则统一为：

若： $g(x_k) \le 0$，则 $w(k+1) = w(k) + cx_k$

⑤ 令 $k = k+1$，返回②。直到 w 对所有训练样本均稳定不变，则结束。一般情况，$0 < c \le 1$。c 值大小会影响收敛速度和稳定性，c 太小收敛速度慢，c 太大，会使 $w(k)$ 的值不稳定。

- 感知器算法实质上是一种赏罚过程
  - 对正确分类的模式则“赏”，实际上是“不罚”，即权向量不变。
  - 对错误分类的模式则“罚”，使w(k)加上一个正比于x_k的分量。
  - 当用全部模式样本训练过一轮以后，只要有一个模式是判别错误的，则需要进行下一轮迭代，即用全部模式样本再训练一次。
  - 如此不断反复直到全部模式样本进行训练都能得到正确的分类结果为止。



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 $\rho_k = \lambda \dfrac{|w^T x|}{x^T x}$ $0 < \lambda \leq 2$

例题：有两类样本

ω_1=（x_1,x_2）={(1,0,1),(0,1,1)}

ω_2=（x_3,x_4）={(1,1,0),(0,1,0)}

解：先求四个样本的增值模式

$x_1=(1,0,1,1)$ $x_2=(0,1,1,1)$

$x_3=(1,1,0,1)$ $x_4=(0,1,0,1)$

假设初始权向量 $w_1=(1,1,1,1)$ $\rho_k=1$

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

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

$w_1^Tx_3=(1,1,1,1)\ (1,1,0,1)^T=3>0$ 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w_3=(0,-1,1,-1),再进行第2,3,…次迭代如下表

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	+	w_1	1 1 1 1	1
x_2 0 1 1 1	+	w_1	1 1 1 1	
x_3 1 1 0 1	+	w_1-x_3	0 0 1 0	
x_4 0 1 0 1	0	w_2-x_4	0 –1 1 -1	
x_1 1 0 1 1	0	w_3+x_1	1 –1 2 0	2
x_2 0 1 1 1	+	w_4	1 –1 2 0	
x_3 1 1 0 1	0	w_4-x_3	0 –2 2 –1	
x_4 0 1 0 1	-	w_5	0 –2 2 -1	
x_1 1 0 1 1	+	w_5	0 –2 2 –1	3
x_2 0 1 1 1	-	w_5+x_2	0 –1 3 0	
x_3 1 1 0 1	-	w_6	0 –1 3 0	
x_4 0 1 0 1	-	w_6	0 –1 3 0	
x_1 1 0 1 1	+	w_6	0 –1 3 0	4
x_2 0 1 1 1	+	w_6	0 –1 3 0	
x_3 1 1 0 1	-	w_6	0 –1 3 0	
x_4 0 1 0 1	-	w_6	0 –1 3 0	

直到在一个迭代过程中权向量相同，训练结束。

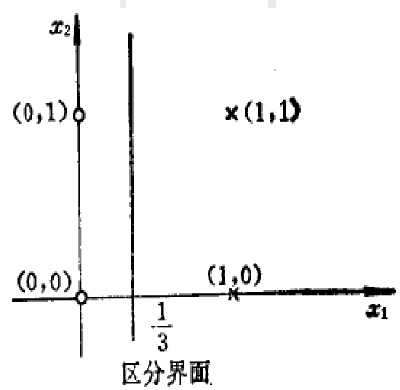
w_6=w=(0,1,3,0)　　　　判别函数$g(x)= -x_2+3x_3$

- 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.

例：一个两类问题，4个训练样本。

w_1： $(0,0)^T, (0,1)^T$

w_2： $(1,0)^T, (1,1)^T$



用感知器求权向量w^*。

解：将训练样本变为增广型的，w_2样本乘以(−1)，得到4个样本：

$X=\{x_1,x_2,x_3,x_4\}$

增广向量：$x_1=(0,0,1)^T$，$x_2=(0,1,1)^T$，$x_3=(-1,0,-1)^T$，$x_4=(-1,-1,-1)^T$。

$$g(x)=w^Tx=(w_1,w_2,\ldots w_d,w_{d+1},)\begin{pmatrix}x_1\\ \vdots\\ x_d\\ 1\end{pmatrix}$$

取初值：$w(0)=(1,1,1)^T$，$c=1$。

$k=0$，$x_k=x_1$，$g(x_k)=w(k)^Tx_k=(1,1,1)\begin{pmatrix}0\\0\\1\end{pmatrix}=1>0$，$w(1)=w(0)$，$w$不变

$k=1$，$x_k=x_2$，$g(x_k)=w(k)^Tx_k=(1,1,1)\begin{pmatrix}0\\1\\1\end{pmatrix}=2>0$，$w(2)=w(1)$

$k=2$，$x_k=x_3$，$g(x_k)=w(k)^Tx_k=(1,1,1)\begin{pmatrix}-1\\0\\-1\end{pmatrix}=-2<0$，$w(3)=w(2)+x_3=(0,1,0)^T$

$$k=3，\quad x_k=x_4，\quad g(x_k)=w(k)^T x_k=(0,1,0)\begin{pmatrix}-1\\-1\\-1\end{pmatrix}=-1<0，\quad w(4)=w(3)+x_4=(-1,0,-1)^T$$

$$k=4，\quad x_k=x_1，\quad g(x_k)=w(k)^T x_k=(-1,0,-1)\begin{pmatrix}0\\0\\1\end{pmatrix}=-1<0，\quad w(5)=w(4)+x_1=(-1,0,0)^T$$

$$k=5，\quad x_k=x_2，\quad g(x_k)=w(k)^T x_k=(-1,0,0)\begin{pmatrix}0\\1\\1\end{pmatrix}=0，\quad w(6)=w(5)+x_2=(-1,1,1)^T$$

$$k=6，\quad x_k=x_3，\quad g(x_k)=w(k)^T x_k=(-1,1,1)\begin{pmatrix}-1\\0\\-1\end{pmatrix}=0，\quad w(7)=w(6)+x_3=(-2,1,0)^T$$

$$k=7，\quad x_k=x_4，\quad g(x_k)=w(k)^T x_k=(-2,1,0)\begin{pmatrix}-1\\-1\\-1\end{pmatrix}=1>0，\quad w(8)=w(7)$$

$$k=8，\quad x_k=x_1，\quad g(x_k)=w(k)^T x_k=(-2,1,0)\begin{pmatrix}0\\0\\1\end{pmatrix}=0，\quad w(9)=w(8)+x_1=(-2,1,1)^T$$

$$k=9\text{，}\quad x_k=x_2\text{，}\quad g(x_k)=w(k)^T x_k=(-2,1,1)\begin{pmatrix}0\\1\\1\end{pmatrix}=2>0\text{，}\quad w(10)=w(9)$$

$$k=10\text{，}\quad x_k=x_3\text{，}\quad g(x_k)=w(k)^T x_k=(-2,1,1)\begin{pmatrix}-1\\0\\-1\end{pmatrix}=1>0\text{，}\quad w(11)=w(10)$$

$$k=11\text{，}\quad x_k=x_4\text{，}\quad g(x_k)=w(k)^T x_k=(-2,1,1)\begin{pmatrix}-1\\-1\\-1\end{pmatrix}=0\text{，}\quad w(12)=w(11)+x_4=(-3,0,0)^T$$

$$k=12\text{，}\quad x_k=x_1\text{，}\quad g(x_k)=w(k)^T x_k=(-3,0,0)\begin{pmatrix}0\\0\\1\end{pmatrix}=0\text{，}\quad w(13)=w(12)+x_1=(-3,0,1)^T$$

$$k=13\text{，}\quad x_k=x_2\text{，}\quad g(x_k)=w(k)^T x_k=(-3,0,1)\begin{pmatrix}0\\1\\1\end{pmatrix}=1>0\text{，}\quad w(14)=w(13)$$

$$k=14\text{，}\quad x_k=x_2\text{，}\quad g(x_k)=w(k)^T x_k=(-3,0,1)\begin{pmatrix}-1\\0\\-1\end{pmatrix}=2>0\text{，}\quad w(15)=w(14)$$

$$k=15，\quad x_k=x_3，\quad g(x_k)=w(k)^T x_k=(-3,0,1)\begin{pmatrix}-1\\-1\\-1\end{pmatrix}=2>0，\quad w(16)=w(15)$$

$$k=16，\quad x_k=x_1，\quad g(x_k)=w(k)^T x_k=(-3,0,1)\begin{pmatrix}0\\0\\1\end{pmatrix}=1>0，\quad w(17)=w(16)$$

从 $k=13\sim16$ 的结果可以看出，使用 $w(13)$ 已经能对所有训练样本正确分类，也就是算法收敛于 $w(13)$。$w(13)$ 即为解向量。

$$w^*=(-3,0,1)^T，\quad g(x)=w^* x=(-3,0,1)\begin{bmatrix}x_1\\x_2\\1\end{bmatrix}=-3x_1+1$$

分界面 $g(x)=0$，也就是：$x_1=\dfrac{1}{3}$，如上图所示。w^* 是不唯一的，w^* 属于解集空间。

3．感知器训练算法的收敛性

若训练样本是线性可分的，则感知器训练算法在有限次迭代后，可以收敛到正确的解向量。下面就对该收敛性进行证明。

假定：(1) 对每个样本进行归一化处理，变为单位向量。$\|x_k\|=1$，$k=1,2,\ldots,n$。

(2) 取常数 $c=1$，取初值 $w(0)=0$。

(3) 对来自 w_j 的训练样本的各个分量均乘以(−1)。

对感知器训练算法变更如下：

① 置 $k=0$，选初值 $w(0)=0$，$c=1$，给较小正数 T。

② 输入训练样本 x_k，$x_k \in \{x_1, x_2, \ldots, x_n\}$

③ 计算： $g(x_k)=w(k)^T x_k$

④ 判断： 若 $g(x_k)>T$， 返回②， 否则继续。

⑤ 令 $w(k+1)=w(k)+x_k$， $x_k=1,2,\ldots,n$， $k=k+1$ 返回②。

若所有样本均不进入⑤，则算法结束。

从算法中不难看出：

$$w(k)=w(0)+\sum_{i=0}^{k-1}x_i=\sum_{i=0}^{k-1}x_i$$

两边乘以解向量 w^*：

$$w_0^{\ *}w(k)=\sum_{i=0}^{k-1}(w^*x_i)>kT$$ ，（由算法的③④知道）

由于： $\|w(k)\|^2=\|w(k-1)+x_{k-1}\|^2<=\|w(k-1)\|^2+\|x_{k-1}\|^2$

归一化 $\|x_{k-1}\|^2=1$，

因此：$\|w(k)\|^2 \le \|w(k-1)\|^2 + 1$

分别计算出 $\|w(k-1)\|^2$， $\ldots\|w(1)\|^2$，代入上式得：

$$\|w(k)\|^2 < k$$

令：$c(k)$表示两个向量w^*与$w(k)$之间的夹角余弦，则：

$$c(k) = \frac{w^* \cdot w(k)}{\|w^*\| \cdot \|w(k)\|} > \frac{kT}{\sqrt{k} \cdot \|w^*\|} > 0$$

又$c(k) < 1$， 令 $T' = \dfrac{T}{\|w^*\|}$， 则有：$k < \dfrac{1}{T'^2}$ $,T' > 0$。

也就是，k是一个有限的值，步骤⑤仅需有限次数就可以得到w^*。换句话说，在线性可分的情况下，感知器训练算法一定收敛。正数T越小，则收敛越慢（k值越大）。

4. 感知器训练算法在多类问题中的应用(推广)

在上一节，介绍了多类问题的三种情况。其中，第三种情况是没有不确定区的，对于c种类型中的某一种类型w_i，存在k个判别函数，如果样本$x \in w_i$，则：$g_i(x) > g_j(x) \quad j=1,2,\ldots k \quad i \neq j$，其中$k$取决于在特征空间里与类型$w_i$相邻的类型数目，$k \leq c$。这样就把多类问题转变成多个没有不确定区的$w_i / w_j$两类问题，可使用最大值判决规则。

假定$k=c$，首先建立c个判决函数：$g_i(x) = w_i^T x$，$i=1,2,\ldots,c$。

判决规则为：若$g_i(x) > g_j(x)$，$j=1,2,\ldots,c$，$j \neq i$，则$x \in w_i$。

将本节2中介绍的感知器训练算法用到这种情况，可以建立如下的算法步骤。

① 赋给初始值：分别赋给 c 个权向量 w_i（$i=1,2,\ldots,c$）任意的初值，选择正常数c，把训练样本变为增广型模式向量，置 $k=0$。

② 输入训练样本 x_k，$x_k \in \{x_1, x_2, \ldots, x_n\}$，假定 $x_k \in w_i$

③ 计算c个判决函数值：$g_i(x_k) = w_i(k)^T x_k \quad i=1,2,\ldots c$

④ 修正权向量。修正规则为：

若 $g_i(x) > g_j(x)$，$j=1,2,\ldots,c$，$j \neq i$，则 $w_i(k+1) = w_i(k)$，$i=1,2,\ldots,c$。

否则，若 $g_l(x_k) > g_i(x_k)$，则：
$$\begin{cases} w_i(k+1) = w_i(k) + c x_k \\ w_l(k+1) = w_l(k) - c x_k \\ w_j(k+1) = w_j(k) \quad j \neq i, j \neq l \end{cases}$$

⑤ 令 $k=k+1$，返回②。直到所有的权向量对所有训练样本都稳定不变时结束。只要模式样本线性可分，则算法迭代有限次后收敛。

例：已知三类训练样本：$w_1=(0,0)^{\mathrm{T}}$，$w_2=(1,1)^{\mathrm{T}}$，$w_3=(-1,1)^{\mathrm{T}}$

试求解 w_1^*，w_2^*，w_3^*。

解：训练样本变成增广型模式向量：

$$x_1=(0,0,1)^T \text{，} \quad x_2=(1,1,1)^T \text{，} \quad x_3=(-1,1,1)^T$$

在此，x 的下标就是它所属类型，且没有一个样本乘以(−1)。置 $k=0$，选 $c=1$。

赋初值：$w_1(0)=(0,0,0)^T$，$w_2(0)=(0,0,0)^T$，$w_3(0)=(0,0,0)^T$，开始迭代：

$k=0$，$x_k=x_1\in w_1$ 但 $g_1(x_k)=0$，$g_2(x_k)=0$，$g_3(x_k)=0$，$g_1(x_k)\not> g_2(x_k)$，$g_1(x_k)\not> g_3(x_k)$。

所以：
$$\begin{cases} w_1(1)=w_1(0)+x_1=(0,0,1)^T \\ w_2(1)=w_2(0)-x_1=(0,0,-1)^T \\ w_3(1)=w_3(0)-x_1=(0,0,-1)^T \end{cases}$$

$k=1$， $x_k = x_2 \in w_2$ 但 $g_1(x_k) = (0,0,1)\begin{pmatrix}1\\1\\1\end{pmatrix} = 1$， $g_2(x_k) = (0,0,-1)\begin{pmatrix}1\\1\\1\end{pmatrix} = -1$，

$g_3(x_k) = -1$， $g_1(x_k) > g_2(x_k)$， $g_2(x_k) \ngtr g_3(x_k)$。

所以：
$$\begin{cases} w_1(2) = w_1(1) - x_2 = (-1,-1,0)^T \\ w_2(2) = w_2(1) + x_2 = (1,1,0)^T \\ w_3(2) = w_3(1) - x_2 = (-1,-1,-2)^T \end{cases}$$

$k=2$， $x_k = x_3 \in w_3$ 但 $g_1(x_k) = (-1,-1,0)\begin{pmatrix}-1\\1\\1\end{pmatrix} = 0$， $g_2(x_k) = (1,1,0)\begin{pmatrix}-1\\1\\1\end{pmatrix} = 0$， $g_3(x_k) =$

-2， $g_3(x_k) < g_1(x_k)$　$g_3(x_k) < g_2(x_k)$。

所以：
$$\begin{cases} w_1(3) = w_1(2) - x_3 = (0,-2,-1)^T \\ w_2(3) = w_2(2) - x_3 = (2,0,-1)^T \\ w_3(3) = w_3(2) + x_3 = (-2,0,-1)^T \end{cases}$$

$k=3$， $x_k=x_1\in w_1$ 但 $g_1(x_k)=(0,-2,-1)\begin{pmatrix}0\\0\\1\end{pmatrix}=-1$， $g_2(x_k)=(2,0,-1)\begin{pmatrix}0\\0\\1\end{pmatrix}=-1$，

$g_3(x_k)=-1$， $g_1(x_k)\ngtr g_2(x_k)$， $g_1(x_k)\ngtr g_3(x_k)$。

所以： $\begin{cases}w_1(4)=w_1(3)+x_1=(0,-2,0)^T\\ w_2(4)=w_2(3)-x_1=(2,0,-2)^T\\ w_3(4)=w_3(3)-x_1=(-2,0,-2)^T\end{cases}$

$k=4$， $x_k=x_2\in w_2$ 但 $g_1(x_k)=(0,-2,0)\begin{pmatrix}1\\1\\1\end{pmatrix}=-2$， $g_2(x_k)=(2,0,-2)\begin{pmatrix}1\\1\\1\end{pmatrix}=0$，

$g_3(x_k)=-4$， $g_2(x_k)>g_1(x_k)$ $g_2(x_k)>g_3(x_k)$

所以： $\begin{cases}w_1(5)=w_1(4)\\ w_2(5)=w_2(4)\\ w_3(5)=w_3(4)\end{cases}$

$k=5$， $x_k=x_3\in w_3$ 但 $g_1(x_k)=(0,-2,0)\begin{pmatrix}-1\\1\\1\end{pmatrix}=-2$， $g_2(x_k)=(2,0,-2)\begin{pmatrix}-1\\1\\1\end{pmatrix}=-4$，

$g_3(x_k)=0$， $g_3(x_k)>g_1(x_k)$ $g_3(x_k)>g_2(x_k)$

所以： $\begin{cases}w_1(6)=w_1(5)\\w_2(6)=w_2(5)\\w_3(6)=w_3(5)\end{cases}$

$k=6$， $x_k=x_1\in w_1$ 但 $g_1(x_k)=(0,-2,0)\begin{pmatrix}0\\0\\1\end{pmatrix}=0$， $g_2(x_k)=(2,0,-2)\begin{pmatrix}0\\0\\1\end{pmatrix}=-2$，

$g_3(x_k)=-2$， $g_1(x_k)>g_2(x_k)$ $g_1(x_k)>g_3(x_k)$

因此，可得到3个解向量： $w_1^*=w_1(4)$， $w_2^*=w_2(4)$， $w_3^*=w_3(4)$。

对应的三个判决函数： $g_1(x)=-2x_2$， $g_2(x)=2x_1-2x_2$， $g_3(x)=-2x_1-2$。

5.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，可以用上述方法解到正确分类的权向量。当样本集线性不可分时，用上述方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：在样本

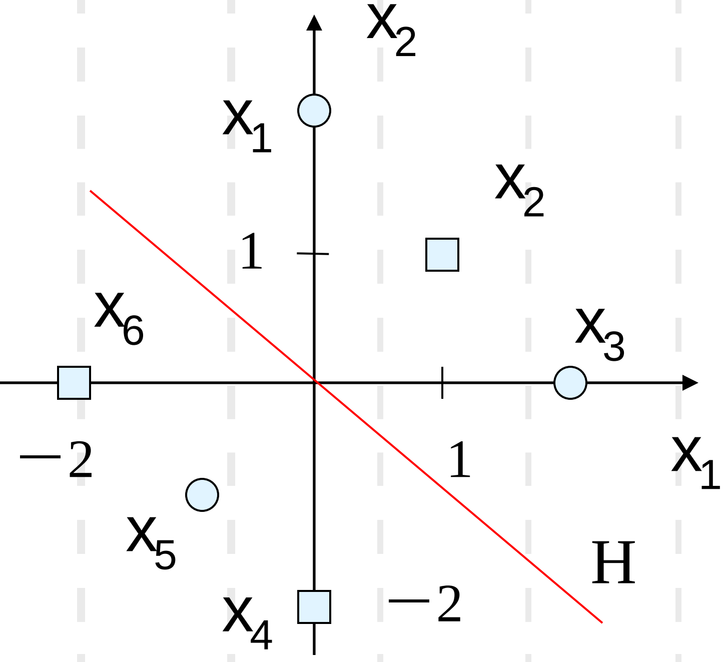
ω_1: X_1 =（0,2） X_3 =（2,0）

X_5 =（-1,-1）

ω_2: X_2 =（1,1） X_4 =（0,-2）

X_6 =（-2,0）

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。

- 作业：已知四个训练样本

  $w_1=\{(0,0),(0,1)\}$

  $w_2=\{(1,0),(1,1)\}$

  使用感知器固定增量法求判别函数

  设$w_1=(1,1,1,1)$　　$\rho_k=1$

  要求编写程序上机运行，写出判别函数，并打出图表。

三、最小平方误差准则(MSE法)---非迭代法

针对 w_i/w_j 二类问题，利用增广型加权向量和增广型模式向量，并把所有来自 w_j 的训练样本的各分量均乘以（－1），则所有模式都应该满足：$w^T\cdot x>0$。

现在任意给定一个小的正数 b，则可在解区里寻找一个解向量 w，使之满足：$w^T\cdot x=b$。

假定有 n 个训练样本，则上式可以写成 n 个联立方程组：

$$\begin{cases} w^T x_1 = b_1 \\ w^T x_2 = b_2 \\ \quad\vdots \\ w^T x_n = b_n \end{cases}$$

式中 $b_i>0, i=1,2,\dots,n$

上述方程组可简写为：$x\cdot w=b$ (是一个超定方程组)。

其中，x为训练样本的增广矩阵：

$$x=\begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}=\begin{pmatrix} x_{11} & x_{12}\cdots\cdots x_{1d} & 1 \\ \vdots & & \\ x_{n_1 1} & x_{n_1 2}\cdots\cdots x_{n_1 d} & 1 \\ x_{(n_1+1)1} & x_{(n_1+1)2}\cdots x_{(n_1+1)d} & -1 \\ \vdots & & \\ x_{n_2 1} & x_{n_2 2}\cdots\cdots x_{n_2 d} & -1 \end{pmatrix}\begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1 \in w_i \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2 \in w_j \end{matrix}$$

$$b=[b_1,\ b_2,\ \ldots,\ b_n]^T$$

注意：假定来自w_i和w_j的样本数目分别为n_1和n_2，且$n=n_1+n_2$，x为$n\times(d+1)$维矩阵，一般$n>d+1$。b为n维列向量。w为$d+1$维列向量。

如果样本是线性可分的，则x的任意$(d+1)\times(d+1)$子阵的秩均等于$d+1$。这时可以通过最小二乘法求解。

定义：误差向量 $e = xw - b$

又定义平方误差准则函数：$J_s(w,x,b) = \| e \|^2 = \| xw - b \|^2 = \sum_{i=1}^{n} (w^T x_i - b_i)^2$

约束条件：要使 J_s 得到的解 w 就是解向量 w^*，必须保证 $w^T x_i - b_i \geq 0$。

下面求 J_s 的极小值：$\frac{\partial J_s}{\partial w} = 2x^T (xw^* - b)$

令 $\frac{\partial J_s}{\partial w} = 0$，得到 $2x^T (xw^* - b) = 0$ → $x^T xw^* - x^T b = 0$

所以有：$w^* = (x^T x)^{-1} x^T b$

上式中：用 $x^\# = (x^T x)^{-1} x^T$ 代入，得到 $w^* = x^\# b$。

其中，$x^\#$ 称为 x 的伪逆，w^* 是伪逆解。

取：

$$b = \begin{bmatrix} N/N_1 \\ \cdots\cdots\cdots \\ N/N_1 \\ N/N_2 \\ \cdots\cdots \\ N/N_2 \end{bmatrix}$$

其中N/N$_1$有N$_1$个，N/N$_2$有N$_2$个

最小平方误差法同Fisher法是一致的

四、韦—霍氏法（LMS法）迭代法

上节得到MSE法的W解为：$W=X^{\#}b$

伪逆$X^{\#}=\left(X^{T}X\right)^{-1}X^{T}$　　计算量很大

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比较大误差

所以要用迭代法来求

求J(W)的梯度

$\nabla J(W)=2X^{T}(XW-b)$　代入迭代公式　W_1任意设定

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

令$\rho_k=\dfrac{\rho_1}{k}$其中ρ_1为任意常数

此法可收敛于W值。W满足: $X^{T}(XW-b)=0$

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$\rho_K = \frac{\rho_1}{k}$

ρ_k随迭代次数k而减少，以保证算法收敛于满意的W值，b值同样是取定的常值。

五、 何—卡氏法（判断迭代过程中是否线性可分）

前面两节中 w^* 还不是最小平方误差准则函数下的伪逆解。因为 w^* 还依赖于b，还要进一步确定b。

根据平方误差准则函数的定义：$J_s(w,x,b)$ 要使 J_s 最小，使用梯度下降法建立 b 的迭代式：

$$b(k+1)=b(k)-c(\frac{\partial J_s}{\partial b})_{b=b(k)}$$

这里使用固定增量法，c 为大于0的常数。

对 J_s 定义求导，$\frac{\partial J_s}{\partial b}=-2(xw-b)$

令 $\frac{\partial J_s}{\partial b}=0$， $xw-b=0$ 这就是 J_s 最小的条件。

对 b 的前后两次迭代来说，$b(k+1)=b(k)+\delta\cdot b(k)$， b 又为正值，又考虑到 $xw-b=0$，则 b 的增量 $\delta\cdot b(k)$ 应该是：

$$\delta b(k)=\begin{cases}0 & ,xw(k)-b\le 0\\ 2c[xw(k)-b(k)] & ,xw(k)-b>0\end{cases}$$

上式也可以改写为：$\delta b(k)=c[xw(k)-b(k)+|xw(k)-b(k)|]$

引入误差向量：$e_k=xw(k)-b(k)$，则有：

$$\delta b(k)=c(e_k+|e_k|)$$

将 $b(k+1)=b(k)+\delta\cdot b(k)$ 代入 $w^*=x^\# b$，则有：

$$w^*(k+1)=x^\# b(k)+x^\#\delta b(k)=w^*(k)+x^\#\delta b(k)$$

至此，我们就建立了一种算法，通常称为LMSE算法，即最小均方误差算法。LMSE：Least Mean Square Error。这种算法也简称H-K算法。其主要内容为：

$$\begin{cases} w^*(k+1)=w^*(k)+cx^\#(e_k+|e_k|) \\ w^*(1)=x^\# b(1) \\ b(1)>0 \end{cases}$$

H-K算法

该算法的步骤如下：

① 由训练样本集构成增广矩阵 x，求伪逆 $x^{\#}=(x^T x)^{-1}x^T$。

② 赋初值 $b(1)$，应使其各分量为正值。选常数 c，置 $k=1$。

③ 计算：$w(k)=x^{\#}b(k)$，$e_k=xw(k)-b(k)$。

④ 判断：若 e_k 的各分量停止变为正值或者不是全部为0，则线性不可分，中止迭代。否则，若 e_k 的各分量均接近于0，即 $e_k\to 0$，则迭代过程完成，结束。否则算法继续。

⑤ 计算：$w(k+1)=w(k)+cx^{\#}[e_k+|e_k|]=w(k)+cx^{\#}|e_k|$

$$b(k+1)=b(k)+c[e_k+|e_k|]$$

『注意推导：

$$x^{\#}e_k=(x^T x)^{-1}\cdot x^T[xw(k)-b(k)]=(x^T x)^{-1}\cdot x^T[xx^{\#}-1]b(k)=(x^T x)^{-1}\cdot(x^T-x^T)b(k)=0$$

其中：$w(k)=x^{\#}b(k)$，$x^T xx^{\#}=(x^T x)(x^T x)^{-1}x^T=x^T$』

⑥ 令$k=k+1$，返回③

注意：第⑤步中，可以先计算$b(k+1)$，然后再由公式$w(k+1)=x^{\#}b(k+1)$求$w(k+1)$。

可以证明：当模式类型可分，且$0<c\le 1$时，H-K算法收敛。证明收敛性的关键在于，在极限情况下，$e_k=xw(k)-b(k)=0$，而H-K算法中指出$b(k)$的各分量为非负的矢量，故若$xw(k)=b(k)$，则$xw(k)>0$（回到本节1中的定义）。

算法没有给出精确的迭代次数，通常在每次迭代后均对$xw(k)$和e_k进行检查，当$xw(k)>0$或$e_k=0$，则有解；反之，若e_k变为非正，则迭代停止，表明模式线性不可分。

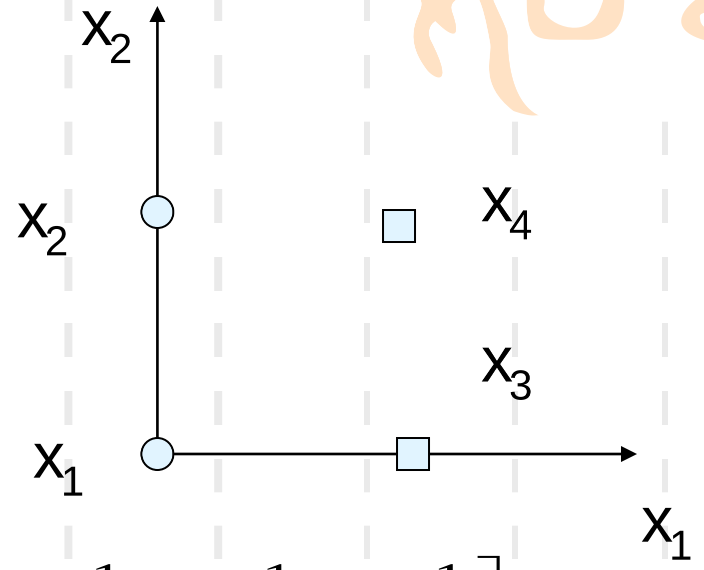
- 例题：

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

- 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的伪逆矩阵为

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

取$b_1=(1,1,1,1)^T$ $c=1$ $W_1=X^{\#}b_1=(-2,0,1)^T$

因为

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

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：

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

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

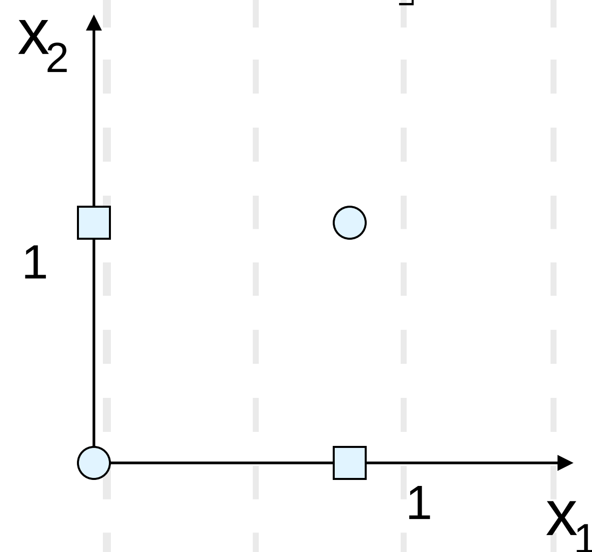
取$b_1=(1,1,1,1)^T$ $c=1$

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

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

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$e_k<0$，迭代就应立即停止。

六、 Fisher分类准则

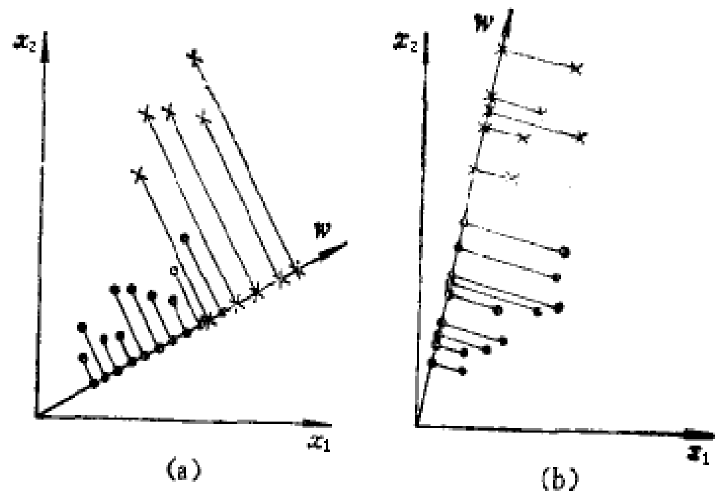
前面讲过的感知器准则、最小平方和准则属于用神经网络的方法解决分类问题。下面介绍一种新的判决函数分类方法。

由于线性判别函数易于分析，关于这方面的研究工作特别多。历史上，这一工作是从R. A. Fisher的经典论文（1936年）开始的。我们知道，在用统计方法进行模式识别时，许多问题涉及到维数，在低维空间行得通的方法，在高维空间往往行不通。因此，降低维数就成为解决实际问题的关键。Fisher的方法，实际上涉及维数压缩。

如果要把模式样本在高(d)维的特征向量空间里投影到一条直线上，实际上就是把特征空间压缩到一维，这在数学上容易办到。另外，即使样本在高维空间里聚集成容易分开的群类，把它们投影到一条任意的直线上，也可能把不同的样本混杂在一起而变得无法区分。也就是说，直线的方向选择很重要。

在一般情况下，总可以找到某个最好的方向，使样本投影到这个方向的直线上是最容易分得开的。如何找到最好的直线方向，如何实现向最好方向投影的变换，是Fisher法要解决的基本问题。这个投影变换就是我们寻求的解向量w^*。

如下图：



抽样试验在直线 w 上的投影

图中画出了直线的两种选择，图(a)中，Y_1和Y_2还无法分开，而图(b)的选择可以使Y_1和Y_2区分开来。所以图(b)的方向是一个好的选择。

1. 线性投影与Fisher准则函数

在 w_1/w_2 两类问题中，假定有 n 个训练样本 $x_k(k=1,2,\ldots.,n)$ 其中 n_1 个样本来自 w_i 类型， n_2 个样本来自 w_j 类型， $n=n_1+n_2$ 。两个类型的训练样本分别构成训练样本的子集 X_1 和 X_2 。

令： $y_k = w^T x_k$ ， $k=1,2,\ldots,n$ (1)

y_k 是向量 x_k 通过变换 w 得到的标量，它是一维的。实际上，对于给定的 w， y_k 就是判决函数的值。

由子集 X_1 和 X_2 的样本映射后的两个子集为 Y_1 和 Y_2 。因为我们关心的是 w 的方向，可以令 $\|w\|=1$， 那么 y_k 就是 x_k 在 w 方向上的投影。使 Y_1 和 Y_2 最容易区分开的 w 方向正是区分超平面的法线方向。

下面讨论怎样得到最佳w方向的解析式。

各类在d维特征空间里的样本均值向量：

$$M_i = \frac{1}{n_i} \sum_{x_k \in X_i} x_k \text{，} \quad i = 1,2 \qquad (2)$$

通过变换w映射到一维特征空间后，各类的平均值为：

$$m_i = \frac{1}{n_i} \sum_{y_k \in Y_i} y_k \text{，} \quad i = 1,2 \qquad (3)$$

映射后，各类样本“类内离散度”定义为：

$$S_i^2 = \sum_{y_k \in Y_i} (y_k - m_i)^2 \text{，} \quad i = 1,2 \qquad (4)$$

显然，我们希望在映射之后，两类的平均值之间的距离越大越好，而各类的样本类内离散度越小越好。因此，定义Fisher准则函数：

$$J_F(w) = \frac{|m_1 - m_2|^2}{s_1^2 + s_2^2} \qquad (5)$$

使J_F最大的解w^*就是最佳解向量，也就是Fisher的线性判别式。

2．求解 w^*

从 $J_F(w)$ 的表达式可知，它并非 w 的显函数，必须进一步变换。

已知：$m_i = \frac{1}{n_i}\sum_{y_k \in Y_i} y_k$， $i = 1,2$， 依次代入(1)和(2)，有：

$$m_i = \frac{1}{n_i}\sum_{x_k \in X_i} w^T x_k = w^T\left(\frac{1}{n_i}\sum_{x_k \in X_i} x_k\right) = w^T M_i, \quad i = 1,2 \qquad (6)$$

所以：
$$|m_1 - m_2|^2 = \| w^T M_1 - w^T M_2 \|^2 = \| w^T (M_1 - M_2) \|^2$$
$$= w^T (M_1 - M_2)(M_1 - M_2)^T w = w^T \vec{S}_b w \qquad (7)$$

其中：$\vec{S}_b = (M_1 - M_2)(M_1 - M_2)^T$ (8)

$\vec{S}_b$ 是原 d 维特征空间里的样本类间离散度矩阵，表示两类均值向量之间的离散度大小，因此，$\vec{S}_b$ 越大越容易区分。

将(6) $m_i = w^T M_i$ 和(2) $M_i = \frac{1}{n_i} \sum_{x_k \in X_i} x_k$ 代入(4) S_i^2 式中：

$$S_i^2 = \sum_{x_k \in X_i} (w^T x_k - w^T M_i)^2$$

$$= w^T \cdot \sum_{x_k \in X_i} (x_k - M_i)(x_k - M_i)^T \cdot w$$

$$= w^T \vec{S_i} w \qquad (9)$$

其中： $\vec{S_i} = \sum_{x_k = X_i} (x_k - M_i)(x_k - M_i)^T$， $i = 1,2$ (10)

因此： $S_1^2 + S_2^2 = w^T (\vec{S_1} + \vec{S_2}) w = w^T \vec{S_w} w$ (11)

显然： $\vec{S_w} = \vec{S_1} + \vec{S_2}$ (12)

$\vec{S_i}$ 称为原 d 维特征空间里，样本“类内离散度”矩阵。

$\vec{S_w}$ 是样本“类内总离散度”矩阵。

为了便于分类，显然 $\vec{S_i}$ 越小越好，也就是 $\vec{S_w}$ 越小越好。

将上述的所有推导结果代入 $J_F(w)$ 表达式：

$$J_F(w)=\frac{w^T\overrightarrow{S_b}w}{w^T\overrightarrow{S_w}w} \quad \text{—— 广义Rayleigh商} \qquad (13)$$

式中 $\overrightarrow{S_b}$ 和 $\overrightarrow{S_w}$ 皆可由样本集 X 计算出。

用lagrange乘子法求解 $J_F(w)$ 的极大值点。

令分母等于非零常数，也就是：$c=w^T\overrightarrow{S_w}w=c\neq 0$。

定义lagrange函数：

$$L(w,\lambda)=w^T\overrightarrow{S_b}w-\lambda(w^T\overrightarrow{S_w}w-c) \qquad (14)$$

L 对 w 求偏导数：

$$\frac{\partial L(w,\lambda)}{\partial w}=2(\overrightarrow{S_b}w-\lambda\overrightarrow{S_w}w)$$

令 $\frac{\partial L(w,\lambda)}{\partial w}=0$ 得到：

$$\overrightarrow{S_b}w^*=\lambda\overrightarrow{S_w}w^* \qquad (15)$$

从上述推导(10)～(12)可知，$\overrightarrow{S_w}$是d维特征的样本协方差矩阵，它是对称的和半正定的。当样本数目$n>d$时，$\overrightarrow{S_w}$是非奇异的，也就是可求逆。

则： $\lambda w^* = \overrightarrow{S_w}^{-1}\overrightarrow{S_b}w^*$ (16)

问题转化为求一般矩阵$\overrightarrow{S_w}^{-1}\overrightarrow{S_b}$的特征值和特征向量。令$\overrightarrow{S_w}^{-1}\overrightarrow{S_b}=A$，则$\lambda$是$A$的特征根，$w^*$是$A$的特征向量。

$$\begin{aligned}\overrightarrow{S_b}w^* &= \{(M_1-M_2)(M_1-M_2)^T\}w^* \\ &= (M_1-M_2)\{(M_1-M_2)^T w^*\} \\ &= (M_1-M_2)\cdot\gamma \qquad (17)\end{aligned}$$

式中：

$$\gamma = (M_1-M_2)^T w^*$$

是一个标量。所以$\overrightarrow{S_b}w^*$总是在(M_1-M_2)方向上。

将(17)代入到(15)，可以得到：

$$w^* = \frac{\gamma}{\lambda}\overrightarrow{S_w}^{-1}(M_1 - M_2)$$

其中，$\frac{\gamma}{\lambda}$ 是一个比例因子，不影响 w^* 的方向，可以删除，从而得到最后解：

$$w^* = \overrightarrow{S_w}^{-1}(M_1 - M_2) \tag{18}$$

w^* 就使 $J_F(w)$ 取得最大值，w^* 可使样本由 d 维空间向一维空间映射，其投影方向最好。$w^* = \overrightarrow{S_w}^{-1}(M_1 - M_2)$ 是一个Fisher线性判断式。

讨论：

如果 $M_1 = M_2$，$w^* = 0$，则样本线性不可分。

$M_1 \neq M_2$，未必线性可分。

$\overrightarrow{S_w}$ 不可逆，未必不可分。

求解Fisher判别函数

由于变换后的模式是一维的，因此判别界面实际上是各类模式所在轴上的一个点，所以可以根据训练模式确定一个阈值 y_t，于是Fisher判别规则为：

$$\vec{u}'\vec{x} = y \gtrless y_t \Rightarrow \vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

判别阈值可取两个类心在u方向上轴的投影连线的中点作为阈值，即：

$$y_t = \frac{\tilde{m}_1 + \tilde{m}_2}{2}$$

Fisher方法实现步骤总结:

（1） 把来自两类ω_1/ω_2的训练样本集X分成与ω_1对应的子集X_1和与ω_2对应的子集X_2。

（2） 由 $\vec{m}_i = \frac{1}{N_i}\sum_j \vec{x}_j^{(i)} \quad (i=1,2)$ ，计算$\boldsymbol{m}_i$。

（3） 由 $S_{W_i} = \sum_j (\vec{x}_j^{(i)} - \vec{m}_i)(\vec{x}_j^{(i)} - \vec{m}_i)'$ ，计算各类的类内离差阵S_{W1}，S_{W2}。

（4） 计算类内总离差阵$S_W=S_{W1}+S_{W2}$。

（5） 计算S_W的逆矩阵S_W^{-1}。

（6） 按 $\vec{u} = S_W^{-1}(\vec{m}_1 - \vec{m}_2)$ 求解μ。

（7） 计算$\tilde{m}_i$。

$$\tilde{m}_i = \frac{1}{N_i}\sum_j y_j^{(i)} = \frac{1}{N_i}\sum_j \vec{u}' x_j^{(i)} = \vec{u}'\vec{m}_i$$

（8） 计算y_t。$y_t = \dfrac{\tilde{m}_1 + \tilde{m}_2}{2}$

（9） 对未知模式x判定模式类。

$$\vec{u}'\vec{x} = y \gtrless y_t \Rightarrow \vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

以100元A面数据和50元A面数据为例

100元A面:(64,76,99,84,98,95,88,83),…

50元A面:(65,67,82,80,89,94,86,92),…

$N_1=N_2=60$算得:

$m_1=(69.3, 61.9, 83.5, 70.8, 97.7, 91.5, 87.6, 82.4)$

$m_2=(59.2, 55.5, 81.9, 63.9, 95.1, 91.0, 91.1, 86.5)$

$$m_1=(69.3,\ 61.9,\ 83.5,\ 70.8,\ 97.7,\ 91.5,\ 87.6,\ 82.4)$$

$$m_2=(59.2,\ 55.5,\ 81.9,\ 63.9,\ 95.1,\ 91.0,\ 91.1,\ 86.5)$$

$$S_{W1}=\begin{pmatrix} 13, & 27.7, & 30.1, & 24.7, & 8.9, & 11.1, & 8.32, & 1.93 \\ 27.7, & 111, & 112, & 89.9, & 20.7, & 37, & 19.2, & 1.23 \\ 30.1, & 112, & 135, & 105, & 22.5, & 33.9, & 24.7, & -2.24 \\ 24.7, & 89.9, & 105, & 99.1, & 13.5, & 24.4, & 20, & 4.36 \\ 8.9, & 20.7, & 22.5, & 13.5, & 14.4, & 13.8, & 10.5, & 6.02 \\ 11.1, & 37, & 33.9, & 24.4, & 13.8, & 19, & 10.7, & 5.88 \\ 8.32, & 19.2, & 24.7, & 20, & 10.5, & 10.7, & 11.2, & 6.73 \\ 1.93, & 1.23, & -2.24, & 4.36, & 6.02, & 5.88, & 6.73, & 12.7 \end{pmatrix}$$

$$m_1=(69.3,\ 61.9,\ 83.5,\ 70.8,\ 97.7,\ 91.5,\ 87.6,\ 82.4)$$

$$m_2=(59.2,\ 55.5,\ 81.9,\ 63.9,\ 95.1,\ 91,\ 91.1,\ 86.5)$$

$$S_{W2}=\begin{pmatrix} 28.3, & 45.7, & 49.6, & 48.6, & 1.59, & 17.4, & 14, & 15.2 \\ 45.7, & 124, & 63.2, & 88.6, & 6.21, & 44.1, & 15.8, & 17.4 \\ 49.6, & 63.2, & 119, & 82.6, & 3.31, & 25.9, & 38, & 28.8 \\ 48.6, & 88.6, & 82.6, & 109, & -2.86, & 33, & 14.1, & 28.6 \\ 1.59, & 6.21, & 3.31, & -2.86, & 4.93, & 3.69, & 5.21, & -1.53 \\ 17.4, & 44.1, & 25.9, & 33, & 3.69, & 17.7, & 7.79, & 6.71 \\ 14, & 15.8, & 38, & 14.1, & 5.21, & 7.79, & 23, & 9.07 \\ 15.2, & 17.4, & 28.8, & 28.6, & -1.53, & 6.71, & 9.07, & 13.9 \end{pmatrix}$$

$$m_1=(69.3,\ 61.9,\ 83.5,\ 70.8,\ 97.7,\ 91.5,\ 87.6,\ 82.4)$$

$$m_2=(59.2,\ 55.5,\ 81.9,\ 63.9,\ 95.1,\ 91,\ 91.1,\ 86.5)$$

$$S_W=\begin{pmatrix} 41.3, & 73.4, & 79.7, & 73.3, & 10.5, & 28.5, & 22.3, & 17.1 \\ 73.4, & 235, & 175, & 178, & 26.9, & 81.2, & 35.1, & 18.6 \\ 79.7, & 175, & 255, & 188, & 25.8, & 59.8, & 62.7, & 26.5 \\ 73.3, & 178, & 188, & 208, & 10.7, & 57.4, & 34.1, & 33 \\ 10.5, & 26.9, & 25.8, & 10.7, & 19.3, & 17.5, & 15.8, & 4.49 \\ 28.5, & 81.2, & 59.8, & 57.4, & 17.5, & 36.8, & 18.5, & 12.6 \\ 22.3, & 35.1, & 62.7, & 34.1, & 15.8, & 18.5, & 34.1, & 15.8 \\ 17.1, & 18.6, & 26.5, & 33, & 4.49, & 12.6, & 15.8, & 26.6 \end{pmatrix}$$

$m_1=(69.3,\ 61.9,\ 83.5,\ 70.8,\ 97.7,\ 91.5,\ 87.6,\ 82.4)$

$m_2=(59.2,\ 55.5,\ 81.9,\ 63.9,\ 95.1,\ 91,\ 91.1,\ 86.5)$

$$S_W^{-1}=\begin{pmatrix}
0.10, & -0.00, & -0.01, & -0.01, & 0.01, & -0.03, & -0.01, & -0.02 \\
-0.00, & 0.05, & -0.00, & -0.02, & 0.03, & -0.10, & -0.00, & 0.04 \\
-0.01, & -0.00, & 0.03, & -0.02, & -0.00, & 0.01, & -0.04, & 0.02 \\
-0.01, & -0.02, & -0.02, & 0.04, & 0.01, & 0.02, & 0.03, & -0.03 \\
0.01, & 0.03, & -0.00, & 0.01, & 0.19, & -0.15, & -0.07, & 0.04 \\
-0.03, & -0.10, & 0.01, & 0.02, & -0.15, & 0.31, & 0.03, & -0.08 \\
-0.01, & -0.00, & -0.04, & 0.03, & -0.07, & 0.03, & 0.13, & -0.06 \\
-0.02, & 0.04, & 0.02, & -0.03, & 0.04, & -0.08, & -0.06, & 0.11
\end{pmatrix}$$

m_1=(69.3, 61.9, 83.5, 70.8, 97.7, 91.5, 87.6, 82.4)

m_2=(59.2, 55.5, 81.9, 63.9, 95.1, 91, 91.1, 86.5)

m_1-m_2=(10.1, 6.4, 1.6, 6.9, 2.6, 0.5, −3.5, −4.1)

$u=S_W^{-1}(m_1-m_2)$ =

(1.09, 0.0284, −0.184, 0.0928, 0.835, −0.832, −0.376, −0.33)

$$\tilde{m}_1 = \mu \cdot m_1 = 13.7$$

$$\tilde{m}_2 = \mu \cdot m_2 = \text{-}2.35$$

y_t=5.67

100元60点y值:(3.62 — 20.5)

3.62,	7.3,	11.9,	14.4,	16.9,	19.0,	20.3,	20.5
19.8,	17.9,	16.8,	15.6,	15.8,	15.1,	15.1,	15.5
13.7,	11.6,	9.08,	9.57,	9.26,	9.78,	8.74,	10.3
11.0,	13.5,	15.1,	14.6,	15.1,	14.6,	16.1,	16.2
17.7,	16.1,	15.2,	13.3,	12.3,	11.9,	11.9,	12.7
14.3,	15.7,	16.5,	14.0,	12.8,	11.5,	10.1,	11.0
10.8,	11.0,	12.6,	13.5,	15.4,	15.4,	13.9,	13.5
13.8,	13.3,	13.6,	13.0,				

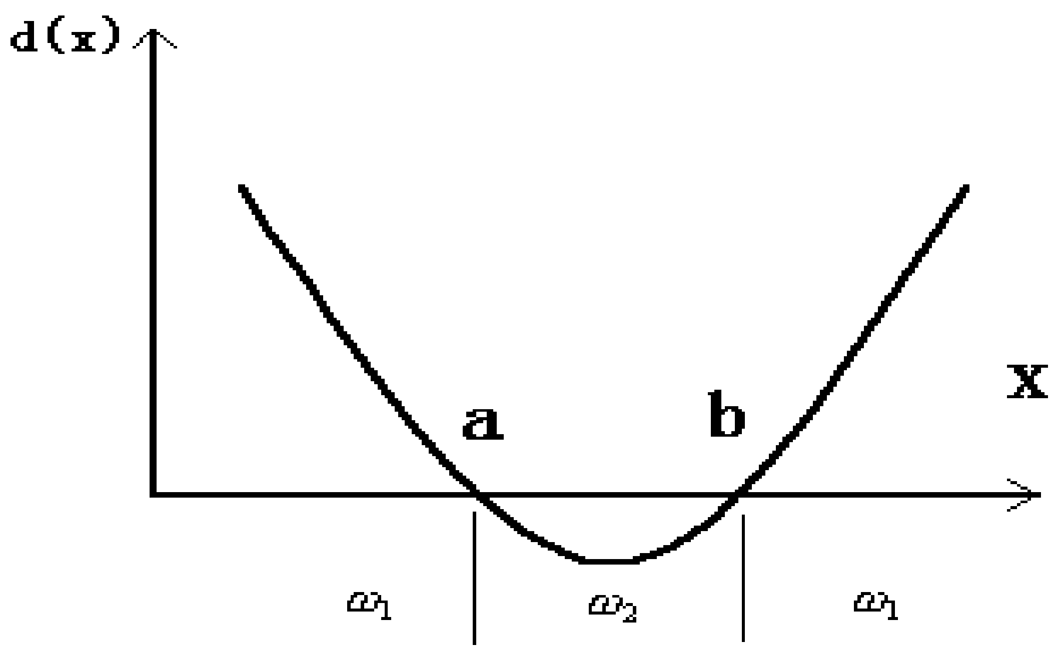
y_t=5.67

50元60点y值: (-9.72 — 3.4)

-1.69,	-0.57,	1.89,	3.4,	1.99,	1.4,	-1.71,	-2.29
-3.43,	-3.5,	-3.56,	-3.84,	-2.34,	0.144,	1.67,	1.99
-1.12,	-3.5,	-3.83,	-4.24,	-3.72,	-2.99,	-3.82,	-3.45
-2.61,	-1.82,	-2.75,	-2.94,	-2.82,	-3.05,	-3.24,	-2.1
-1.88,	-1.92,	-1.55,	-1.55,	-1.47,	-2.82,	-2.3,	-2.65
-3.04,	-4.75,	-3.25,	-5.5,	-7.78,	-9.72,	-9.18,	-5.42
-2.55,	0.413,	1.18,	1.23,	-1.99,	-2.48,	-2.31,	-1.86
-0.38,	-1.21,	-2.55,	-5.08,				

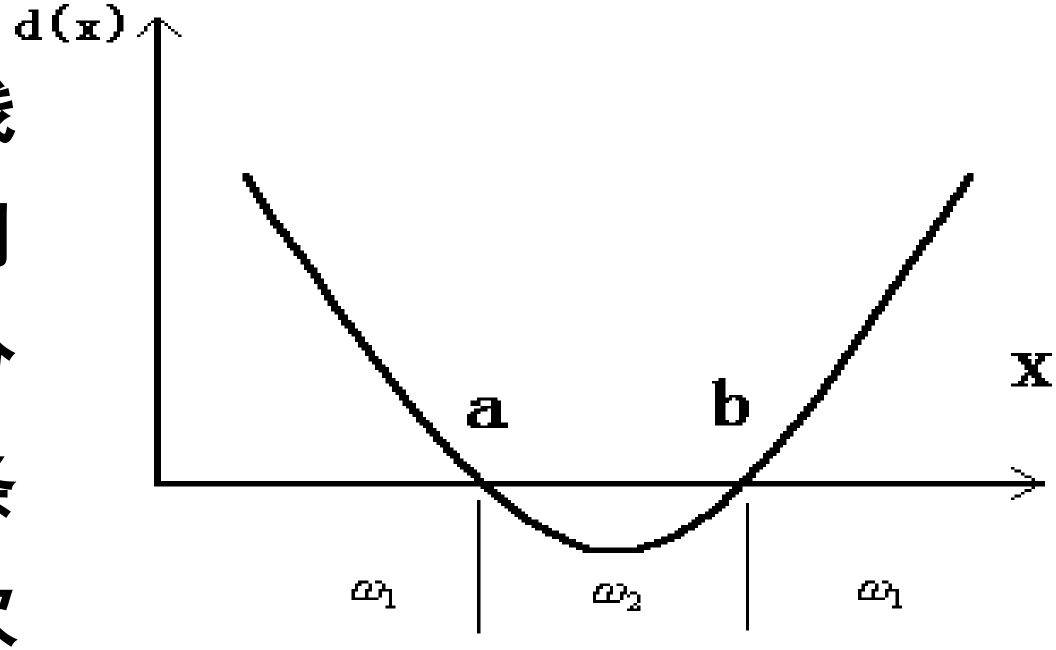
§2-5、广义线性判别函数

设一维两类模式 x 在一维空间——坐标轴上分布如图3-9-1所示，两类的类域为 $\Omega_1:(-\infty,a)$（b,∞）和 $\Omega_2:(a,b)$



一维特征空间中非线性可分的图示

显然，这两类不是线性可分的，因不能用一个分界点将两类分开。但是，对于一条穿过a、b两点的二次曲线，



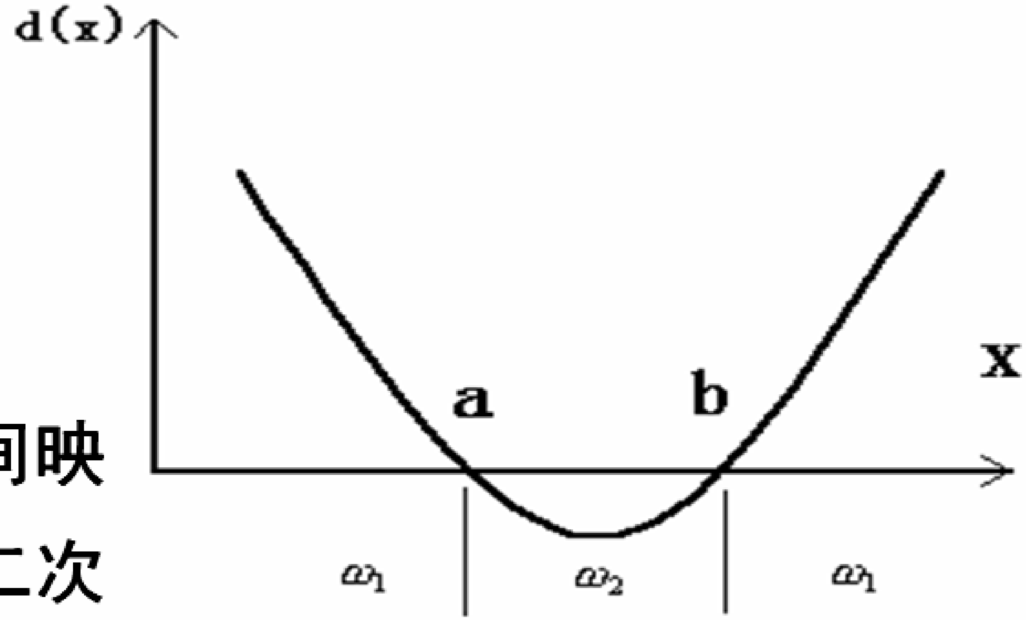
$$d(x)=(x-a)(x-b)=x^2-(a+b)x+ab$$

当 $x\in\omega_1$ 即 $x<a$ 或 $x>b$ 时，$d(x)>0$

当 $x\in\omega_2$ 即 $a<x<b$ 时，$d(x)<0$

如果作非线性变换：

$$\begin{cases} y_1 = x^2 \\ y_2 = x \end{cases}$$



则原来的一维特征空间映射为二维特征空间，二次判别函数 $d(x)$ 变换后在二维特征空间中映射成一次判别函数： $d(\vec{y}) = y_1 - (a+b)y_2 + ab$

原来的一维非线性可分的模式在所映射的二维特征空间中是线性可分的，即：

$$\begin{cases} \vec{y} \in \omega_1 & d(\vec{y}) > 0 \\ \vec{y} \in \omega_2 & d(\vec{y}) < 0 \end{cases}$$

将上述方法一般化：

设 n 维模式特征集 $\{\vec{x}_i\}$ 在特征空间 X^n 中是非线性可分的，对各模式作非线性变换

$$T: X^n \Rightarrow Y^d \quad d > n$$

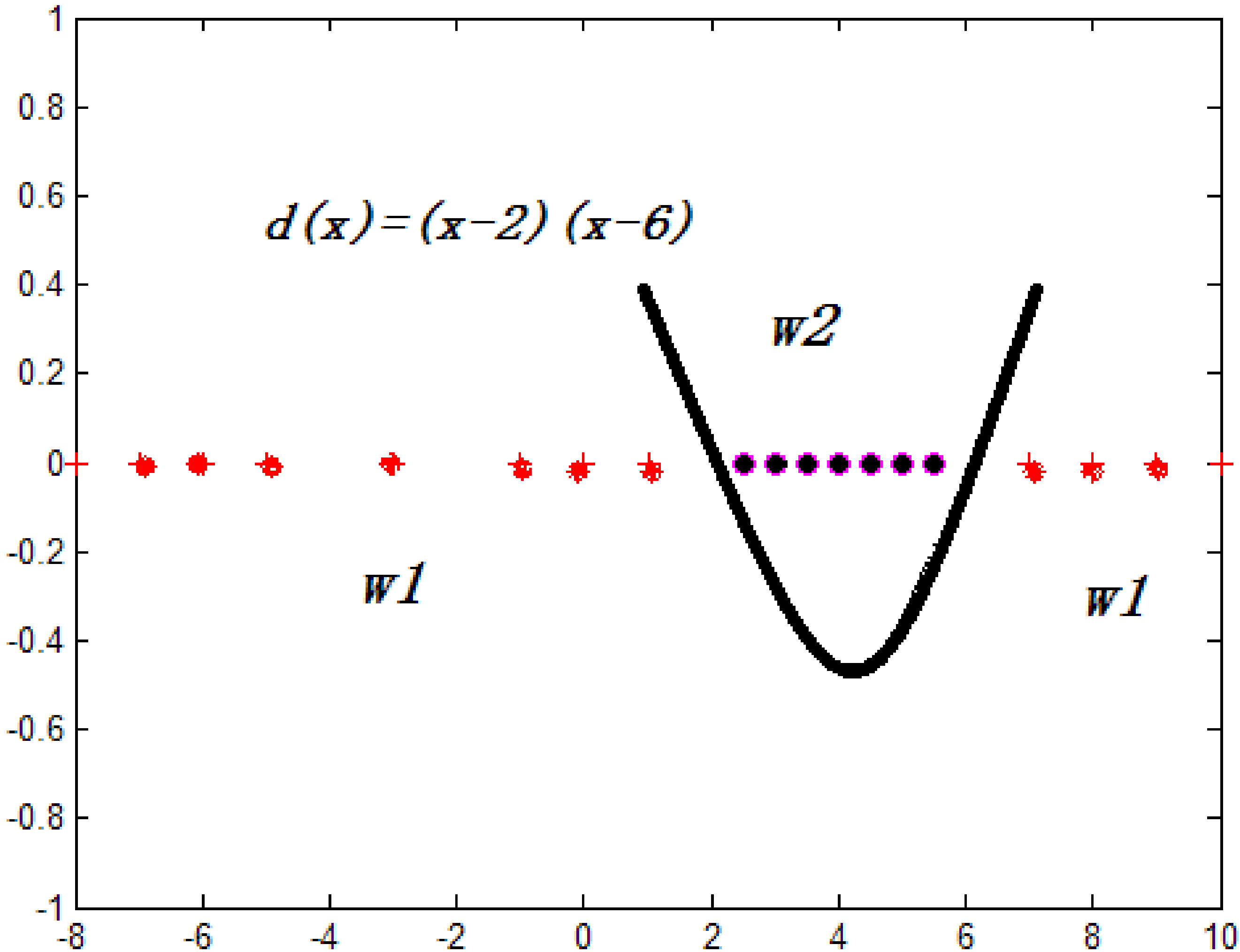
$$\vec{x} = (x_1, x_2, \cdots, x_n)' \quad \vec{y} = (y_1, y_2, \cdots, y_d)'$$

其中 $y_i = f_i(\vec{x})$ 是 $\vec{x}$ 的单值函数，选取适当的函数

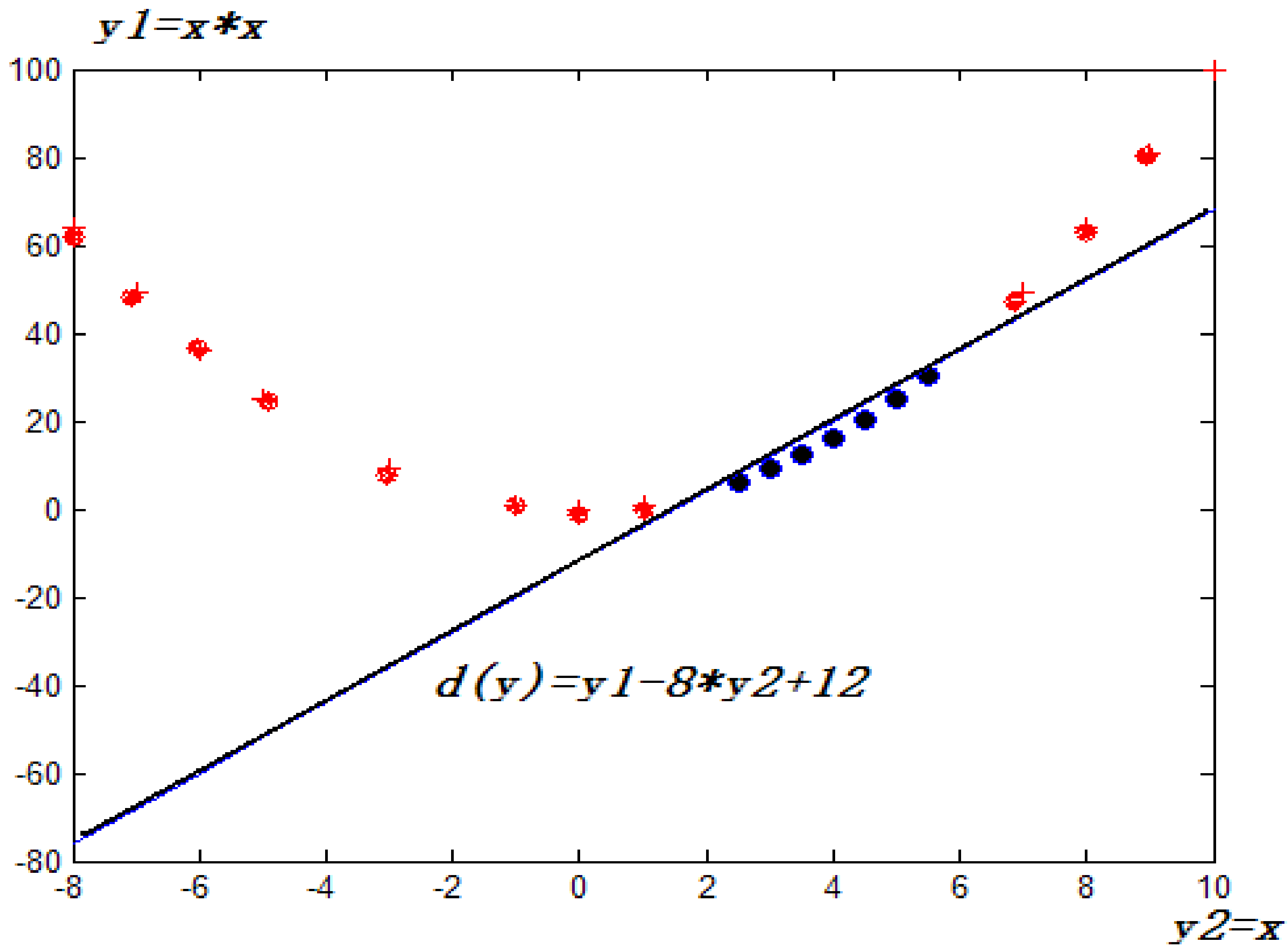
$$f_i(\vec{x}) \quad (i = 1, 2, \cdots, d)$$

使 $\{\vec{y}_j = (f_1(\vec{x}_j), f_2(\vec{x}_j), \cdots, f_d(\vec{x}_j))'\}$ 在特征空间是线性可分的，称其为**广义线性判别函数**。

下图所示两类模式是线性不可分的。



经过非线性变换，两类模式是线性可分的。



可以将 X^n 中非线性判别函数 $d(\vec{x})$ 的一般形式和 Y^d 中相应的线性判别函数 $\tilde{d}(\vec{y})$ 的关系表示如下：

$$d(\vec{x}) = w_1 f_1(\vec{x}) + w_2 f_2(\vec{x}) + \cdots + w_d f_d(\vec{x}) + w_{d+1}$$
$$= w_1 y_1 + w_2 y_2 + \cdots + w_d y_d + w_{d+1} \equiv \vec{w}'\vec{y} = \tilde{d}(\vec{y})$$

式中 $\vec{w} = (w_1, w_2, \cdots, w_{d+1})'$

$$\vec{y} = (y_1, y_2, \cdots, y_d, 1)' = (f_1(\vec{x}), f_2(\vec{x}), \cdots, f_d(\vec{x}), 1)'$$

$y_i = f_i(\vec{x})(i = 1, 2, \cdots, d)$ 是 $\vec{x}$ 的单值实函数。

下面讨论两种典型的 $f_i(\vec{x})$ 的形式

1、$d(\vec{x})$ 为二次多项式

此时 $d(\vec{x})$ 的一般形式为：

$$d(\vec{x})=\sum_{i=1}^{n} w_{ii}x_i^2+\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{ij}x_ix_j+\sum_{i=1}^{n} w_ix_i+w_{n+1}$$

取 $y_i=f_i(\vec{x})$ 为二次或一次式, 可使 $d(\vec{y})$ 成为线性函数。

下面讨论两种典型的 $f_i(\vec{x})$ 的形式

$$d(\vec{x})=\sum_{i=1}^{n}w_{ii}x_i^2+\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}w_{ij}x_ix_j+\sum_{i=1}^{n}w_ix_i+w_{n+1}$$

在非线性判决函数中，一种常见的情况是 $f_i(x)$ 取 x 的平方函数，且 x 是二维的，此时：

$g(x)=w_{11}x_1^2+w_{12}x_1x_2+w_{22}x_2^2+w_1x_1+w_2x_2+w_3$ 为二次多项式。

令 $y_1=x_1^2$， $y_2=x_1x_2$， $y_3=x_2^2$， $y_4=x_1$， $y_5=x_2$

则： $g(x)\Leftrightarrow g(y)=w_{11}y_1+w_{12}y_2+w_{22}y_3+w_1y_4+w_2y_5+w_3$ 。

也就是：二维二次多项式判决函数变成五维线性判决函数。

当非判决函数为线性判决之后，其中的参数 w 可以用感知器、H—K、Fisher等算法来计算。

下面讨论两种典型的 $f_i(\vec{x})$ 的形式

$$d(\vec{x}) = \sum_{i=1}^{n} w_{ii} x_i^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} x_i x_j + \sum_{i=1}^{n} w_i x_i + w_{n+1}$$

上式右边前两项是x各分量的二次项求和式，显然它们的项数分别为n和 n(n-1)/2。

第三项是x的各分量一次项求和式，项数为n。

所以d(x)的项数为：

n+n(n-1)/2+n+1=(n+1)(n+2)/2

由于有一个常数项，变换后的特征空间维数为：

(n+1)(n+2)/2-1=n(n+3)/2

2 $d(\vec{x})$ 为 r 次多项式

此时 $d(\vec{x})$ 的一般形式为：

$$d(\vec{x}) = d^r(\vec{x}) + d^{r-1}(\vec{x}) + \cdots + d^1(\vec{x}) + d^0(\vec{x}) \triangleq d^{(r)}(\vec{x})$$

$$d^0(\vec{x}) = w_{n+1} = d^{(0)}(\vec{x})$$

$$d^1(\vec{x}) = \sum_{i=1}^{n} w_i x_i$$

$$d^2(\vec{x}) = \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} w_{i_1 i_2} x_{i_1} x_{i_2}$$

$$d^r(\vec{x}) = \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} \cdots \sum_{i_r=i_{r-1}}^{n} w_{i_1 i_2 \cdots i_r} x_{i_1} x_{i_2} \cdots x_{i_r}$$

1
C_n^1
C_{n+1}^2
……
C_{n+r-1}^r

令 $f_i(\vec{x}) = x_1^{s_1} x_2^{s_2} \cdots x_n^{s_n}$

其中 $s_1 + s_2 + \cdots + s_n \le r$

$\vec{y}$ 的维数 $d = (n+r)!/n!r! - 1$

非线性判别函数 $d(\vec{x})$ 变换成线性判别函数 $d(\vec{y})$ 后，特征空间的维数由原来的n维增长到d维（d>n），增大了二分能力。

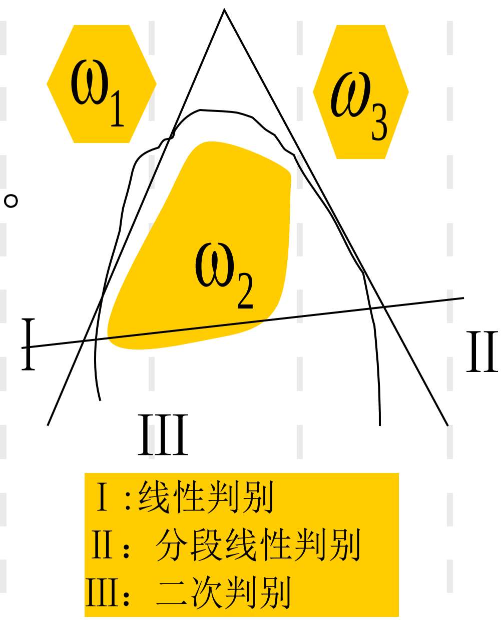
§ 2-6、非线性判别函数

- 1.**分段线性判别函数**（用线性无法分开,可用分段线性判别函数）

①、**基于距离的分段线性判别函数。**（用均值代表一类，通过均值连线中点的垂直线分开）

把ω $_i$类可以分成l_i个子类：

$$\therefore\ \omega_i = (\omega_i^1, \omega_i^2, \ldots, \omega_i^l)$$ 分成l个子类。



- 现在定义子类判别函数：

$$g_i(x) = \min_{l=1,2,\ldots,l} \left\| x - \mu_i^l \right\|$$

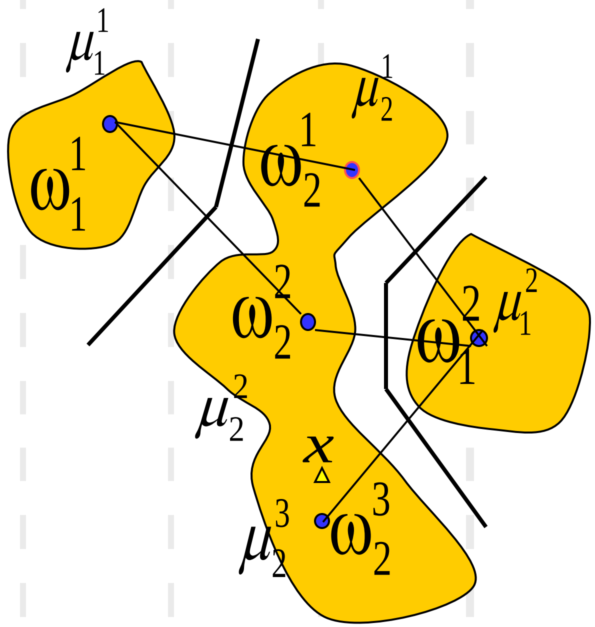
- 在同类的子类中找最近的均值。
- 判别规则：

$$g_j(x) = \min g_i(x), i = 1,2,\ldots, M$$

- 这是在M类中找最近均值。则把x归于ω_j类完成分类。

1、分段线性判别函数

- 例：未知x,如图：
- 先与ω_1类各子类的均值比较，即 $\|x-\mu_1^l\|$ ，找一个最近的 $g_1(x)=\|x-\mu_1^2\|$ 与ω_2各子类均值比较取最近的 $g_2(x)=\|x-\mu_2^3\|$ 因$g_2(x)< g_1(x)$，所以$x\in\omega_2$类。



1、分段线性判别函数（续）

②、基于函数的分段线性判别函数

利用均值代表一类有时有局限性，如图所示。

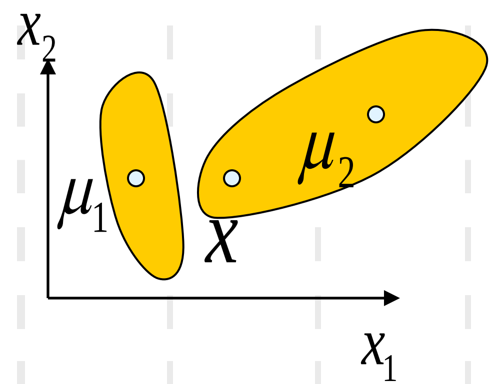
若用 线性判别函数代表一类，就会克服上述情况。

- 设$\omega = \omega_1, \omega_2, \ldots\ldots \omega_m$
- 而每一类又可以分为$\omega_i = (\omega_i^1, \omega_i^2, \ldots\ldots \omega_i^l)$ 子类。
- 对每个子类定义一个线性判别函数为：

$$g_i^l(x) = w_i^l x,\text{其中}w_i^l\text{为}\omega_i^l\text{子类的权向量。}$$

- 则定义ω_i类的线性判别函数为：

$$g_i(x) = \max_{l=1,2,\ldots\ldots,l} g_i^l(x)$$



1、分段线性判别函数（续）

- 在各子类中找最大的判别函数作为此类的代表，则对于M类，可定义M个判别函数$g_i(x), i=1,2,\ldots\ldots M$，因此，决策规则：

$$g_j(x)=\max_{i=1,2,\ldots\ldots,M} g_i(x), 则 x\in\omega_j$$

- 对未知模式x，把x先代入每类的各子类的判别函数中，找出一个最大的子类判别函数，M类有M个最大子类判别函数，在M个子类最大判别函数中，再找一个最大的，则x就属于最大的子类判别函数所属的那一类。

1、分段线性判别函数（续）

③、 基于凹函数的并分段线性判别函数（针对多峰情况）

设l_i子类判别函数，i=1,2,…..r则分段线性判别函数有如下特性：

- (a)：$l_1,l_2,\ldots\ldots l_r$都是分段线性判别函数
- (b)：若A,B都是分段线性判别函数，则： $A\wedge B$ ， $A\vee B$ 也是分段线性判别函数。 $A\wedge B$取最小， $A\vee B$取最大。
- (c)：对任何分段线性函数都可以表示成如下二种形式：

- 1)、析取范式(这是经常采用的形式)

$P=(L_{11}\wedge L_{12}\wedge\ldots\wedge L_{1m})\vee\ldots\vee(L_{q1}\wedge L_{q2}\wedge\ldots\wedge L_{qm})$

- 2)、合取范式

$Q=(L_{11}\vee L_{12}\vee\ldots\vee L_{1m})\wedge\ldots\wedge(L_{q1}\vee L_{q2}\vee\ldots\vee L_{qm})$

- 每个$(L_{11}\vee L_{12}\vee\ldots\vee L_{1m})$ 都称为凹函数。

1、分段线性判别函数（续）

- 对于多峰二类问题：设第一类有q个峰，则有q个凹函数。
- 即$P=P_1 \vee P_2 \vee \ldots\ldots \vee P_q$
- 每个凹函数P_i由m 个线性判别函数来构成。
- $\therefore P_i=L_{i1} \wedge L_{i2} \wedge \ldots \wedge L_{im}$
- 假设对于每个子类线性判别函数L_{ij}都设计成：

$$L_{ij} = w_{ij}x \begin{cases} > 0, x \in \omega_1, i = 1,2,\ldots, q \text{子类。} \\ < 0, x \in \omega_2, j = 1,2,\ldots, m \text{每个子类的判别函数数。} \end{cases}$$

$$\text{判别规则:} \begin{cases} P > 0, \text{则} x \in \omega_1 \\ P \le 0, \text{则} x \in \omega_2 \end{cases}$$

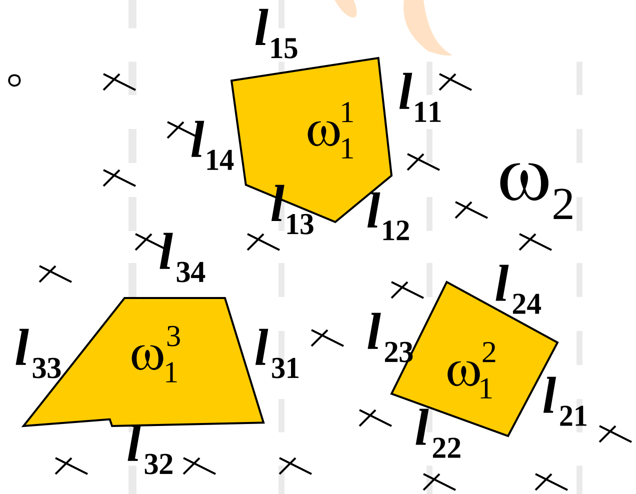
1、分段线性判别函数（续）

- 例、设如图

ω_1分三个峰，$q=3$这样它有三个子类。

判别函数个数：$\begin{cases} m_1 = 5 \\ m_2 = 4 \\ m_3 = 4 \end{cases}$

$\therefore$有13个分段判别函数



$\therefore P=(L_{11} \wedge L_{12} \wedge L_{13} \wedge L_{14} \wedge L_{15}) \vee (L_{21} \wedge L_{22} \wedge L_{23} \wedge L_{24}) \vee (L_{31} \wedge L_{32} \wedge L_{33} \wedge L_{34})$

$\therefore P = \max\{\min(l_{11}, l_{12}, ..., l_{15}), \min(l_{21}, ..., l_{24}), \min(l_{31}, ..., l_{34})\}$

若$P > 0$,则$x \in \omega_1$。若$P \le 0$,则$x \in \omega_2$。

2、二次判别函数

❖二次判别函数一般可表示成：

$$g(x) = X^T \overline{W} X + W^T X + W_0$$

$$= \sum_{i=1}^{n} w_{ii} x_i^2 + 2\sum_{j=1}^{n-1} \sum_{i=j+1} w_{ji} x_j x_i + \sum_{j=1}^{n} w_j x_j + W_0$$

其中：$\overline{W}$是$n \times n$维的权向量。

W为n维权向量

$g(x)$的系数一共有$l = \frac{1}{2} n(n+3) + 1$(非常复杂,计算量很大）

二次决策面为超二次曲 面。(超球面，超双曲面等）

(1)若已知样本ω_1分布比较集中，形成单 峰，ω_2分布分散。如下图：

定义ω_1判别函数：

$g(x) = k^2 - (x - \overline{\mu}_1)^T \sum_1^{-1} (x - \overline{\mu}_1)$,$k$的大小，决定超平面的 大小。

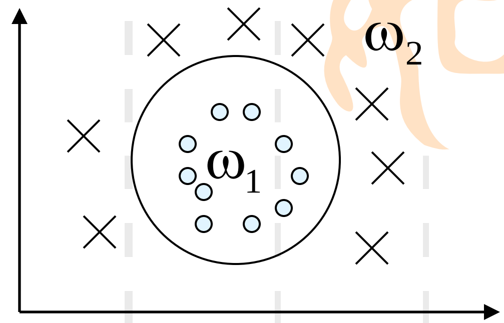
其中：$\overline{\mu}_1$为ω_1均值,$\sum_1$为ω_1协方差

2、二次判别函数(续)

判别规则:$\begin{cases} g(x) > 0, x \in \omega_1 \\ g(x) < 0, x \in \omega_2 \end{cases}$

判别平面：$g_1(x) = 0$是个超球面

由k控制大小



ω_1类比较集中

（2）如果ω_1，ω_2都比较集中，那么定义两个判别函数：

$$g_i(x) = k_i^2 - (x - \overline{\mu}_i)^T \sum\nolimits_i^{-1} (x - \overline{\mu}_i), \quad i = 1,2$$

其中：$\overline{\mu}_i$为ω_1，ω_2均值，$\sum_i$为ω_1，ω_2协方差

➢2、二次判别函数(续)

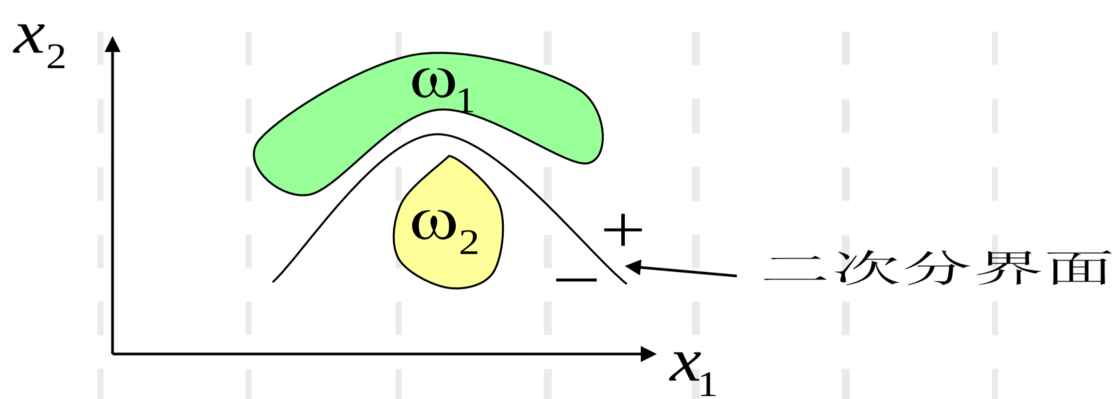
判别平面方程：$g(x)=g_1(x)-g_2(x)$

$$=-x^2(\sum_1^{-1}-\sum_2^{-1})x+2(\overline{\mu}_1^T\sum_1^{-1}-\overline{\mu}_2^T\sum_2^{-1})x-$$

$$(\overline{\mu}_1^T\sum_1^{-1}\overline{\mu}_1-\overline{\mu}_2^T\sum_2^{-1}\overline{\mu}_2)+(k_1^2-k_2^2)=0$$

判别规则：$g(x)\begin{cases}>0, x\in\omega_1\\<0, x\in\omega_2\end{cases}$

k_1,k_2可用来调整二类错误率。



❖关于二次判别函数，我们将在贝叶斯分类器中详细论述。

3、势函数法

势函数法的特点就是可以直接确定判决函数，它不但适用于线性分类器，而且还适用于非线性分类器。这里着重介绍势函数法，以示范一个非线性分类器。

势函数法借用电场的概念，来解决模式的分类问题。在这种算法中，把属于w_1类的样本点看作正电荷，而属于w_2类的样本点看作负电荷，从而把模式的分类转变为正、负电荷的转移的问题，电位为0的等位线（面）就是判别界线。

3、势函数法（续）

势函数的选择

令 $K(x,x_k)$ 表示 x 与 x_k 之间的电位势函数，则 $K(x,x_k)$ 的选择应满足以下三个条件：

① $K(x,x_k)=K(x_k,x)$，当 $x=x_k$ 时，$K(x,x_k)$ 达到最大值；

② 当矢量 x 与 x_k 之间的距离趋于无穷大时，$K(x,x_k)=0$；

③ $K(x,x_k)$ 是光滑函数，且是 x 与 x_k 之间距离的单调下降函数。

所有样本点 x_k（$k=1,2,\ldots\ldots,$）的电位势函数 $K(x,x_k)$ 在任一点 x 产生的电位总和被称为积累电位势函数，表示为 $K(x)$。如果 $K(x)$ 能够把样本正确分类，可取为判决函数 $g(x)$，即：$g(x)=K(x)$。

下面介绍2类满足要求的电位势函数。

第一类势函数：

$$K(x,x_k)=\sum_{i=1}^{\infty}\varphi_i(x)\varphi_i(x_k)$$

它是对称的有限项多项式。式中，$\{\varphi_i(x), i=1,2,....\}$ 在 x 的定义域里是正交函数系。

第二类势函数：

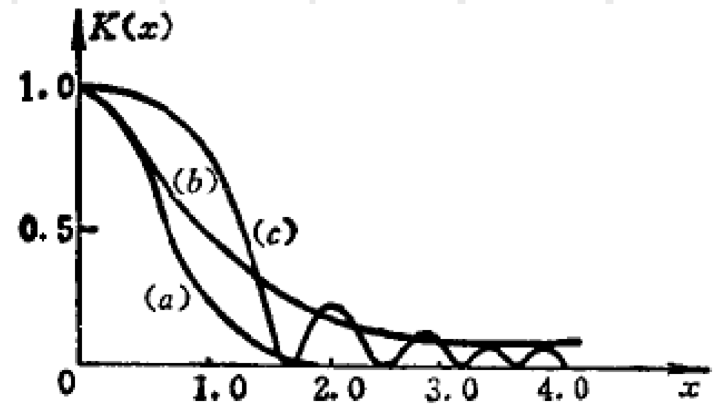
选择双变量 x 和 x_k 的对称函数，即 $K(x,x_k)=K(x_k,x)$，且可以展开成无穷级数，如：

$$K(x,x_k)=\exp(-\alpha|x-x_k|^2) \qquad (4.6\text{-}1)$$

$$K(x,x_k)=\frac{1}{1+\alpha|x-x_k|^2} \qquad (4.6\text{-}2)$$

$$K(x,x_k)=\left|\frac{\sin\alpha|x-x_k|^2}{\alpha|x-x_k|^2}\right| \qquad (4.6\text{-}3)$$

上述3个函数都是自原点开始随距离下降的分布，因为$|x-x_k|^2$实际上是欧式距离。对于(4.6-3)式，含有正弦函数，具有振荡的特点，因此，只有第一个振荡周期，才是可用的范围。如下图所示：



(a) 式(4.6-1)曲线 (b) 式(4.6-2)曲线 (c) 式(4.6-3)曲线

上图为一维情况下的曲线，对于二维情况，它们是以原点为中心的同心园状分布。

§ 2-7 非线性分类器的设计

一、分段线性分类器的设计

先求子类的权向量W_i^l，再求总的权向量W_i

1. 已知子类划分时的设计方法

把每一个子类作为独立类，利用每个子类的训练样本，求每个子类的线性判别函数，总的判别函数就可获得。

子类的划分可用以下方法：

① 用先验知识直接划分

② 用聚类分析，聚成多个子类

2. 已知子类的数目的设计方法

① 设各个子类的初始权向量：$W_i^1, W_i^2 \ldots W_i^{li}$

$i = 1,2,\ldots M$ W_i中有L_i个子类

② 若第K步迭代时ω_j类样本X_j同ω_j类某个子类的权向量$W_j^n(k)$的内积值最大，

即$W_j^n(k)^l x_j = \max\{ W_j^n(k)^l x_j \}$ $n = 1,2,\ldots l_j$

并且满足条件$W_j^{n}(k)\,x_j > W_i^{n}(k)^l x_j$

$i=1,2,\ldots M$类　　$j=1,2,\ldots l_i$子类　$i\neq j$

则权向量$W_i^{1}(k)$，$W_i^{2}(k)$，… ,$W_i^{l_i}(k)$不影响分类，所以权向量不需要修正。

若有某个或某几个子类不满足条件即：

存在$W_i^{n}(k)$使$W_j^{n}(k)\,x_j \le W_i^{n}(k)^l x_j \quad i\neq j$

所以x_j错分类，要修改权向量。

设$W_i^{n}(k)^l x_j = \max\{ W_i^{n}(k)^l x_j \} \quad n = 1,2,\ldots l_i \quad i\neq j$

则修改权向量$W_j^{n}(k+1)= W_j^{n}(k) \pm \rho_k x_j$

③　重复以上迭代,直到收敛,此法类似于固定增量法.

3.未知子类数目时的设计方法

当每类应分成的子类数也不知时，这是最一般情况，方法很多，举例如下。

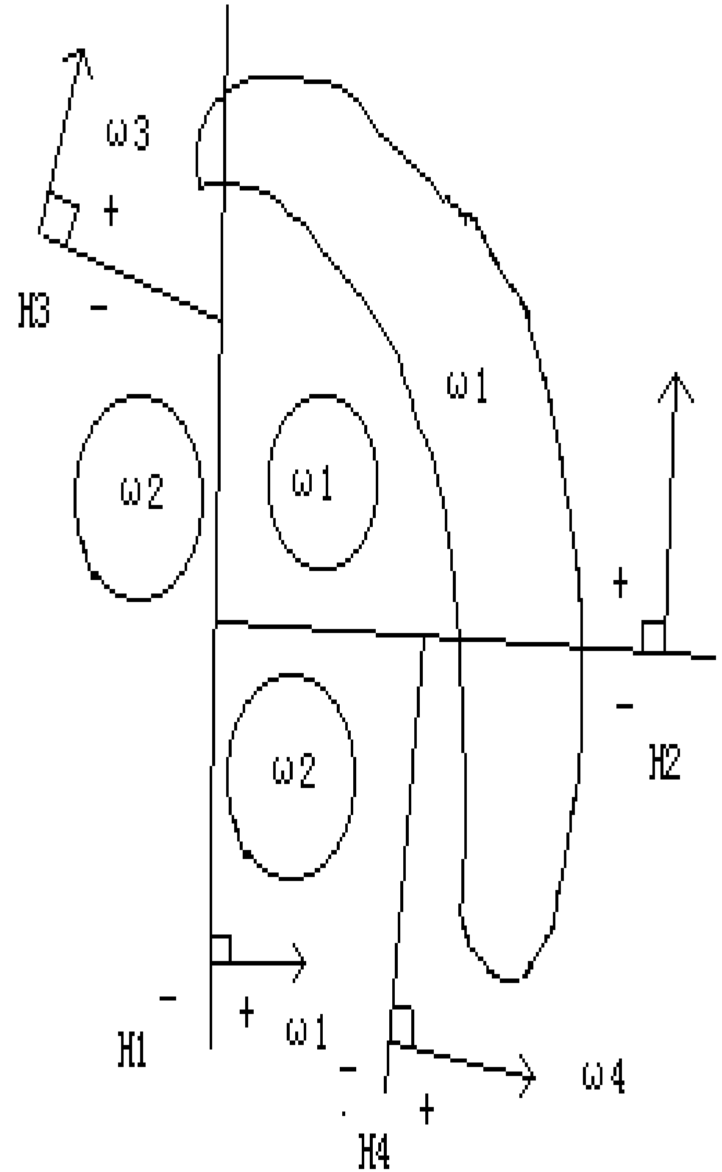
树状分段线性分类器:

设两类情况ω_1，ω_2。如图所示

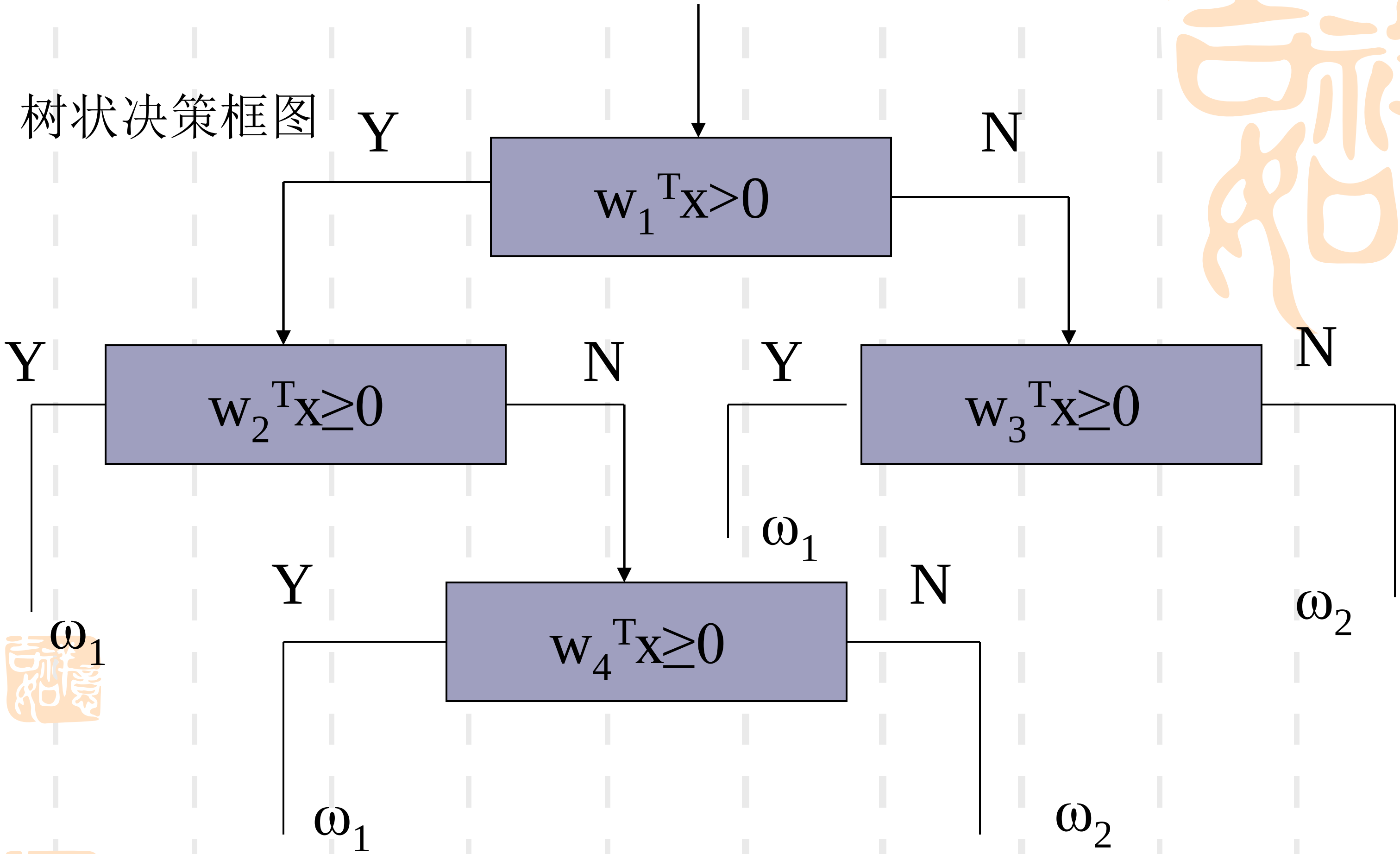
① 先用两类线性判别函数求出W_1,超平面H_1分成两个区间,每个区间包含两类。

②再利用二类分类求出$W_2(H_2)$ $W_3(H_3)$。

③ 如果每个部分仍包含两类,继续上面的过程。



树状决策框图



- 关键是初始权向量W_1的选择:一般先选两类中距离最近的两个子类的均值连线做垂直线作为$H_1(w_1)$初始值再求最优解。

二、 势函数分类器的设计

势函数算法的训练过程，是利用势函数在逐个样本输入时，逐步积累电位的过程。对于两类问题来说，如势积累方程能以其运算结果的正负来区别两类样本，则训练结束。

势函数算法过程为：

设初始电位为 $K_0(x)=0$。

第一步：输入样本 x_1，计算其积累电位 $K_1(x)$

$$K_1(x)=\begin{cases} K_0(x)+K(x,x_1), & \text{若 } x_1 \in w_1 \\ K_0(x)-K(x,x_1) & \phantom{\text{若 }} x_1 \in w_2 \end{cases}$$

$K_1(x)$ 描述了加入第一个样本后的边界划分，当样本属于 w_1。势函数为正，当样本属于 w_2，势函数为负。

第二步：加入第二个训练样本 x_2，则有三种情况：

① 若 $x_2 \in w_1$，且 $K_1(x_2) > 0$，或 $x_2 \in w_2$ 且 $K_1(x_2) < 0$，表示分类正确, 则势函数不变, 即： $K_1(x) = K_2(x)$ 。

② 若 $x_2 \in w_1$, 但 $K_1(x_2) \le 0$, 则修改势函数: $K_2(x) = K_1(x) + K(x, x_2)$

③ 若 $x_2 \in w_2$, 但 $K_1(x_2) \ge 0$, 则也应修改势函数： $K_2(x) = K_1(x) - K(x, x_2)$

由分析可知，②，③两情况属于错分，必须修改势函数。将上述情况推广到第 $(k+1)$ 步，此时已输入了 k 个样本，进到第 $(k+1)$ 次迭代，此时的积累势函数也会有三种情况：

① 若 $x_{k+1} \in w_1$, 且 $K_k(x_{k+1}) > 0$ 或 $x_{k+1} \in w_2$, 且 $K_k(x_{k+1}) < 0$, 则:

$K_{k+1}(x) = K_k(x)$ 。

② 若 $x_{k+1} \in w_1$, 且 $K_k(x_{k+1}) \le 0$， 则： $K_{k+1}(x) = K_k(x) + K(x, x_{k+1})$

③ 若 $x_{k+1} \in w_2$, 且 $K_k(x_{k+1}) \ge 0$， 则： $K_{k+1}(x) = K_k(x) - K(x, x_{k+1})$

上述三种情况，可归纳成一个方程：

$$K_{k+1}(x)=K_k(x)+r_{k+1}\cdot K(x,x_{k+1})$$

其中 r_{k+1} 的取值为：

当 $x_{k+1}\in w_1$， 且 $K_k(x_{k+1})>0$ 时， $r_{k+1}=0$ 。

当 $x_{k+1}\in w_2$， 且 $K_k(x_{k+1})<0$ 时， $r_{k+1}=0$ 。

当 $x_{k+1}\in w_1$， 且 $K_k(x_{k+1})\le 0$ 时， $r_{k+1}=1$ 。

当 $x_{k+1}\in w_2$， 且 $K_k(x_{k+1})\ge 0$ 时， $r_{k+1}=-1$ 。

如果从所给的样本集中 $\{x_1,x_2,\ldots\ldots,x_k,\ldots\ldots\}$ 省略那些并不使积累势函数发生变化的样本，则可得到一简化了的训练样本系列 $\{x_1,x_2,\ldots,x_j,\ldots\}$，它们完全是校正错误的模式样本，因此势函数的修改可归纳为：

$$K_{k+1}(x)=\sum_{x_j}\alpha_j K(x,x_j)$$

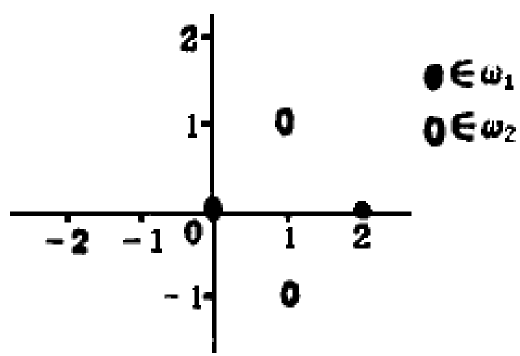
式中：$\alpha_j=\begin{cases}1, & 对于x_j\in w_1\\ -1, & 对于x_j\in w_2\end{cases}$

因此，由 $k+1$ 个训练样本产生的积累势函数，等于 w_1 类和 w_2 类两者中，校正错误样本的总位势之差。

积累势函数不必作任何修改就可以用作判别函数。对于一个二类问题，其判别函数：

$$g(x)=K(x)$$

例. 设一个两类训练样本：$w_1=\{(0,0)^T,(2,0)^T\}$，$w_2=\{(1,1)^T,(1,-1)^T\}$ 样本分布如下图所示：



解：如样本分布图所示，这两类样本是非线性可分的。因为特征向量是二维的，势函数为：$K(x,x_k)=\exp\{[(x_1-x_{k_1})^2+(x_2-x_{k_2})^2]\}$，取 $\alpha=1$，$K_0(x)=0$， $k=0$。

第一步：输入 $x_1=(0,0)^T$， $x_1\in w_1$， $K_1(x)=K(x,x_1)=\exp[-(x_1^2+x_2^2)]$

第二步：输入 $x_2=(2,0)^T$， $x_2\in w_1$， $K_1(x_2)=e^{-4}>0$ 分类正确, 不修正

$$K_2(x)=K_1(x)=\exp[-(x_1^2+x_2^2)]$$

第三步：输入 $x_3=(1,1)^T$， $x_3\in w_2$， $K_1(x_3)=e^{-2}\nless 0$，分类错，修正

$$K_3(x)=K_2(x)-K(x,x_3)=\exp[-(x_1^2+x_2^2)]-\exp\{-[(x_1-1)^2+(x_2-1)^2]\}$$

第四步：输入 $x_4=(1,-1)^T$， $x_4\in w_2$， $K_3(x_4)=e^{-2}-e^{-4}\nless 0$，分类错，修正

$$K_4(x)=K_3(x)-K(x,x_4)$$

$$=\exp[-(x_1^2+x_2^2)]-\exp\{-[(x_1-1)^2+(x_2-1)^2]\}-\exp\{-[(x_1-1)^2+(x_2+1)^2]\}$$

第五步：再次输入 $x_1=(0,0)^T$，$x_1\in w_1$，$K_4(x_1)=e^0-e^{-2}-e^{-2}>0$，不修正

$$K_5(x)=K_4(x)$$

第六步：再次输入 $x_2=(2,0)^T$，$x_2\in w_1$，$K_5(x_2)=e^{-4}-2e^{-2}\ngtr 0$，修正

$$K_6(x)=K_5(x)+K(x,x_2)$$

$$=\exp[-(x_1^2+x_2^2)]-\exp\{-[(x_1-1)^2+(x_2-1)^2]\}-\exp\{-[(x_1-1)^2+(x_2+1)^2]\}$$

$$+\exp\{[(x_1-2)^2+x_2^2]\}$$

第七步：再次输入 $x_3=(1,1)^T$，$x_3\in w_2$，$K_6(x_3)=2e^{-2}-1-e^{-4}<0$，不修正

$$K_7(x)=K_6(x)$$

第八步：再次输入 $x_4=(1,-1)^T$，$x_4\in w_2$，$K_7(x_4)=e^{-2}-e^{-2}-1+e^{-4}<0$，不修正：$K_8(x)=K_7(x)$。

第九步：重新输入 $x_1=(0,0)^T$， $x_1\in w_1$， $K_8(x_1)=e^0-e^{-2}-e^{-2}+e^{-4}>0$， 不修正： $K_9(x)=K_8(x)$。

第十步：重新输入 $x_2=(2,0)^T$， $x_2\in w_1$， $K_9(x_2)=e^{-4}-e^{-2}-e^{-2}+e^0>0$， 不修正： $K_{10}(x)=K_9(x)$。

至此，所有训练样本皆被正确分类，分类器设计完毕。

判决函数：

$$
\begin{aligned}
g(x)&=K_{10}(x)\\
&=\exp[-(x_1^2+x_2^2)]-\exp\{-[(x_1-1)^2+(x_2-1)^2]\}-\exp\{-[(x_1-1)^2+(x_2+1)^2]\}\\
&\quad+\exp\{[(x_1-2)^2+x_2^2]\}
\end{aligned}
$$

势函数迭代公式，具有很强的分类能力，但修正次数增多，势函数方程的项数将会增加，从而使计算机的计算量大增。

$$d(\vec{x}) = K_{10}(\vec{x})$$

$$d(\vec{x}) = \exp[-(x_1^2 + x_2^2)] - \exp[-((x_1 - 1)^2 + (x_2 - 1)^2)]$$

$$-\exp[-((x_1 - 1)^2 + (x_2 + 1)^2)] + \exp[-((x_1 - 2)^2 + x_2^2)]$$

令 $d(\vec{x}) = 0$

$$\exp[-(x_1^2 + x_2^2)] - \exp[-((x_1 - 1)^2 + (x_2 - 1)^2)] -$$

$$\exp[-((x_1 - 1)^2 + (x_2 + 1)^2)] + \exp[-((x_1 - 2)^2 + x_2{}^2)]$$

$$= \exp[-(x_1^2 + x_2^2)]\{1 - \exp[-2(1 - x_1 - x_2)] -$$

$$\exp[-2(1 - x_1 + x_2)] + \exp[-4(1 - x_1)]\}$$

$$= \exp[-(x_1^2 + x_2^2)]\exp[-2(1 - x_1)]\{\exp[2(1 - x_1)] +$$

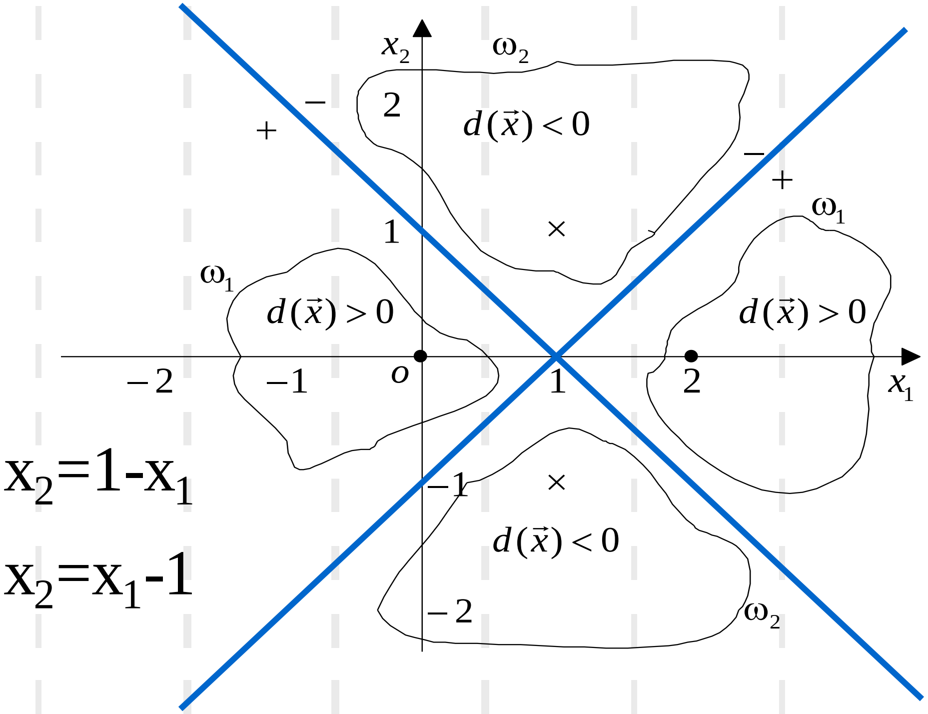
$$\exp[-2(1 - x_1)] - \exp[2x_2] - \exp[-2x_2]\} = 0$$

$$\because \exp[-(x_1{}^2 + x_2{}^2)] \neq 0, \exp[-2(1 - x_1)] \neq 0$$

$$\therefore \exp[2x_2] + \exp[-2x_2] = \exp[2(1 - x_1)] + \exp[-2(1 - x_1)]$$

$$\begin{cases} 2x_2 = 2 - 2x_1 \\ 2x_2 = -(2 - 2x_1) \end{cases} \Rightarrow \begin{cases} x_2 = 1 - x_1 \\ x_2 = x_1 - 1 \end{cases}$$

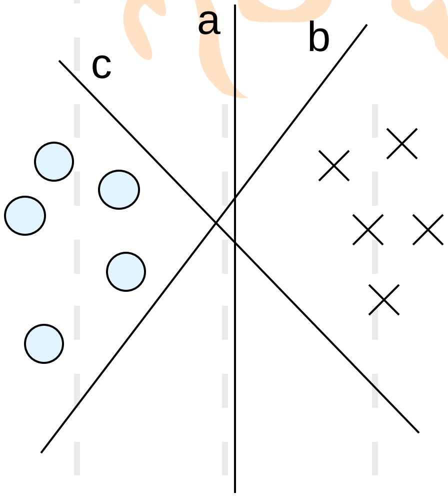
即 $ch[2x_2] = ch[2-2x_1]$，(双曲余弦函数：$ch(x) = \cosh(x) = \dfrac{e^x + e^{-x}}{2}$)



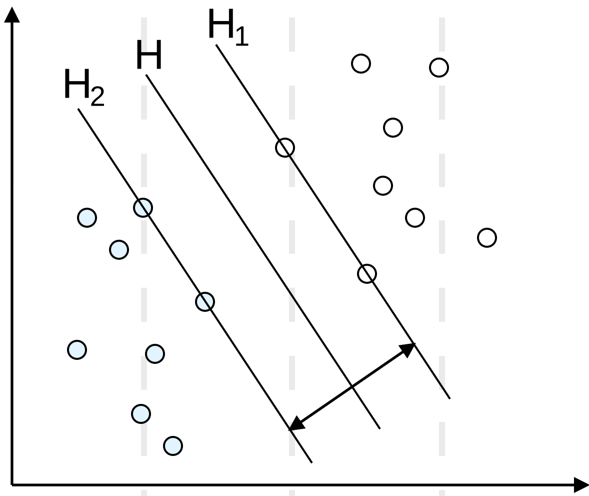
§2-8、最优分类超平面

- 线性可分情况

先考虑二维情况下的线性可分的两类样本(○, ×), 如图所示, 存在很多条可能的分类线能够将训练样本分开。显然分类线a最好, 因为它更远离每一类样本, 风险小。而其他的分类线离样本较近, 只要样本有较小的变化, 将会导致错误的分类结果。因此分类线a是代表一个最优的线性分类器。



所谓最优分类线就是要求分类线不但能将两类无误地分开, 而且要使两类的分类间隔最大。



图中H是最优分类线, H1和H2分别为过各类样本中离分类线最近的点且平行于分类线的直线, H1和H2之间的距离叫做两类的分类空隙或者分类间隔(margin)。

将二维推广到高维, 最优分类线就成为最优分类超平面。

2.8、最优分类超平面

- 线性可分情况（续）

设线性可分样本集为(x_i, y_i), $i=1, 2, \ldots n$, $x \in R^d$, $y \in \{+1, -1\}$是类别号。d维空间中线性判别函数的一般形式为$g(x)=w \bullet x+b$，则分类超平面方程为：

$$w \bullet x+b=0$$

其中，w为分类超平面的法线，是可调的权值向量；b为偏置，决定相对原点的位置。当两类样本是线性可分时，满足条件：

$$(w \bullet x_i)+b \geqslant +1 \qquad y_i=+1$$

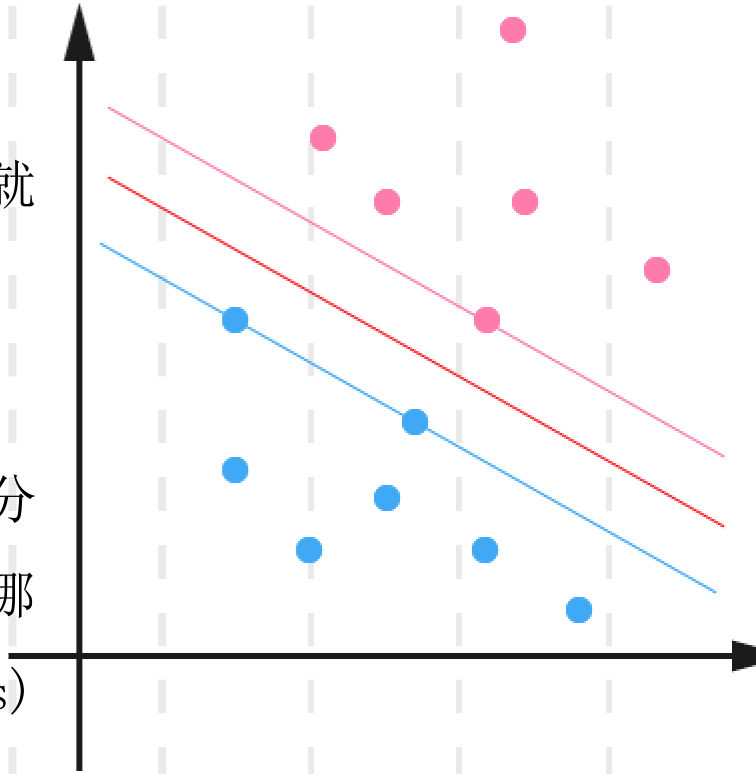
$$(w \bullet x_i)+b \leqslant -1 \qquad y_i=-1$$

超平面$(w \bullet x_i)+b=+1$距离原点的垂直距离为$\frac{|1-b|}{\|w\|}$，而超平面$(w \bullet x_i)+b=-1$距离原点的垂直距离为$\frac{|-1-b|}{\|w\|}$，因此分类间隔就等于$\frac{|1-b+1+b|}{\|w\|}=\frac{2}{\|w\|}$，所以使间隔最大等价于使$\|w\|$（或$\|w\|^2$）最小。

若要求分类线对所有样本正确分类，则要求它满足：

$$y_i[(w \cdot x_i)+b]-1 \geqslant 0，i=1,2,\ldots,n$$

因此满足该条件且使 $\|w\|^2$ 最小的分类超平面就是最优分类超平面。



过两类样本中离分类超平面最近点且平行于最优分类面的超平面的训练样本就是使等号成立的哪些样本，它们叫做支持向量(Support Vectors)。

2.8、最优分类超平面

- 线性可分情况（续）

最优分类超平面问题可以表示成如下约束优化问题

$$\min \Phi(w) = \min_{w,b} \frac{1}{2}\|w\|^2 = \min_{w,b} \frac{1}{2}(w \cdot w)$$

其约束条件为

$$y_i[(w \bullet x_i)+b]-1 \geqslant 0, \quad i=1,2,\ldots,n$$

定义Lagrange函数：

$$L(w,b,\alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{n} \alpha_i \{y_i[(w \cdot x_i)+b]-1\}$$

其中，α_i>0为Lagrange系数。分别对w和b求偏微分并令它们等于0，得

$$\frac{\partial L(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{n} y_i \alpha_i x_i = 0 \quad \Rightarrow \quad w = \sum_{i=1}^{n} y_i \alpha_i x_i$$

$$\frac{\partial L(w,b,\alpha)}{\partial b} = \sum_{i=1}^{n} y_i \alpha_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} y_i \alpha_i = 0$$

带入原始Lagrange函数，得

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$
\begin{aligned}
\mathcal{L}(\mathrm{w}, \mathrm{b}, \alpha) &= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(w^T x^{(i)} + b \right) - 1 \right] \\
&= \frac{1}{2}\mathrm{w}^T\mathrm{w} - \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i \\
&= \frac{1}{2}\mathrm{w}^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i \\
&= \frac{1}{2}\mathrm{w}^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - w^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i \\
&= -\frac{1}{2}\mathrm{w}^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i \\
&= -\frac{1}{2}\mathrm{w}^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - b \sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha_i \\
&= -\frac{1}{2}\left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - b \sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha_i \\
&= -\frac{1}{2}\sum_{i=1}^{m} \alpha_i y^{(i)} \left(x^{(i)} \right)^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - b \sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha_i \\
&= -\frac{1}{2} \sum_{i=1,j=1}^{m} \alpha_i y^{(i)} \left(x^{(i)} \right)^T \alpha_j y^{(j)} x^{(j)} - b \sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha_i
\end{aligned}
$$

2.8、最优分类超平面

- 线性可分情况（续）

因此，原问题转换为对偶问题：

在约束条件：

$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

$$\alpha_i \geq 0, i = 1,2,\cdots,n$$

之下对α $_i$求解下列函数的最大值：

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

对偶问题完全是根据训练数据来表达的。所得到的解α $_i$只有一部分(通常是少部分)不为零，对应的样本就是支持向量。

若 α_i^* 为最优解，则

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i \ , \quad b^* = \frac{1}{y_s} - w^* \cdot x_s$$

其中，x_S为任一支持向量。最后得到的最优分类函数为：

$$f(x) = \mathrm{sgn}\left[\left(w^* \cdot x\right) + b^*\right] = \mathrm{sgn}\left[\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*\right]$$

小结

本章介绍了“确定性模式”分类器的几种分类训练方法，同时还介绍了广义线性判别函数。其中：线性判别函数方法都是求解向量 w^* 包含下列四种方法：

① 感知器算法：是梯度下降法的特例，固定增量法。

② 梯度下降。

③ H—K算法：最小均方误差算法，LMSE算法。

④ Fisher：不用迭代运算，但要求 $\vec{S}_w$ 类内总离散度矩阵。

非线性判别函数包含下列三种方法：

①分段线性判别函数。

②二次判别函数

③用势函数法直接求 $g(x)$。