第 4 章非参数判别分类方法

第4章 非参数判别分类方法

- 贝叶斯分类器采用错误率最小或风险最小作为指标，构造判别函数和决策面，这给出了一般情况下“最优”分类器的设计方法，对各种不同的分类器的设计具有指导意义。
- 直接使用贝叶斯决策理论需要已知有关样本总体分布的知识，如各类的先验概率、类条件概率密度函数，然后计算出样本的后验概率，并据此设计出相应的判别函数与决策面。
- 实际问题中很难获取准确的统计分布，需要借助第3章的理论，进行较为困难的参数估计（或非参数估计）。

第4章 非参数判别分类方法

- 为避免上述困难，本章我们将跳过参数估计，依据不同的准则函数，由样本直接设计出满足要求的分类器（判别函数、决策面）。
- 这种分类器设计技术称为基于非参数方法的分类器设计技术。
- 之所以称为非参数方法，是指该方法不寻求获取样本的统计分布，也不需要估计统计分布的参数。
- 注意：本章所谓的非参数方法不代表没有参数，而是指没有统计分布的参数。事实上，非参数方法中的判别函数也是有参数的，只是没有统计分布的参数。

4.1线性分类器

- 在本节中，我们假设所有类别的模式向量都可以用线性分类器来正确分类，将讨论线性判别函数的定义和计算方法，以及线性分类器的设计方法。
- 4.1.1 线性判别函数的基本概念
- 首先考虑两类问题的线性判别函数，设模式向量X是一个d维向量，两类问题中线性判别函数的一般形式则可表示为

$$d(X) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + w_{d+1} = W_0^{\mathrm{T}} X + w_{d+1} \quad (4\text{-}1)$$

- 式中，$W_0 = [w_1, w_2, \cdots, w_d]^{\mathrm{T}}$ 称为权向量或参数向量；$X = [x_1, x_2, \cdots, x_d]^{\mathrm{T}}$ 是d维特征向量，又称模式向量或样本向量；w_{d+1} 是常数，称为阈值。为了简洁起见，式（4-1）也可写成

$$d(X) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + w_{d+1} \cdot 1$$

$$= \left[w_1, w_2, \cdots, w_d, w_{d+1} \right] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \\ 1 \end{bmatrix} = W^{\mathrm{T}} X \qquad (4\text{-}2)$$

- 其中，$W = \left[w_1, w_2, \cdots, w_d, w_{d+1} \right]^{\mathrm{T}}$ 为增广权向量；$X = \left[x_1, x_2, \cdots, x_d, 1 \right]^{\mathrm{T}}$ 为增广特征向量，增广特征向量的全体称为增广特征空间。
- 在给出线性判别函数后，如果满足如下决策规则

$$\begin{cases} d(X) > 0, & X \in \omega_1 \\ d(X) < 0, & X \in \omega_2 \\ d(X) = 0, & \text{不确定} \end{cases} \qquad (4\text{-}3)$$

- $d(X)=0$就是相应的决策面方程，在线性判别函数条件下，它对应*d*维空间的一个超平面。
- 对于两类问题，如果模式样本为二维特征向量，则所有分布在二维平面的模式样本可以用一条直线进行划分，这条直线就可以作为一个分类依据，其决策面方程可以表示为

$$d(X)=w_1x_1+w_2x_2+w_3=0 \qquad (4\text{–}4)$$

- 式中，x_1,x_2为坐标变量；w_1,w_2,w_3为方程参数，决策规则依然为式（4-3），两类二维模式的分布示意图见4-1。

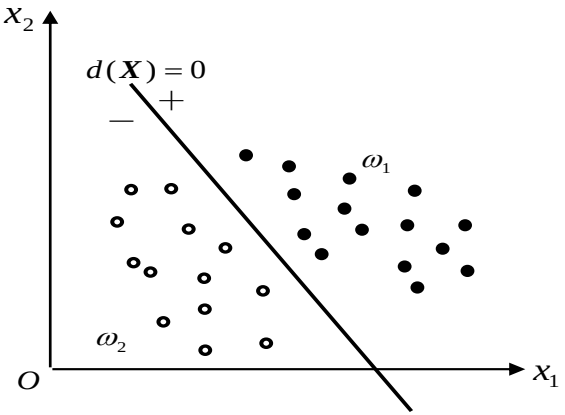

图4-1 两类二维模式的分布

- 4.1.2 多类问题中的线性判别函数
- 下面将讨论多类问题的解决方案，将两类问题进行推广可将应用扩展到多类情况。假设样本整体有 $\omega_1, \omega_2, \cdots, \omega_c$ 共 c 个模式类，且 $c \geq 3$ 。
- 为了把所有的类别分开，存在三种不同的技术途径。
- 一、$\omega_i / \overline{\omega_i}$ 两分法
- $\omega_i / \overline{\omega_i}$ 两分法的基本思想是通过一个线性判别函数，将属于 ω_i 类的模式与其余不属于 ω_i 类的模式分开。对于 c 类问题，如果样本模式是完全线性可分的，则需要 $c-1$ 个独立的判别函数。为了方便，可建立 c 个判别函数，形如

- $$d_i(X) = W_i^{\mathrm{T}} X \ , \ i = 1, 2, \cdots, \ c \qquad (4\text{–}5)$$
- 其中，每一个判别函数具有以下功能
- $$\begin{cases} d_i(X) > 0 & X \in \omega_i \\ d_i(X) < 0 & X \notin \omega_i \end{cases} \quad i = 1, 2, \cdots, \ c \qquad (4\text{–}6)$$
- 通过这类判别函数，把 c 类问题转为 c 个属于 ω_i 和不属于 ω_i 的问题。若把不属于 ω_i 记为 $\overline{\omega_i}$ ，上述问题就成了 c 个属于 ω_i 还是属于 $\overline{\omega_i}$ 的两类问题，因此称为 $\omega_i / \overline{\omega_i}$ 二分法。

- 由上述分析知道，决策面 $d_i(X)=W_i^{\mathrm{T}}X=0$ 把空间划分成两个区域，一个属于ω_i ，另一个属于 $\overline{\omega_i}$。再考察另一个决策的判别函数 $d_j(X)=W_j^{\mathrm{T}}X,\ (j\neq i)$ ，其决策面 $W_j^{\mathrm{T}}X=0$ 同样把特征空间划分成两个区域，一个属于ω_j，另一个属于 $\overline{\omega_j}$ 。这两个决策面分别确定的属于ω_i 和ω_j 的区域可能会有重叠，该重叠区域不能由这两个判别函数确定类别。同样，$\overline{\omega_i}$ 和 $\overline{\omega_j}$ 也可能出现重叠。如果由 c个决策面确定的 c个属于 $\overline{\omega_i}(i=1,\cdots,c)$ 的区域有一个共同的重叠区域，当样本落入该区域时，判别函数不能对它所属的类别做出判决。

- 因此，在使用上述判别函数时，可能会出现两个或两个以上的判别式都大于零，或所有的判别式都小于零的情况。也即特征空间会出现同属于两个类别以上的区域和不属于任何类别的区域，样本落入这些区域时，就无法作出最后判断，这样的区域就是不确定区，用IR标记。样本的类别越多，不确定区IR就越多。
- 示例见书中图4-2.

- 在二维空间里，图4-2给出了3个类别的决策面 $d_i(X)=0$（$i=1,2,3$），出现了4个不确定区。由于不确定区的存在，仅有 $d_i(X)>0$ 不能做出最终判决 $X\in\omega_i$，还必须检查另外的判决函数 $d_j(X)$ 的值。若 $d_j(X)\le 0$，且 $j\neq i$ 才能确定。所以此时判决规则为
- 如果 $\begin{cases} d_i(X)>0 \\ d_j(X)\le 0 \quad j\neq i \end{cases}$ 则 $X\in\omega_i$。（4–7）

- 二、ω_i / ω_j两分法
- ω_i / ω_j 两分法的基本思想是对 c 个类别中的任意两个类别 ω_i 和 ω_j 建立一个判别函数 $d_{ij}(X)$，决策面方程为 $d_{ij}(X)=0$，它能把 ω_i 和 ω_j 两个类别区分开，但对其他类别的区分则不提供任何信息。在 c 个类别中，任取两个类别的组合数为 $c(c-1)/2$，其中能分开 ω_i 和 ω_j 这两类的判别函数为
- $$d_{ij}(X)=W_{ij}^{\mathrm{T}}X \ , i, j=1, 2, ..., c \text{ 且 } i \neq j \qquad (4\text{-}8)$$
- 此时，判别函数具有性质
- $$d_{ij}(X)=-d_{ji}(X) \qquad (4\text{-}9)$$
- 每个判别函数具有以下功能
- $$d_{ij}(X)\begin{cases} >0 & X \in \omega_i \\ <0 & X \in \omega_j \end{cases} \qquad (4\text{-}10)$$

- 从（4-8）式可知，这类判别函数也是把 c 类问题转变为两类问题，与 $\omega_i / \overline{\omega_i}$ 两分法不同的是，其两类问题的数目不是 c 个，而是 $c(c-1)/2$ 个，且每个两类问题不是 $\omega_i / \overline{\omega_i}$，而是 ω_i / ω_j 。也就是，此时转变成了 $c(c-1)/2$ 个 ω_i / ω_j 两类问题。
- 只有一个决策面 $d_{ij}(X)=0$ 无法最后决定样本 X 的最终归属，因为 $d_{ij}(X)$ 只涉及 ω_i 和 ω_j 的关系，只能判定 X 是位于含有 ω_i 类的空间，还是位于含有 ω_j 类的空间，而对它们和别的类别 ω_k（$k=1, 2, \cdots, c$, 且 $k \neq i, k \neq j$）的关系不提供任何信息。要得到 X 的判别结论，必须考察 $c-1$ 个判决函数。即有判决规则：
- 如果 $d_{ij}(X)>0,\ j=1, 2, \cdots, c$, 且 $j \neq i$ ，则 $X \in \omega_i$ 。 （4−11）
- 例4-2表明，该方法仍存在不确定区，只是不确定区的数量减少了。

- 三、没有不确定区域的 ω_i / ω_j 两分法
- 该方法的思想是对c类中的每一个类别，均建立一个判决函数，即
- $d_i(X)=W_i^{\mathrm{T}}X,\ i=1,\ 2,\ \cdots,\ c$ 。（4-12）
- 为了区分出其中的某一个类别 ω_i，则需要k个判决函数（ $k \le c$ ），其判别规则为
- 如果 $d_i(X)>d_j(X),\ j \ne i$，则 $X \in \omega_i$ 。（4-13）
- 上述判决规则也可以有另一种表示形式为
- 如果 $d_i(X)=\max\limits_{j=1,2,\cdots,k}\{d_j(X)\}$，则 $X \in \omega_i$ 。（4-14）

- 显然，对不同的 ω_i，k的取值不尽相同，k值的选择与类别之间的相邻关系密切相关。下面举例说明，如图4-4所示，特征空间里有一个五类问题，五个不同的类别可用分段线性函数分开。从图中可以看出，类别 ω_1与其余4个类别均相邻，ω_2 分别与 ω_1和ω_3相邻，ω_5 分别与ω_1 、ω_3和ω_4相邻。 k的取值取决于与所考察的类别相邻的类别的个数，如：ω_1，$k=4$；ω_2 ，$k=2$；ω_5，$k=3$。

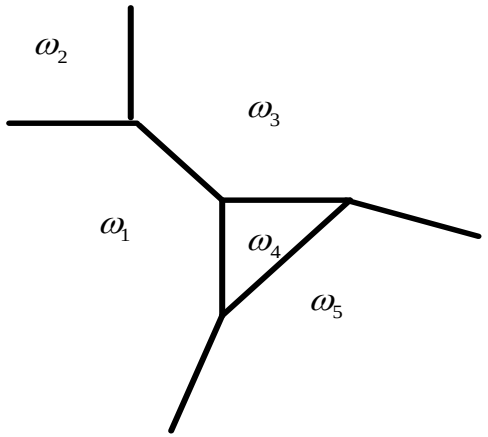

图4-4 五类问题

- 下面进一步讨论，该类方法与ω_i / ω_j两分法的区别，假定c=3，且已建立下列3个判决函数。

$$\begin{cases} d_1(X) = W_1^{\mathrm{T}} X \\ d_2(X) = W_2^{\mathrm{T}} X \\ d_3(X) = W_3^{\mathrm{T}} X \end{cases}$$

- 三个类别的区域均相邻，有

$$d_{12}(X) = d_1(X) - d_2(X) = (W_1^{\mathrm{T}} - W_2^{\mathrm{T}})X = W_{12}^{\mathrm{T}} X$$

- 同理 $d_{13}(X) = W_{13}^{\mathrm{T}} X$ ，$d_{23}(X) = W_{23}^{\mathrm{T}} X$。
- 又由

$$\begin{aligned} d_{23}(X) &= d_1(X) - d_1(X) + d_2(X) - d_3(X) \\ &= [d_1(X) - d_3(X)] - [d_1(X) - d_2(X)] \\ &= d_{13}(X) - d_{12}(X) \end{aligned}$$

- 可知 $d_{23}(X)$是$d_{13}(X)$和$d_{12}(X)$ 的线性组合，换句话说，若$d_{13}(X)$和$d_{12}(X)$是独立的，则 $d_{23}(X)$是非独立的，且在二维空间中，三个判决函数必须相交于一点，如图4-5。

从书中图4-5三个类别的分布情况来看，该类方法与传统ω_i / ω_j两分法的区别在于该方法没有不确定区。对于c个类别来说，该类方法的判决函数的数量为c个，不是传统ω_i / ω_j 两分法的$c(c\text{-}1)/2$个。尽管有此差别，该类方法的判别式$d_i(X) > d_j(X)$与 ω_i / ω_j 两分法的判别式 $d_{ij}(X) > 0$ 等价。

三类方法的关系：

（1）没有不确定区的ω_i / ω_j两分法更像是$\omega_i / \overline{\omega_i}$两分法和传统$\omega_i / \omega_j$两分法的结合。从判别函数形式上看，没有不确定区的$\omega_i / \omega_j$两分法与$\omega_i / \overline{\omega_i}$两分法类似，都有$c$个判别函数。区别在于，$\omega_i / \overline{\omega_i}$两分法的判别规则要求只有一个判别函数大于0，其他判别函数小于等于0；没有不确定区的ω_i / ω_j两分法则是通过比较$k(k \leq c)$个判别函数的大小来作出判决，与单个判别函数大于0还是小于等于0没关系。

三类方法的关系：

（2）没有不确定区的ω_i / ω_j两分法和传统 ω_i / ω_j 两分法的判别函数存在某种等价关系。没有不确定区的 ω_i / ω_j 两分法的判别函数 $d_i(X) > d_j(X)$ 与传统 ω_i / ω_j 法的判别函数 $d_{ij}(X) > 0$ 等价。

（3）三类方法中，只有没有不确定区的ω_i / ω_j 两分法没有不确定区，其他两类方法都有不确定区，但传统 ω_i / ω_j 两分法的不确定区数量少于 $\omega_i / \overline{\omega_i}$ 两分法的不确定区数量。

- 4.1.3广义线性判别函数
- 线性判别函数是形式最为简单的判别函数，但在实际应用中有较大的局限性，对稍复杂一些的情况，线性判别函数就有可能失效。例如，在一维空间中的两类模式，其分布如图4-6所示，两类模式的分布类域为$\omega_1:(-\infty,b)\cup(a,+\infty)$；$\omega_2:(b,a)$。若要将两类模式正确分类，则需设计一个一维分类器，满足如下性能

$$\text{如果}\begin{cases} X<b\text{或}X>a, & X\in\omega_1 \\ b\le X\le a, & X\in\omega_2 \end{cases} \tag{4-15}$$

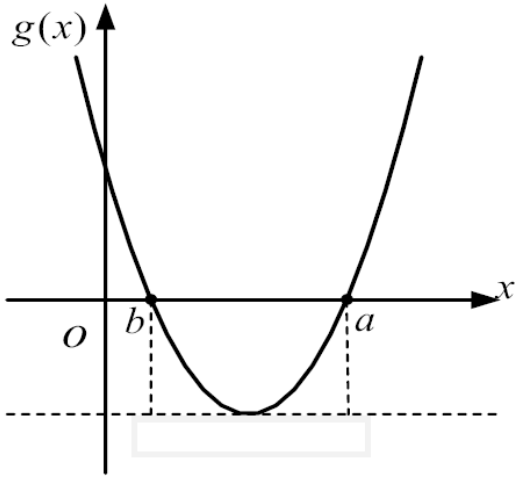

图4-6 二次判别函数举例

显然，这两类模式不是线性可分的，式（4-15）的分类器无法采用线性判别函数实现，针对这种情况，可设计二次判别函数

$$d(X)=(X-a)(X-b)=X^2-(a+b)X+ab \qquad (4-16)$$

其相应的决策规则为

$$\begin{cases} d(X)>0, X\in\omega_1 \\ d(X)\le 0, X\in\omega_2 \end{cases} \qquad (4-17)$$

- 如图4-6所示，此时$d(X)$是X的非线性函数。由于线性判别函数形式简单、计算方便，因此人们希望能找到一种能将非线性可分问题转化为线性可分问题的方法。其思路是选择一种映射$X \to Y$，即将原样本特征向量X映射成另一向量Y，从而可以对Y采用线性判别函数来分类。例如对于图4-6的二次函数情况，其一般式可表示成
- $$d(X) = c_0 + c_1 x + c_2 x^2 \qquad (4\text{-}18)$$
- 如果采用映射$X \to Y$，使 $$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix}$$
- 则判别函数$d(X)$又可表示成
- $$d(X) = a^{\mathrm{T}} Y = \sum_{i=1}^{3} a_i y_i \qquad (4\text{-}19)$$
- 其中 $a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix}$
- 由此可见，样本原来在一维空间中线性不可分，但当其转换到二维空间后（此处二维空间指的是出现了X的二次项），样本就变成线性可分了。

- 此时$d(X)$被称为广义线性判别函数，a称为广义权向量。因此一个原属二次函数的非线性分类问题就转化为一个线性分类问题。事实上，可以将这类方法一般化，任何形式的高次判别函数都可转化成线性判别函数来处理。设样本集$\{X_i\}$在原始的n维特征空间是非线性可分的，对各模式X_i进行非线性变换$T: X^n \to Y^m$，$m > n$，使得样本模式在特征空间Y^m中是线性可分的，即分类面是线性的。

- 设有一个n维的训练样本模式集在模式空间X中线性不可分。
- 定义广义形式的模式向量为
- $$Y=[y_1,y_2,\cdots,y_m,1]^{\mathrm{T}}=\left[f_1(X),f_2(X),\cdots,f_m(X),1\right]^{\mathrm{T}} \quad (4\text{-}21)$$
- 这里 $f_i(X)$ 通常为X 的非线性函数，Y空间的维数m高于X空间的维数n，判别函数可写为 $$d(Y)=W^{\mathrm{T}}Y \quad (4\text{-}22)$$
- 其中$W=\left[w_1,w_2,\cdots,w_m,w_{m+1}\right]^{\mathrm{T}}$是增广权向量， Y是增广模式向量，其所在的空间是一个m维的空间，称为Y空间。这里$d(Y)$ 也称为广义线性判别函数，它本质上是特征空间Y空间中的线性判别函数，原始空间X空间中的非线性判别函数。

- 用广义线性判别函数虽然可以将非线性问题转化为简单的线性问题来处理，但是实现这种转化的非线性变换的形式可能非常复杂。另外，在原空间X中模式样本是n维向量，在新空间Y中，模式样本是m维向量，通常m比n大很多，经过上述变换，维数大大增加了。
- 模式样本特征维数的增加会导致计算量的迅速增加，以致计算机难以处理，这就是所谓的“维数灾难”。

- 4.1.4线性分类器的主要特性及设计步骤
- 1．线性分类器的主要特性
- 1）模式空间与超平面
- 设有d维模式向量X，则以X的d个分量为坐标变量的欧式空间称为模式空间。在模式空间里，模式向量可以表示成一个点，也可以表示成从原点出发到这个点的一个有向线段。当模式类别线性可分时，判别函数的形式是线性的，剩下的问题就是确定一组系数，从而确定一个符合条件的超平面。对于两类问题，利用线性判别函数$d(X)$进行分类，就是用超平面$d(X)=0$ 把模式空间分成两个决策区域。
- 设判别函数为
- $$d(X)=w_1x_1+w_2x_2+\cdots+w_dx_d+w_{d+1}=W_0^{\mathrm{T}}X+w_{d+1} \quad (4\text{-}23)$$
- 式中，$W_0=[w_1,w_2,\cdots,w_d]^{\mathrm{T}}$，$X=[x_1,x_2,\cdots,x_d]^{\mathrm{T}}$，则超平面为
- $$d(X)=W_0^{\mathrm{T}}X+w_{d+1}=0 \quad (4\text{-}24)$$

- 为了说明线性判别函数中向量 W_0的意义，假设在该决策平面上有两个特征向量X_1 与 X_2 ，如图（4-7）（a）所示，将 X_1 与 X_2 代入式（4-23），则有
- $$W_0^{\mathrm{T}} X_1 + w_{d+1} = W_0^{\mathrm{T}} X_2 + w_{d+1} \qquad (4\text{-}25)$$
- 也即
- $$W_0^{\mathrm{T}} (X_1 - X_2) = 0 \qquad (4\text{-}26)$$
- 其中$(X_1 - X_2)$ 也是一个向量。式(4-26) 的几何意义是向量 W_0 与该平面上任两点组成的向量$(X_1 - X_2)$ 正交。也就是说，W_0 就是超平面 $d(X) = 0$ 的法向量，方向由超平面的负侧指向正侧。设超平面的单位法向量为U，则有
- $$U = W_0 / \| W_0 \| \qquad (4\text{-}27)$$
- 式中$\| W_0 \|$可理解为向量 W_0 的模值，由下式计算得到
- $$\| W_0 \| = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2} \qquad (4\text{-}28)$$

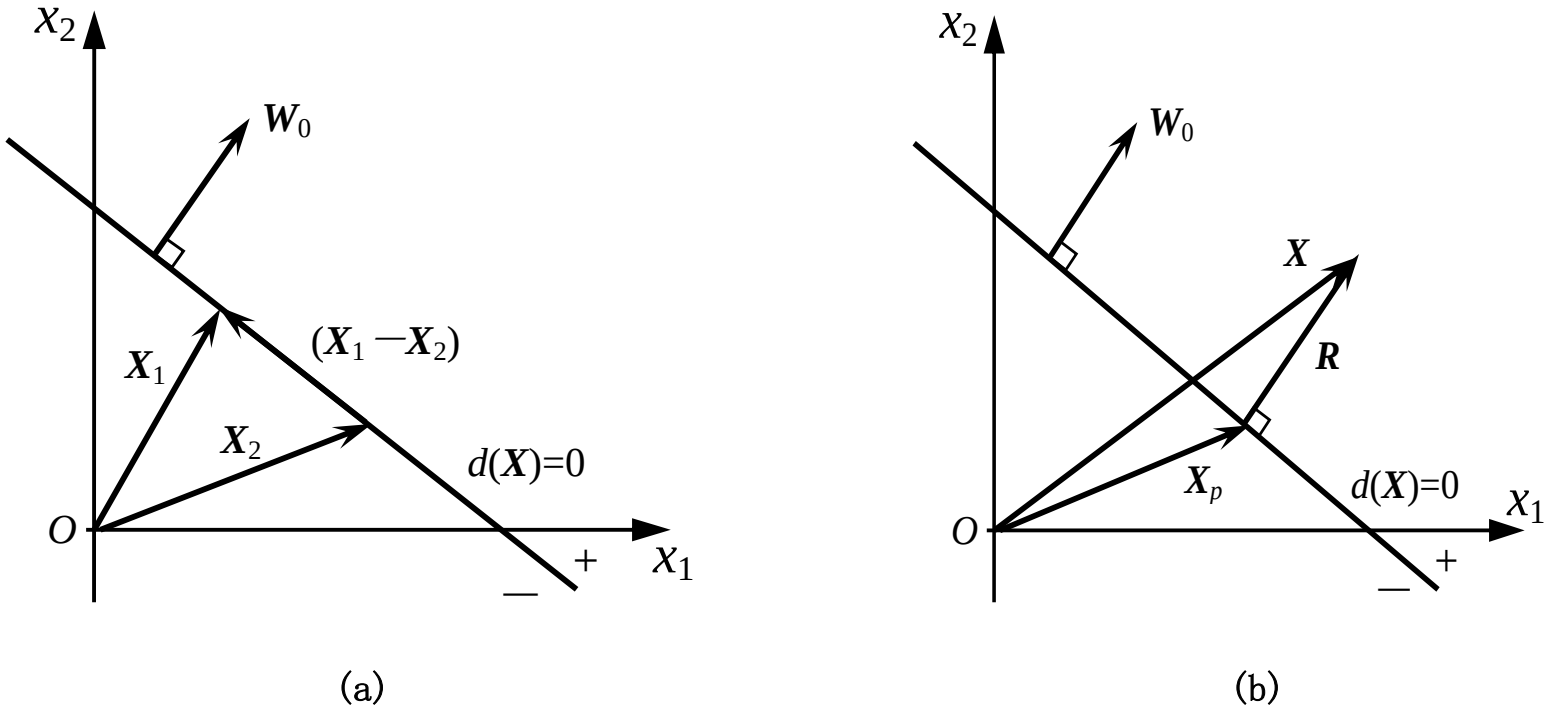

图4-7 点到超平面的距离

- 设X为不在超平面上的模式点，将X向超平面投影得向量X_p，并构造向量R，如图（4-7）（b）所示，由式（4-25）有

$$R = r \cdot \boldsymbol{U} = r W_0 / \| W_0 \|$$

- 式中，r为X到超平面的垂直距离。这样，X就可以表示成

$$X = X_p + R = X_p + r W_0 / \| W_0 \| \qquad (4\text{-}29)$$

- 将（4-29）代入式（4-23）得到

$$d(X) = W_0^{\mathrm{T}}(X_p + r W_0 / \| W_0 \|) + w_{d+1} = (W_0^{\mathrm{T}} X_p + w_{d+1}) + W_0^{\mathrm{T}} \cdot r W_0 / \| W_0 \| \qquad (4\text{-}30)$$

- 因X_p位于超平面上，故式（4-30）中第一项为零，应用$W_0^{\mathrm{T}} W_0 = \| W_0 \|^2$，得

$$d(X) = r \| W_0 \| \qquad (4\text{-}31)$$

- 因此，X到超平面的距离为

$$r = \frac{d(X)}{\| W_0 \|} \qquad (4\text{-}32)$$

- 图（4-7）（b）中X位于超平面的正侧，因而$d(X)>0$；若位于超平面的负侧，则 $d(X)<0$ 。式（4-32）表明点到超平面的距离（带正负号）正比于$d(X)$的函数值。也可以看出，对于两类问题，可按两类样本到决策面距离的正负号确定其类别。
- 对于式（4-23），当X为原点时，$d(X)=w_{d+1}$，原点到超平面的距离为

$$r=\frac{w_{d+1}}{\|W_0\|} \tag{4-33}$$

- 该式说明决策面的位置是由权值 w_{d+1} 决定的，当 $w_{d+1}=0$ 时，该决策面过原点，而 $w_{d+1}\neq 0$ 时，则 $w_{d+1}/\|W_0\|$表示原点到该决策面的距离。如果 $w_{d+1}>0$ ，原点在超平面的正侧；如果$w_{d+1}<0$ ，原点在超平面的负侧。

2）权空间与权向量解

在模式识别过程中，有时需要将判别函数绘制在权向量空间中。设有式（4-23）的线性判别函数

$$d(X) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + w_{d+1}$$

则以 $w_1, w_2, \ldots, w_d, w_{d+1}$ 为坐标变量构成的空间称为权空间。在权空间里，$d+1$ 维增广权向量 $W = [w_1, w_2, \ldots, w_d, w_{d+1}]^{\mathrm{T}}$ 对应该空间中的一个点，也可以用从原点出发到这个点的一条有向线段来表示。

当样本类别线性可分时，根据已知训练样本确定 $d(X)$ 的任务等同于确定符合条件的权向量 $W = [w_1, w_2, \ldots, w_d, w_{d+1}]^{\mathrm{T}}$ 。

下面以两类问题为例，讨论线性判别函数 $d(X)$ 中权向量的求解问题。

- 设ω_1类有$X_{11}, X_{12}, \cdots, X_{1p}$共$p$个增广样本向量，$\omega_2$类有 $X_{21}, X_{22}, \cdots, X_{2q}$

共q个增广样本向量。建立判别函数的任务是确定$d(X)$把ω_1 类和ω_2 类分开，若线性判别函数为$d(X)=W^{\mathrm{T}}X$ ，则有如下不等式成立

$$\begin{cases} d(X_{1i})>0,\ i=1,2,\cdots,p \\ d(X_{2i})<0,\ i=1,2,\cdots,q \end{cases} \qquad (4\text{-}24)$$

- 式（4-24）共包含$p+q$个不等式，如果将 ω_2 的q个增广样本向量都乘以-1，则式（4-24）可写为

$$\begin{cases} d(X_{1i})>0,\ i=1,2,\cdots,p \\ d(-X_{2i})>0,\ i=1,2,\cdots,q \end{cases} \qquad (4\text{-}25)$$

这样就可以不管原样本的类别，将把两类模式分开的条件统一写为 $d(X)>0$ ，其中

$$X=\begin{cases} X_{1i}, & i=1,2,\cdots,p \\ -X_{2i}, & i=1,2,\cdots,q \end{cases} \qquad (4\text{-}26)$$

- 这一过程叫做样本的规范化过程，上式中的X称为规范化增广样本向量。
- 在不等式（4-25）中，X是已知的，增广权向量W的各个分量是未知的。我们需要做的就是寻找一个W使得式（4-25）中的$p+q$个不等式同时成立，因此满足条件的W必然位于$p+q$个超平面的正侧的重叠区域里，这个区域就是W的解区。
- 二维空间中权空间和权向量解区如图4-8所示，在该例子中超平面均为过原点的直线，图中阴影部分就是解区。

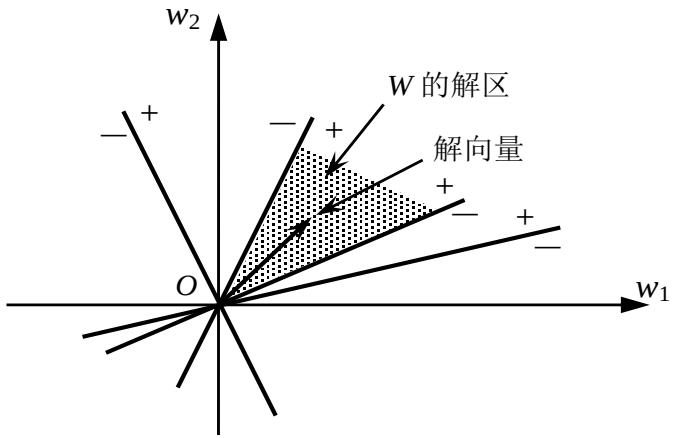

图4−8 权向量的解区和解向量

理论上来讲，解区中的任意权向量都满足式（4-25），但是，从新样本模式的适应性来看，选取位于解区中央部分的权向量更可靠，对新样本正确分类的概率更高。

在高维权空间里，满足式（4-25）的权向量的解区对应着多个超平面围成的超锥面体的内部区域。

- 2 线性分类器的设计步骤
- 通常来说，当属于两个类别的模式样本在特征空间中能够被一个超平面区分时，则这两个类别是线性可分的。进一步推论，对于一个已知容量为N的样本集，如果有一个超平面能把每一个样本正确分类，则称该样本集是线性可分的。
- 要确定超平面的权向量W，需首先采集一些样本，并把它们表示为特征空间中的模式向量，这些模式向量称为训练样本。
- 训练样本集是若干训练样本的集合，基于此所做的实验为训练实验。如果训练样本集中每个训练样本的类别都是预先已知的，这样的训练实验就是监督试验。
- 由训练样本集所确定的权向量称为解向量W^*。

- 一般来说，学习线性分类器的主要步骤是：
- 第一步：采集训练样本，构成训练样本集。
- 第二步：确定一个准则函数 $J = J(W, X)$，该准则函数能反映分类器的性能。
- 第三步：设计求解算法，求使准则函数J取最大(J为增益函数)或最小(J为代价函数/损失函数)的权值，得到解向量W^*。

- 4.1.5感知器算法
- 一、感知器算法的原理
- 对线性判别函数而言，当模式样本的维数已知时，判别函数的形式就已经确定。例如，若模式样本的维数等于三，则相应的线性判别函数为

$$d(X) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 = W^{\mathrm{T}} X$$

- 其中，$X = [x_1, x_2, x_3, 1]^{\mathrm{T}}$为增广模式(样本)向量，$W = [w_1, w_2, w_3, w_4]^{\mathrm{T}}$为增广权向量。只要求出增广权向量$W$，分类器的设计即告完成。
- 本节介绍如何通过感知器算法，利用已知类别的模式样本学习出增广权向量W。

- 感知器算法的基本思想：首先设置一个初始的增广权向量，然后用已知类别的模式样本去检验增广权向量的合理性，若不合理则对其进行修正，一般需经过多次修正直到合理为止。修正的方法常采用梯度下降法。
- 设有两类线性可分的模式类ω_1和ω_2，判别函数为 $d(X)=W^{\mathrm{T}}X$，其中$W=\left[w_1,w_2,\cdots,w_d,w_{d+1}\right]^{\mathrm{T}}$，则$d(X)$ 应具有如下性质

$$d(X)=W^{\mathrm{T}}X\begin{cases}>0 & X\in\omega_1\\ <0 & X\in\omega_2\end{cases} \qquad (4\text{-}31)$$

- 若对样本进行规范化处理，即将 ω_2 类的全部样本都乘以-1，则判别函数的性质为 $d(X)=W^{\mathrm{T}}X>0$ （4-32）
- 感知器算法通过对已知类别的训练样本集的学习，寻找一个满足式（4-31）或式（4-32）的增广权向量。

- 设有训练样本集 $X=\{X_1, X_2, ...X_n\}$，其中每个样本属于ω_1类或ω_2类，且类别属性已知。为了确定最优的增广权向量W^*，感知器训练算法的具体步骤如下：
- ① 令k=0，给增广权向量$W(k)$赋任意初值(即分别给增广权向量每个分量赋任意值)，取常数$1 \geq c > 0$.
- ② 输入一个训练样本 X_k ， $X_k \in \{X_1, X_2, ..., X_n\}$.
- ③ 计算判决函数值： $d(X_k) = [W(k)]^{\mathrm{T}} X_k$.
- ④ 根据以下规则修正增广权向量：

  若 $X_k \in \omega_1$ 且 $d(X_k) \leq 0$，则令 $W(k+1) = W(k) + cX_k$.

  若 $X_k \in \omega_2$ 且 $d(X_k) > 0$，则令 $W(k+1) = W(k) - cX_k$.

  如果$X_k \in \omega_2$，且X_k的各分量均乘以-1，则修正规则统一为：

  若 $d(X_k) \leq 0$ ，则令 $W(k+1) = W(k) + cX_k$.

- ⑤ 若W对所有训练样本均稳定不变，则结束。否则令$k=k+1$，返回②继续修正。
- 注意：常数c的取值范围为$1\geq c>0$。c值太小，会导致算法收敛速度变慢； c值太大， 会使$W(K)$的值不稳定。
- 感知器算法对权值的修正过程本质上是一种赏罚过程，若 $X_k \in \omega_1$，但 $d(X_k)\leq 0$，说明当前权向量构造的判别函数对样本 X_k 做出了错误的分类，应修正权向量，使 $W(k+1)=W(k)+cX_k$ ，其目的是使 $[W(k+1)]^{\mathrm{T}} X_k > [W(k)]^{\mathrm{T}} X_k$ ；同样，若 $X_k \in \omega_2$，但 $d(X_k)>0$，同样说明当前权向量构造的判别函数对样本 X_k 做出了错误的分类，应修正权向量，使 $W(k+1)=W(k)-cX_k$ ，令 $[W(k+1)]^{\mathrm{T}} X_k < [W(k)]^{\mathrm{T}} X_k$.

- 由此可见，当发生分类错误时，感知器算法通过修改权向量做出“惩罚”，从而使其向正确的方向转换；分类正确时，则对其进行“奖励”——这里表现为“不罚”，即权向量不变。
- 如果经过有限次迭代运算后，求出了一个使训练集中所有样本都能正确分类的权向量，则称算法是收敛的。
- 可以证明感知器算法是收敛的。对于感知器算法，只要模式类别是线性可分的，就可以在有限的迭代步数里求出权向量的解。

例4.6 一个两类问题，4 个训练样本分别为 $\omega_1:(0,0)^T,(0,1)^T$，$\omega_2:(1,0)^T,(1,1)^T$，试用感知器算法求权向量 W^*。

解：将训练样本变为增广型的，ω_2 样本乘以(-1)，得到 4 个样本 $\{X_1,X_2,X_3,X_4\}$ 的增广向量：

$$X_1=(0,0,1)^t,\quad X_2=(0,1,1)^t,\quad X_3=(-1,0,-1)^t,\quad X_4=(-1,-1,-1)^t$$

判别函数为

$$d(X)=W^TX=(w_1,w_2,\ldots w_d,w_{d+1},)\begin{pmatrix}x_1\\ \vdots\\ x_d\\ 1\end{pmatrix}$$

取初值

$$W(0)=(1,1,1)^T,\quad c=1$$

则有

$k=1$，$X_k=X_1$，$d(X_k)=W^T(k)X_k=1>0$，$W(1)=W(0)$，W 不变

$k=2$，$X_k=X_2$，$d(X_k)=W^T(k)X_k=2>0$，$W(2)=W(1)$，W 不变

$k=3$， $X_k=X_3$， $d(X_k)=W^T(k)X_k=-2<0$， $W(3)=W(2)+X_3=(0,1,0)^T$

$k=4$， $X_k=X_4$， $d(X_k)=W^T(k)X_k=-1<0$， $W(4)=W(3)+X_4=(-1,0,-1)^T$

$k=5$， $X_k=X_1$， $d(X_k)=W^T(k)X_k=-1<0$， $W(6)=W(5)+X_1=(-1,0,0)^T$

$k=6$， $X_k=X_2$， $d(X_k)=W^T(k)X_k=0$， $W(7)=W(6)+X_2=(-1,1,1)^T$

$k=7$， $X_k=X_3$， $d(X_k)=W^T(k)X_k=0$， $W(8)=W(7)+X_3=(-2,1,0)^T$

$k=8$， $X_k=X_4$， $d(X_k)=W^T(k)X_k=1>0$， $W(9)=W(8)$

$k=9$， $X_k=X_1$， $d(X_k)=W^T(k)X_k=0$， $W(9)=W(8)+X_1=(-2,1,1)^T$

$k=10$， $X_k=X_2$， $d(X_k)=W^T(k)X_k=2>0$， $W(10)=W(9)$

$k=11$， $X_k=X_3$， $d(X_k)=W^T(k)X_k=1>0$， $W(11)=W(10)$

$k=12$， $X_k=X_4$， $d(X_k)=W^T(k)X_k=0$， $W(13)=W(12)+X_4=(-3,0,0)^T$

$k=13$， $X_k=X_1$， $d(X_k)=W^T(k)X_k=0$， $W(14)=W(13)+X_1=(-3,0,1)^T$

$k=14$， $X_k=X_2$， $d(X_k)=W^T(k)X_k=1>0$， $W(15)=W(14)$

$k=15$， $X_k=X_3$， $d(X_k)=W^T(k)X_k=2>0$， $W(16)=W(15)$

$k=16$， $X_k=X_4$， $d(X_k)=W^T(k)X_k=2>0$， $W(17)=W(16)$

$k=17$， $X_k=X_1$， $d(X_k)=W^T(k)X_k=1>0$， $W(18)=W(17)$

从 $k=14-17$ 的结果可以看出，使用 $W(14)$ 已经能对所有训练样本正确分类，也就是算法收敛于 $W(14)$， $W(14)$ 即为解向量，即

$$W^*=(-3,0,1)^T$$

对应的判别函数为

$$d(X)=W^{*T}X=(-3,0,1)\begin{bmatrix}x_1\\x_2\\1\end{bmatrix}=-3x_1+1$$

决策面为 $d(X)=0$，也就是 $x_1=\frac{1}{3}$。

- ## 4.1.6 Fisher线性判别函数
- 为了使所设计的线性分类器在性能上满足一定的要求，通常将这种要求表示成一种准则函数，并通过优化准则函数来确定线性分类器中的权向量。
- Fisher准则函数是R.A.Fisher在1936年提出的一种线性判别方法。其基本原理是，将d维空间中的样本投影到一维坐标上，如果能找到一个投影方向，使得不同类别的样本在该投影方向上的交叠最少，则可以较容易地从投影后的特征中区分出样本的类别，从而得到更好的分类效果。
- 在79页图4-9中有两种不同类别的样本，显然沿向量W_2进行投影能使这两类样本明显分开，而沿向量W_1进行投影，则使两类样本部分交叠在一起，无法找到一个能将它们截然分开的界面 。

- 一、两类问题Fisher准则
- 在 ω_i/ω_j 两类问题中，假定有 N 个训练样本 $X_k(k=1,2,....,N)$，其中类别 ω_i 有 N_i 个样本，类别 ω_j 有 N_j 个样本，$N=N_i+N_j$ 。ω_i 和 ω_j 的训练样本分别构成训练样本的子集 Ω_i 和 Ω_j 。可以用投影向量 W 与原特征向量作数量积（内积），其表达式为
- $$Y_k = W^{\mathrm{T}} X_k, \quad k=1,2,\cdots,N \qquad (4\text{-}39)$$
- 相应地，$Y_k(k=1,2,....,N)$ 也构成两个子集 Ψ_i 和 Ψ_j。通常只考虑投影向量 W 的方向不考虑其长度，即默认其长度为单位1。Fisher准则的目的就是寻找最优投影方向 W^*，使得两类样本在投影之后尽可能的分开。

- 一、两类问题Fisher准则
- 分析79页图4-9的两个投影方向，W_2方向之所以比W_1方向优越，是因为沿W_2方向投影后的类间离散程度更大，类内离散程度更小。
- 因此，可以归纳出这样一个准则，即投影向量的方向选择应能使投影后两类样本的均值之差尽可能大，而类内样本（同类样本）的离散程度尽可能小。这就是Fisher准则函数的基本思想。为了将这个思想表示成为可计算的函数值，下面先对一些基本参量下定义。

- 在原始的d维特征空间中， 各类样本均值向量m_k定义为

$$m_k = \frac{1}{N_k} \sum_{X \in \Omega_k} X, \quad k = i, j \tag{4-40}$$

- 样本的类内离散度矩阵S_k与总类内离散度矩阵S_w分别为

$$S_k = \sum_{X \in \Omega_k} (X - m_k)(X - m_k)^{\mathrm{T}}, \quad k = i, j \tag{4-41}$$

$$S_w = S_i + S_j \tag{4-42}$$

- 样本的类间离散度矩阵S_b为

$$S_b = (m_i - m_j)(m_i - m_j)^{\mathrm{T}} \tag{4-43}$$

- 同理，在投影后的一维空间Y中各类样本的均值$\tilde{m}_k$ 定义为
- $$\tilde{m}_k = \frac{1}{N_k}\sum_{Y\in\psi_k} Y = \frac{1}{N_k}\sum_{X\in\Omega_k} W^{\mathrm{T}}X = W^{\mathrm{T}}m_k, \quad k=i,j \tag{4-44}$$
- 样本的类内离散度$\tilde{S}_k$与总类内离散度$\tilde{S}_w$分别为
- $$\tilde{S}_k = \sum_{Y\in\psi_k}(Y-\tilde{m}_k)^2 = \sum_{X\in\Omega_k}(W^{\mathrm{T}}X - W^{\mathrm{T}}m_k)(W^{\mathrm{T}}X - W^{\mathrm{T}}m_k)^{\mathrm{T}} = W^{\mathrm{T}}S_kW, \quad k=i,j \tag{4-45}$$
- $$\tilde{S}_w = \tilde{S}_i + \tilde{S}_j = W^{\mathrm{T}}(\tilde{S}_i + \tilde{S}_j)W = W^{\mathrm{T}}S_wW \tag{4-46}$$
- 样本的类间离散度$\tilde{S}_b$ 为
- $$\tilde{S}_b = (\tilde{m}_i - \tilde{m}_j)^2 = (W^{\mathrm{T}}m_1 - W^{\mathrm{T}}m_2)(W^{\mathrm{T}}m_1 - W^{\mathrm{T}}m_2)^{\mathrm{T}} = W^{\mathrm{T}}S_bW \tag{4-47}$$

- 在定义了上述参量后，就可以根据它们给出Fisher准则的函数形式。根据Fisher准则选择投影方向W的原则是：原始样本向量在该方向上的投影，既能使类间样本尽可能分开，又能使类内样本尽可能聚集。换句话说，就是希望类内离散度 $\tilde{S}_w$ 越小越好，类间离散差度 $\tilde{S}_b$ 越大越好。根据这一原则，用来评价投影方向W的函数为
- $$J_F(W)=\frac{\tilde{S}_b}{\tilde{S}_w}=\frac{W^{\mathrm{T}}S_bW}{W^{\mathrm{T}}S_wW} \qquad (4\text{-}52)$$
- 这个函数称为Fisher准则函数。根据其含义，Fisher准则函数取最大值时对应的W即为最优投影向量W^*。
- 最大化(4-52)的求解过程可以采用拉格朗日乘子法，即保持(4-52)的分母为一非零常数c的条件下，求其分子的极大值，其相应的拉格朗日函数为 $L(W,\lambda)=W^{\mathrm{T}}S_bW-\lambda(W^{\mathrm{T}}S_wW-c)$ (4-53)
- 其中 λ 为拉格朗日乘子。

- 将拉格朗日函数关于W求偏导，并令偏导数等于0，即可求解最优投影向量W^*。可得

$$\frac{\partial L(W^*,\lambda)}{\partial W^*} = S_b W^* - \lambda S_w W^* = 0 \qquad (4\text{-}55)$$

- 对上式化简可得 $$S_b W^* = \lambda S_w W^* \qquad (4\text{-}56)$$
- 当S_w非奇异（可逆）时，将(4-56)两边乘以S_w^{-1}可得
- $$S_W^{-1} S_b W^* = \lambda W^* \qquad (4\text{-}57)$$
- 式(4-57)为关于矩阵 $S_W^{-1}S_b$ 的特征值求解问题。
- 根据式(4-43)关于S_b的定义，可得

$$S_b W^* = (m_i - m_j)(m_i - m_j)^{\mathrm{T}} W^* \qquad (4\text{-}58)$$

- 其中，$(m_i - m_j)^{\mathrm{T}} W^*$ 是一个标量，可用数值R表示，则式(4-58)可写成 $S_b W^* = (m_i - m_j)R$，代入式(4-57)可得
- $$W^* = \frac{R}{\lambda} S_W^{-1}(m_i - m_j) \qquad (4\text{-}59)$$
- 实际上我们关心的只是向量 W^* 的方向，其数值大小对分类器的性能没有影响。因此，在忽略了数值因子 $\frac{R}{\lambda}$ 后，可得
- $$W^* = S_w^{-1}(m_i - m_j) \qquad (4\text{-}60)$$
- 上式的W^*可使Fisher准则函数$J_F(W)$达到最大值，即使用Fisher准则求最佳投影向量的解。使用W^*可将样本X从d维空间投影到一维特征空间($W^{*\mathrm{T}} X$ 为一维向量)，这种投影运算称为线性变换。

- 从上式可以看出，最佳投影W^*的方向是由 $m_i - m_j$ 和 S_W^{-1} 共同决定的。向量 $m_i - m_j$ 为两类样本均值的差分向量，考虑到Fisher准则既要求两类样本的类间距离较大，又要求类内密集程度较高，因此只向 $m_i - m_j$ 方向作投影是不够的，还需根据类内分布的离散程度对投影方向作相应的调整，这就体现在对向量 $m_i - m_j$ 按 S_w^{-1}作修正，使Fisher准则函数取得极大值。
- 以上仅给出了使Fisher准则函数取极大值的d维向量W^*的计算方法，即仅完成了投影工作。但在类别的判定中还需要确定一个阈值Y_0，当满足下面的不等式时

$$\begin{cases} W^{\mathrm{T}}X \geq Y_0, & X \in \omega_i \\ W^{\mathrm{T}}X < Y_0, & X \in \omega_j \end{cases} \tag{4-61}$$

- 样本的类别属性即可以确定了，一般可采用以下几种方法确定Y_0。一种简单方法是把阈值取为投影后的两个类心的连线中点，即

$$Y_0 = \frac{\tilde{m}_i + \tilde{m}_j}{2} \tag{4-62}$$

- 或者取为以类的频率为权值的投影后两个类心的加权算术平均

$$Y_0 = \frac{N_i\tilde{m}_i + N_j\tilde{m}_j}{N_i + N_j} = \tilde{m} \tag{4-63}$$

- 或者当先验概率$p(\omega_i)$与$p(\omega_j)$已知时，可取为

$$Y_0 = \left[\frac{\tilde{m}_i + \tilde{m}_j}{2} + \frac{\ln[p(\omega_i)/p(\omega_j)]}{N_i + N_j - 2} \right] \tag{4-64}$$

- 在实际工作中，还可通过对Y_0进行逐次修正的方式，选择不同的Y_0值，分别计算它们在训练样本集上的错误率，找到错误率较小的Y_0值。(4-62)算式中只考虑采用均值连线的中点作为阈值点，相当于贝叶斯决策中先验概率 $p(\omega_i)$ 与 $p(\omega_j)$ 相等的情况，而(4-63)与(4-64)则是以不同方式考虑 $p(\omega_i)$ 与 $p(\omega_j)$ 不等的影响，以减小先验概率不等时的错误率。其中(4-63)以不同类别样本的数量 N_i 与 N_j 来估计 $p(\omega_i)$ 与 $p(\omega_j)$ 。

- 综上所述，当 $W^* = S_w^{-1}(m_i - m_j)$ 时， 根据Fisher准则求解向量 W^* 的步骤为
- ① 按类别把来自两类的训练样本集X分成 ω_i 类和 ω_j 类两个子集：Ω_i 和 Ω_j 。
- ②计算各类的类心 $m_k = \frac{1}{N_k}\sum_{X\in\Omega_k} X$ ， $k = i, j$ 。
- ③计算各类的类内离散度矩阵 $S_k = \sum_{X\in\Omega_k}(X - m_i)(X - m_i)^{\mathrm{T}}$ ， $k = i, j$ 。
- ④ 计算总类内离散度矩阵 $S_w = S_i + S_j$ 。
- ⑤ 计算 S_w 的逆矩阵 S_w^{-1} 。
- ⑥ 求解 $W^* = S_w^{-1}(m_i - m_j)$ 。

- 4.2 非线性判别函数
- 4.2.1 非线性判别函数与分段线性判别函数
- 实际应用中，样本在特征空间的分布较为复杂，这些情况下采用线性判别函数通常不能取得令人满意的分类效果。例如，不同类别的分布区域互相交错，这时采用线性判别函数往往会产生大量错分，而采用非线性判别函数则能取得较好的分类效果。
- 以图4.10为例，图中给出了两类物体在二维特征空间中的分布，采用线性判别函数无法取得令人满意的分类效果，采用分段线性判别函数或二次判别函数，效果则会好得多。无论是分段线性判别函数，还是二次判别函数，都属于非线性判别函数。本小节只讨论分段线性判别函数的设计中的一些基本问题。

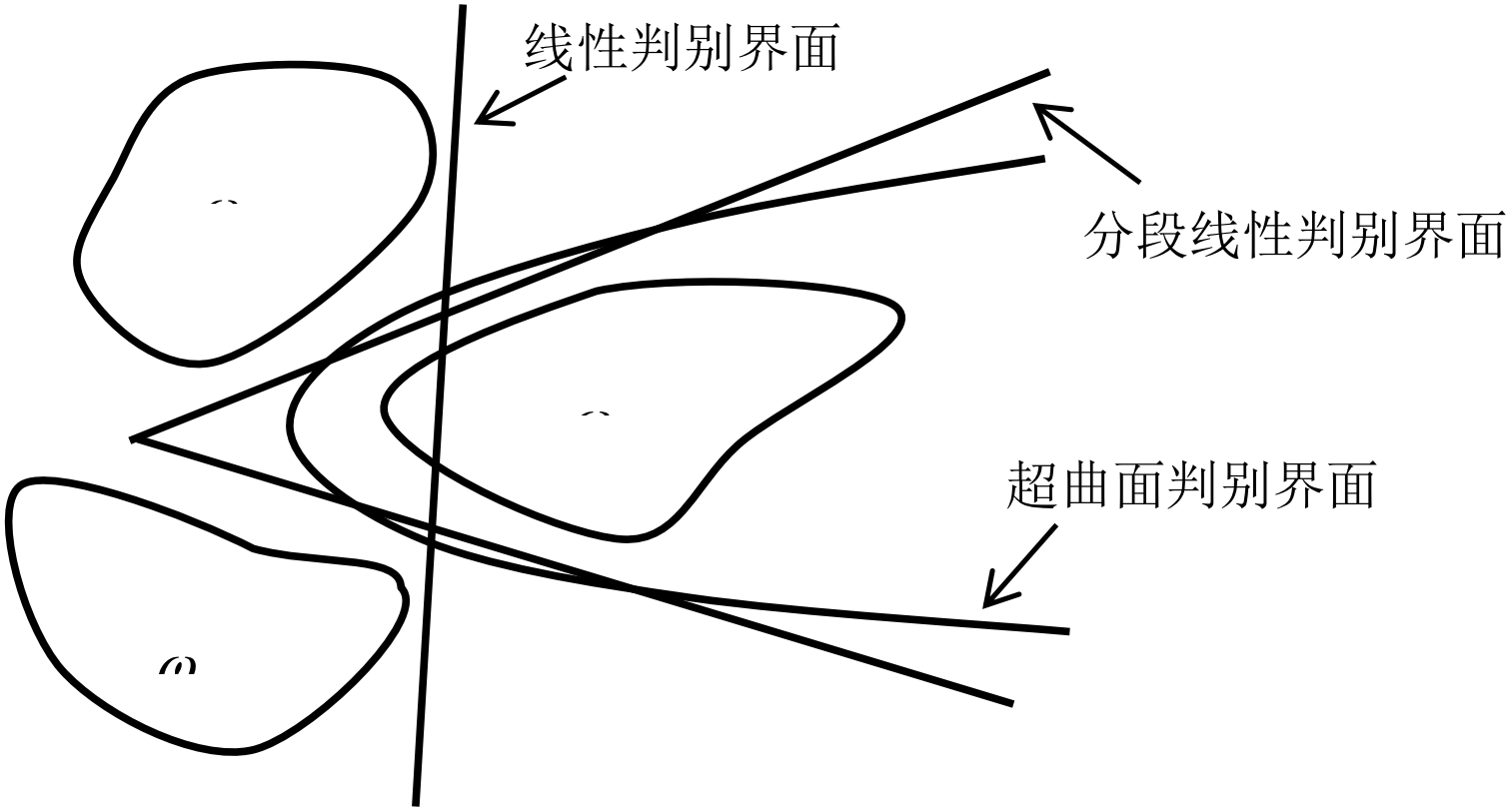

图4.10 不同判别函数

- 在分段线性判别函数的设计中首先要解决确定分段数的问题，这是一个与样本分布有关的问题。
- 若分段数过少，就如图4.10中采用一个线性判别函数(段数为1)的情况，其分类效果必然要差；但分段数又要尽可能少，以免判别函数过于复杂，增加分类决策的计算量。
- 在有些实际问题中，同一类样本可以用若干个子类来描述，子类的数目就可作为确定分段数的依据。但多数情况下样本分布及子类划分均未知，此时需要先采用某种聚类方法，将样本划分为相对密集的子类，然后再设计各分段的线性判别函数。
- 聚类问题将在第五章进行讨论，本章主要讨论在样本分布及子类划分大体已知的情况下，如何设计分段线性判别函数的问题，并着重讨论几种典型的设计原理。

假设样本整体有 $\omega_1, \omega_2, ..., \omega_c$ 共C个模式类，每个类别又可划分为若干个子类，即

$$\omega_i = \omega_i^1 \bigcup \omega_i^2 \bigcup \cdots \bigcup \omega_i^{l_i}, i = 1, 2, \cdots, \ c$$

$$\omega_i^j \bigcap \omega_i^m = \Phi, \ j \neq m$$

分段线性判别函数的一般形式可定义为

$$d_i^l(X) = W_i^{l\mathrm{T}} + w_{i0}^l, \quad l = 1, 2, \cdots, l_i; \quad i = 1, 2, \cdots, c \qquad (4\text{-}65)$$

其中，$d_i^l(X)$表示第i类第l段线性判别函数，l_i为第i类所具有的判别函数个数，W_i^l 与 w_{i0}^l 分别是第l段的权向量与阈值权。

定义第i类的判别函数为 $d_i(X) = \max\limits_{i=1,2,\cdots l} d_i^l(X)$ （4-66）

则相应的判别规则为，如果 $d_j(X) = \max\limits_i d_i(X)$ ，则决策 $X \in \omega_j$ 。

- 至于分类的决策面方程取决于相邻的决策区域，如第i类的第n个子类与第j类的第m个子类相邻，则由它们共同决定的决策面方程为
- $$d_i^n(X) = d_j^m(X) \qquad (4\text{-}68)$$
- 当每类的模式样本在特征空间中呈复杂分布时，使用线性判别函数会产生很差的效果，如果能将它们分割成子集，而每个子集在空间聚集成团，那么子集与子集的线性划分就能取得比较好的效果。因此分段线性判别的主要问题是如何将样本划分成子集的问题，这是第五章着重讨论的内容。