模 式 识 别

（一）

统 计 模 式 识 别

（理论和方法）

下 册

高 林 编写

清华大学自动化系
信息处理及模式识别教研组

1981.7.

第六章　线性判别函数

——模式识别中的实用方法之一

在第四章中，已经证明，Bayes分类器在使错误率或风险为最小的意义上是最佳的。但正如已经指出的，为构造Bayes分类器，必须已知类条件概率密度函数的形式，并利用样本估计未知参数。这虽不是不可能的，但往往非常复杂，且为得到精确结果要求有大量的样本。此外即使我们可以得到这些密度，Bayes分类器还可能难以实现。这是因为Bayes分类器的分界面往往是超曲面，形状复杂，难以构造。再则，即使得到这样的分类器，也往往由于要求的计算时间和存储量都很大而难于实现。

因此，在实际设计中，只提出使错误率或风险最小的要求往往是不够的，它会造成很多上述实际困难。所以，实际上设计者总是针对不同的实际情况，提出不同的设计要求，而力求设计出满足这些要求的"最佳"分类器。由于所提要求的不同，设计结果也将各异。这种设计思想产生的因果关系在日常生活中也常见到。如买东西，若只提出优质的要求，则选择结果可能是高档商品；若只提出廉价要求，则结果可能为低档商品。这种相反的结果是因提出的要求不同造成的，但每种结果对所提要求却是相对'最佳的'。

这种要求在分类器设计中常常表现为一个特定的函数形式，我们称之为准则函数。满足设计要求的'最佳'结果总能在准则函数的极值点上找到。从第三章最优化技术基础中看到，这正是最优化技术所研究的问题。这样我们就可以将模式识别问题利用最优化技术来解决。

实际上，第四章讨论的Bayes分类器就可以看成以错误率或风险为准则函数的最佳分类器，这种分类器使准则函数——错误率或风险——达到最小。常称这种使错误率或风险为最小的分类器，即Bayes分类器为最优分类器，而在其它准则函数下'最佳'的分类器称'次优'的。但从实用的角度看，Bayes分类器的优

实往々由于它使用中的实际困难所暗然失色。这是因为实际问题往々要求判别函数，决策区尽易简单；计算易，存储易尽可能小。众所周知，最简单的判别函数是线性函数；最简单的决策区是超平区。这样在实际中我们宁可牺牲风险最小的优点而获得线性判别函数的好处。也就是说我们可以提出各种准则函数，然后求它们的极值解而得到线性判别函数。这样，如何提出准则函数和解出线性判别函数，将是本章要解决的主要问题。线性判别函数是模式识别中重要的实用方法之一，它是以下三章模式识别中的实用设计方法的第一个分。

6－1 线性判别函数概念

我们将首先研究两类情况的线性判别函数问题，然后在本章最后一节将所有这些概念推广到多类情况。

线性判别函数的一般表达式如

$$g(X) = W^T X + w_0 \qquad (6-1)$$

此处 W 称为权向量，w_0 称为阈值权，一个两类线性分类回可采用下述决策规则：

$$\begin{aligned} &\text{如果} \quad g(X) > 0 \quad \text{决策 } \omega_1 \\ &\qquad\quad g(X) < 0 \quad \text{决策 } \omega_2 \\ &\text{而} \quad\quad g(X) = 0 \quad \text{则拒绝决策} \end{aligned} \qquad (6-2)$$

方程 $g(X)=0$ 定义了一个决策区，它把归类于 ω_1 的矢回归类于 ω_2 的矢分划开来。当 $g(X)$ 为线性函数时，这个决策区便是超平区。如果 X_1 和 X_2 都在决策区上，则

$$W^t X_1 + w_0 = W^t X_2 + w_0 \qquad (6-3)$$

或

$$W^t (X_1 - X_2) = 0 \qquad (6-4)$$

这表明 W 和超平区上任一向量正交，一般说来一个超平区 H 把特征空间分成两个半空间，即对 ω_1 的决策域 R_1 和对 ω_2 的决策域 R_2。因为当 X 在 R_1 中时，$g(X) > 0$，所以决策区的法向量是指向 R_1 中的。所以有时称 R_1 中的任何 X 在 H 的正侧，相应地 R_2

中的任何 X 在 H 的负侧。

判别函数 $g(X)$ 是特征空间某点 X 到超平面的欧氏距离的一种代数度量。若把 X 表示成

$$X = X_p + r\frac{W}{\|W\|} \qquad (6-5)$$

其中 X_p X 在 H 上投影向量

r X 到 H 垂直距离

$W/\|W\|$ W 方向单位向量

将 (6-5) 代入 (6-1)

$$\begin{aligned} g(X) &= W^T\left(X_p + r\frac{W}{\|W\|}\right) + w_0 \\ &= W^T X_p + w_0 + r\frac{W^T W}{\|W\|} \\ &= r\|W\| \end{aligned} \qquad (6-6)$$

或者

$$r = \frac{g(X)}{\|W\|} \qquad (6-7)$$

若 X 为原点 X_0，则

$$g(X) = W^T X_0 + w_0 = w_0 \qquad (6-8)$$

将 (6-8) 代入 (6-7) 就得到从原点到超平面 H 的距离

$$r_0 = \frac{w_0}{\|W\|} \qquad (6-9)$$

如果 $w_0 > 0$ 则原点在 H 的正侧，若 $w_0 < 0$ 则原点在 H 的负侧。若 $w_0 = 0$ 则 $g(X)$ 具有齐次形式 $W^T X$，且超平面 H 过原点。图 6-1 对这些结果作了几何解释。

总之，利用线性判别函数的决策是由一个超平面把特征空间分割开来的，超平面的方向由法线方向 W 决定，它的位置由阈值 w_0 决定。判别函数 $g(X)$ 正比于 X 点到超平面的代数距离

（带正负号）。当 $\overline{X}$ 在正侧时 $g(\overline{X})>0$，在负侧时 $g(\overline{X})<0$。

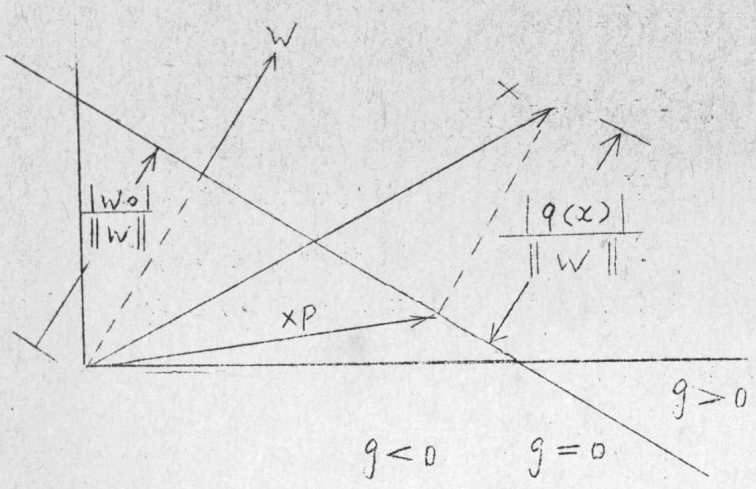

图 3-1 线性决策界面 $g(x)=w^T x+w_0=0$

至此我们已了解了线性判别函数的构造，那么所谓设计线性分类器，就是寻找线性判别函数（6-1），使它满足所提出的设计要求，也就是使准则函数达到极值点。从（6-1）看出，求知的仅是权向量 $\overline{w}$ 和阈值权 w_0，这样设计问题就变成寻找最佳 $\overline{w}$ 和 w_0 的问题了，所谓最佳是指最好的满足设计要求之点。

在实际使用中，有时我们还希望将（6-1）式的线性判别函数再简化一些，即企图将（6-1）的非齐次线性函数写成齐次形式。对（6-1），令 $\overline{X}$ 和 $\overline{w}$ 为 d 维向量，即

$$\overline{X}=\begin{bmatrix}x_1\\x_2\\\vdots\\x_d\end{bmatrix} \qquad \overline{w}=\begin{bmatrix}w_1\\w_2\\\vdots\\w_d\end{bmatrix}$$

即（6-1）式可表示成

$$g(\overline{X})=\overline{w}^T\overline{X}+w_0=\sum_{j=1}^{d} w_j x_j+w_0 \qquad (6-10)$$

（6-10）是一个 $\overline{X}$ 空间的非齐次线性函数。现在如果我们令

$$y=\begin{bmatrix}x_1\\x_2\\\cdots\\x_d\\1\end{bmatrix}=\begin{bmatrix}X\\1\end{bmatrix}\qquad a=\begin{bmatrix}w_1\\w_2\\\cdots\\w_d\\w_0\end{bmatrix}=\begin{bmatrix}w\\w_0\end{bmatrix}$$

则 $g(X)$ 可以写成

$$g(X)=\begin{bmatrix}w^T & w_0\end{bmatrix}\begin{bmatrix}X\\1\end{bmatrix}=a^Ty \qquad (6-11)$$

此时 $g(X)$ 为 $d+1$ 维 Y 空间的齐次线性函数。从非齐次到齐次的变换可看成从 d 维 X 空间到 $d+1$ 维 Y 空间的映射，在 $d+1$ 维 Y 空间决策面 $\hat{H}$ 的方程为

$$a^Ty=0 \qquad (6-12)$$

显然在 Y 空间阈值权为零，所以超平面 $\hat{H}$ 过原点。Y 空间任何点 y_i 到 $\hat{H}$ 的距离可根据（6-7）求出为

$$\hat{r}=\frac{g(X)}{\|a\|}=\frac{a^Ty}{\|a\|} \qquad (6-13)$$

从 X 空间到 Y 空间只增加一维，映射方法在数学上也是很简单的，而给我们带来的好处是把设计中寻找最佳 w 和 w_0 的问题简化为仅仅寻找 a 的问题了。我们将（6-11）给出的齐次线性判别函数称为广义线性判别函数，加权 a 称为广义权向量。在以后的线性分类器设计中我们经常使用广义线性判别函数简化问题的处理。

6-2 Bayes 线性判别函数和最小错误率线性判别函数

6-2-1 Bayes 线性判别函数

在第四章中我们讨论了各种 Bayes 分类器，它使错误概率或风险为最小，但它们的判别函数一般是非线性的，这给计算上带来了复杂性。现在要问这两者是否能统一起来呢？也就是说是否可以设计出即使错误率为最小（具有最优性），又使判别函数

为线性（只有简单性）的两者兼顾的分类回呢？答案实际已包含在第四章中了。这就是仅在某些特殊条件下是可以做。现将这些特殊条件下的Bayes线性判别函数再概括如下：

1、条件：正态分布，各特征是统计独立的且每一特征都具有同一方差 σ^2 的情况。也就是

$$\Sigma_i = \sigma^2 I \qquad i = 1, 2 \quad *$$

的情况，此时各类判别函数为

$$g_i(X) = W_i X + w_{i0} \qquad i = 1, 2 \qquad (6-14)$$

其中

$$W_i = \frac{1}{\sigma^2} \mu_i$$

$$w_{i0} = -\frac{1}{2\sigma^2} \mu_i^T \mu_i + \ln P(\omega_i)$$

总的判别函数可写成

$$g(X) = g_1(X) - g_2(X) \qquad (6-15)$$

2、条件：正态分布且各类协方差矩阵相等即：

$$\Sigma_i = \Sigma \qquad i = 1, 2 \qquad (6-16)$$

各类判别函数为

$$g_i(X) = W_i^T X + w_{i0} \qquad i = 1, 2 \qquad (6-17)$$

其中

$$W_i = \Sigma^{-1} \mu_i$$

$$w_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln P(\omega_i)$$

总的判别函数仍为 $g(X) = g_1(X) - g_2(X)$。

此外在其它一些情况下也还有可能得到Bayes线性判别函数，它具有最小错误率及线性两者兼顾之故。但应指出的是这只是一些特殊情况下的结果，一般是不行的。

* 这里 $i = 1, 2$ 是因为现在我们仅考虑两类情况，实际上这些判别函数也适用于多类情况。

6-2-2 最小错误率线性判别函数

现实我们已经知道，从理论上说，Bayes分类器是使错误率最小的最优分类器；而就实现而言，线性分类器则是最简单的。且两者只在少数情况下可以兼顾，比如等协方差分布。但在实际应用中，类似等协方差正态分布的假设往往是不适用的，而做非等协方差正态分布假设则较为合理。从第四章的理论可知，这种情况下的Bayes分类器当然是非线性的（图6-2）。

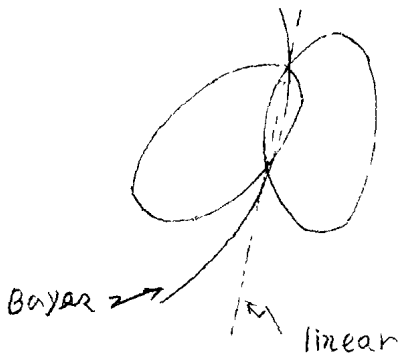

图6-2 非等协方差正态分布时的Bayes决策器和最小错误率线性决策器。

那么在非等协方差正态分布或甚至非正态分布条件下，我们能否牺牲Bayes分类器最优的特点，而采用线性分类器分类，同时我们在线性条件下，使错误率为最小的分类器呢？本节将解决这一问题。

我们首先给定线性判别函数的形式及决策规则为：

$$g(X)=W^TX+w_0\gtrless 0\longrightarrow X\in\begin{cases}w_1\\w_2\end{cases}\qquad(6-18)$$

若X为正态分布，由于$g(X)$是X的函数，所以$g(X)$也是正态分布；若X不是正态分布，当X的维数d较大时，根据中心极限定理，$g(X)$可以是接近正态分布的。因此，最小错误率线性判别函数设计的基本思想是把判别函数$g(X)$看成随机变量，当X是正态或接近正态分布时，我们可以把错误率ε写成W和w_0的函数，这样就可以用数学分析中的极值理论寻找使$\hat{\varepsilon}=\min\varepsilon$的$W$和$w_0$。

为达到这一目的，首先要求出随机变量$g(X)$的期望μ_i和差σ_i^2。

$$\eta_i = E\{g(X)/\omega_i\} = E\{W^T X + w_0/\omega_i\}$$
$$= E\{W^T X/\omega_i\} + w_0 = W^T E\{X/\omega_i\} + w_0$$
$$= W^T \mu_i + w_0 \qquad (6-19)$$

其中 μ_i 为 ω_i 类随机变量 X 的期望

$$\sigma_i^2 = Var\{g(X)/\omega_i\} = E\{(g(X)-\eta_i)(g(X)-\eta_i)^T/\omega_i\}$$
$$= W^T E\{(X-\mu_i)(X-\mu_i)^T/\omega_i\} W$$
$$= W^T \Sigma_i W \qquad (6-20)$$

其中 Σ_i 为 ω_i 类随机变量 X 的协方差矩阵

下面我们就可以改为于将错误率 ε 表成 W 和 w_0 的函数形式了。已知错误率可以写成

$$\varepsilon = P(\omega_1)\varepsilon_1 + P(\omega_2)\varepsilon_2 \qquad (6-21)$$

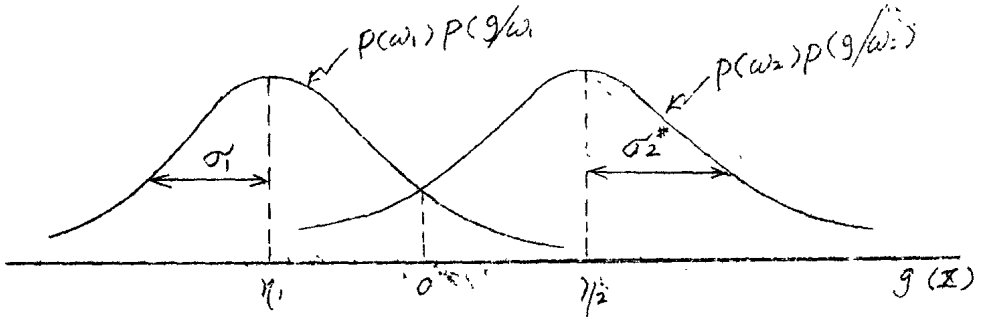

图 6-3 判别函数 $g(X)$ 的分布

其中

$$\varepsilon_1 = \int_0^\infty p(g/\omega_1)\,dg = \int_0^\infty \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left[-\frac{(g-\eta_1)^2}{2\sigma_1^2}\right] dg$$

$$(6-22)$$

(6-22) 中令 $\xi = \frac{g-\eta_1}{\sigma_1}$ $\quad dg = \sigma_1 d\xi$ 则

参 Fukunaje "Introductim to statistical Patlern Recognition" 4.2

$$\varepsilon_1=\int_{-\eta_1/\sigma_1}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\xi^2}{2}\right)d\xi \tag{6-23}$$

同理可推出：

$$\varepsilon_2=\int_{-\infty}^{-\eta_2/\sigma_2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\xi^2}{2}\right)d\xi \tag{6-24}$$

将(6-23)(6-24)代入(6-21)得

$$\varepsilon=p(\omega_1)\int_{-\eta_1/\sigma_1}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\xi^2}{2}\right)d\xi + p(\omega_2)\int_{-\infty}^{-\eta_2/\sigma_2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\xi^2}{2}\right)d\xi \tag{6-25}$$

(6-25)中的ε是ξ的函数；ξ是η_i，σ_i的函数；$\eta_i\sigma_i$又是$\overline{w}$，w_0的函数，因此ε也是$\overline{w}$和w_0的函数。这样从(6-25)中求出使$\varepsilon=\min\varepsilon$的$\overline{w}$和$w_0$的问题就完全等价于一个函数极值问题了，即将(6-25)求对$\overline{w}$和w_0的偏导数并使其等于零。从而解出$\overline{w}$和w_0。

$$\frac{\partial\varepsilon}{\partial\overline{w}}=-\frac{P(\omega_1)}{\sqrt{2\pi}}\exp\left(-\frac{\eta_1^2}{2\sigma_1^2}\right)\frac{\partial}{\partial\overline{w}}\left(-\frac{\eta_1}{\sigma_1}\right) + \frac{P(\omega_2)}{\sqrt{2\pi}}\exp\left(-\frac{\eta_2^2}{2\sigma_2^2}\right)\frac{\partial}{\partial\overline{w}}\left(-\frac{\eta_2}{\sigma_2}\right) \tag{6-26}$$

(6-26)中的偏导数项

$$\frac{\partial}{\partial\overline{w}}\left(\frac{\eta_i}{\sigma_i}\right)=\frac{\sigma_i\partial\eta_i/\partial\overline{w}-\eta_i\partial\sigma_i/\partial\overline{w}}{\sigma_i^2} \tag{6-27}$$

$$\because\ \frac{\partial\eta_i}{\partial\overline{w}}=\frac{\partial}{\partial\overline{w}}(\overline{w}^T\mu_i+w_0)=\mu_i \tag{6-28}$$

$$\frac{\partial\sigma_i}{\partial\overline{w}}=\frac{\partial}{\partial\overline{w}}(\overline{w}^T\Sigma_i\overline{w})^{\frac{1}{2}}=\frac{1}{\sigma_i}\Sigma_i\overline{w} \tag{6-29}$$

(6-28)(6-29)代入(6-27)

$$\frac{\partial}{\partial w}\left(\frac{\eta_1}{\sigma_1}\right)=\frac{\sigma_1^2\mu_1-\frac{\eta_1}{\sigma_1^2}\Sigma_1 w}{\sigma_1^4}=\frac{1}{\sigma_1}\left(\mu_1-\frac{\eta_1}{\sigma_1^2}\Sigma_1 w\right)$$

(6-30)

将(6-30)代入(6-26)

$$\frac{\partial}{\partial w}=\frac{P(\omega_1)}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{\eta_1^2}{2\sigma_1^2}\right)\left(\mu_1-\frac{\eta_1}{\sigma_1^2}\Sigma_1\hat{w}\right)$$

$$-\frac{P(\omega_2)}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{\eta_2^2}{2\sigma_2^2}\right)\left(\mu_2-\frac{\eta_2}{\sigma_2^2}\Sigma_2\hat{w}\right)=0$$

(6-31)

再将ε对w_0求偏导

$$\frac{\partial\varepsilon}{\partial w_0}=-P(\omega_1)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\eta_1^2}{2\sigma_1^2}\right)\frac{\partial}{\partial w_0}\left(-\frac{\eta_1}{\sigma_1}\right)$$

$$+P(\omega_2)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\eta_2^2}{2\sigma_2^2}\right)\frac{\partial}{\partial w_0}\left(-\frac{\eta_2}{\sigma_2}\right)$$

$$=\frac{P(\omega_1)}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{\eta_1^2}{2\sigma_1^2}\right)-\frac{P(\omega_2)}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{\eta_2^2}{2\sigma_2^2}\right)$$

$$=0 \qquad (6-32)$$

如将(6-31)(6-32)适当化简并联立得

$$\mu_2-\mu_1=\left[\frac{\eta_2}{\sigma_2^2}\Sigma_2-\frac{\eta_1}{\sigma_1^2}\Sigma_1\right]\hat{w} \qquad (6-33)$$

$$\frac{P(\omega_1)}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{\eta_1^2}{2\sigma_1^2}\right)=\frac{P(\omega_2)}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{\eta_2^2}{2\sigma_2^2}\right)$$

(6-34)

原则上可解得$\hat{w}$和$\hat{w}_0$，它们分别是使错误率ε达到最小的线性判别函数的权向量和阈值权。遗憾的是其中η_i，σ_i^2是w和w_0的函数，因此很难得到方程组的闭式解。一般来说，我们可以用迭代法克服这一困难，但是值得指出的是既使用迭代法求解也因计算量大而难于实现。因此，虽然可以说这种考虑在一般正态分布或高维条件下可以设计最小错误率线性分类[*]。但由于实际的

困难，使得它们的理论价值远远高于实际使用价值了。

6－3　Fisher 线性判别函数

在上节中，我们把错误率 ε 看成准则函数，并把它写成 ω_1，ω_2 的函数，求使 $\hat{\varepsilon}=\min\varepsilon$ 时的 $\hat{\omega}_1$ 和 $\hat{\omega}_2$，从而得到最小错误率线性判别函数 $g(X)=\hat{\omega}^T X+\hat{\omega}_0$。但问题是由于计算上的困难，使这种方法的实际使用价值不大。但它却给我们一个重要启示，能否不直接用错误率 ε 作为准则函数，而是找一个能从侧面反映分类错误情况，且又比较简单易于计算的函数了作为准则函数呢？能否这样找出 $J=\min J$ 时的 $\hat{\omega}$ 和 $\hat{\omega}_0$ 作为线性判别函数的权向量呢？回答将是肯定的，且本章下面将要讨论的设计线性判别函数的一些实用算法都是基于这样一种考虑，这里首先讨论一种过度性方法，即 Fisher 判别法，它是判别分析中十分著名的经典方法之一。

用统计法解决模式识别问题时，一再碰到的问题之一就是"维数"问题，往往在低维空间解析上或计算上行得通的办法在高维空间就行不通。因此，降低维数将是处理实际问题的关键。关于这一点我们还将在第九章特征选择和抽取中详细讨论，这里是从线性判别的角度给出一种降低维数的方法。

我们可以考虑把 d 维空间的样本投影到一直线上，形成一维空间，即把维数压缩到一维。这在数学上总是容易办到的。可是，即使样本在 d 维空间里形成分得很开的许多紧凑的集群，但如把它们投影到一根任意的直线上就可能使几类样本混在一起而变得无法识别。但有时将直线转动几个角度，就可能找到某一个方向，使在这方向的直线上，样本的投影能较好地分开。问题是如何结合实际情况，找到这条最佳投影线，也就是使样本分得最开，从而最易于分类的投影线呢？！这将是 Fisher 法所要解决的基本问题。（图 6－3）

首先我们研究从 d 维空间到一维空间的一般数学转换方法。

假设有一集合包括 d 维样本 $X_1, X_2, \cdots, X_N$ 共 N 个，其中 N_1 个属于 ω_1 类的样本记为子集 χ_1；N_2 个属于 ω_2 类的样本记为子集 χ_2。若对 X_i 的分量作线性变换，可得标量

$$y_i = w^t X_i^* \qquad i = 1, 2, \cdots, N_i \qquad (6-35)$$

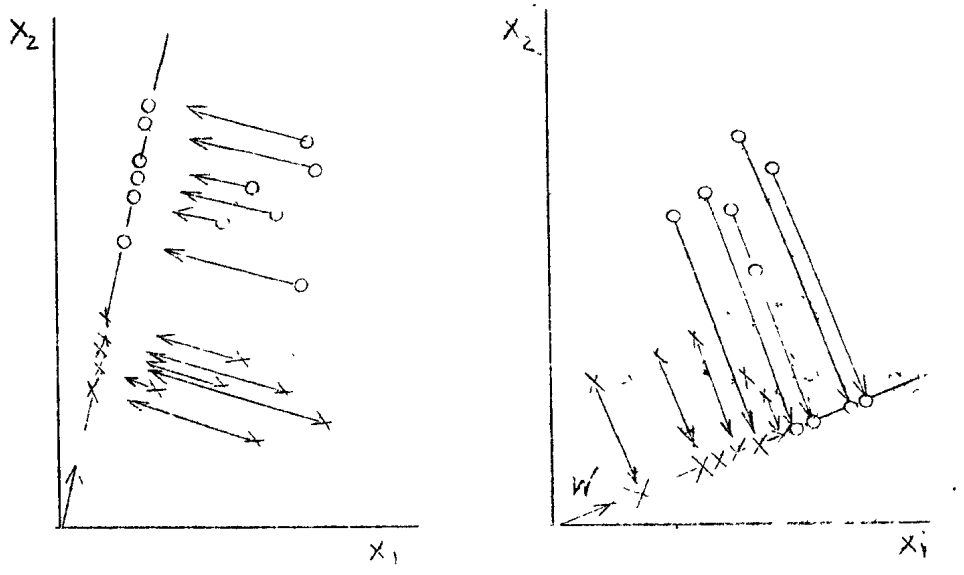

图 6-3 样本在一直线上的投影

这样得到相应于 N 个样本的集合，它们是一维样本，并可分为两个子集 $\mathcal{Y}_1$ 和 $\mathcal{Y}_2$。从几何上看，如果 $\|w\| = 1$，则每个 y_i 就是相对应的 X_i 到其方向为 w 的直线上的投影。实际上 w 的绝对值是无关紧要的，因它仅使 y 改变一个比例而已，重要的是寻找 w 的方向。w 方向选择的不同，将使两类样本在 w 直线投影后的可分离程度不同，从而直接影响识别效果。因此前述所谓寻找最佳投影线的问题，在数学上就是寻找最佳变换向量 $\hat{w}$ 的问题。图 6-3 说明在二维空间里选择二个不同的 w 值所产生的不同投影效果。

在定义 Fisher 准则函数之前，我们先定义几个必要的基本概念。

---

* 注意，这里的 y_i 不是前已 6-1 节所述广义样本 y.

(一) d维X空间

1、d维X空间样本均值 m_i

$$m_i = \frac{1}{N_i}\sum_{X\in \chi_i} X \qquad i=1,2 \qquad (6-36)$$

2、d维X空间 ω_i 类样本类内离散度矩阵 S_i 和总离散度矩阵 S_w

$$S_i = \sum_{X\in \chi_i} (X-m_i)(X-m_i)^T \qquad i=1,2 \qquad (6-37)$$

$$S_w = S_1 + S_2 \qquad (6-38)$$

3、d维X空间样本类内离散度矩阵 S_B

$$S_B = (m_1 - m_2)(m_1 - m_2)^T \qquad (6-39)$$

其中 S_w 是对称半正定矩阵，且当 $N>d$ 时通常是非奇异的。S_B 也是对称半正定矩阵，因其为两个向量的外积，所以 S_B 是奇异的。

(二) 一维Y空间

1、一维Y空间投影点样本均值 $\tilde{m}_i$

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y\in \Omega_i} y \qquad i=1,2 \qquad (6-40)$$

2、一维Y空间投影点 ω_i 类样本类内离散度 $\tilde{S}_i^2$

$$\tilde{S}_i^2 = \sum_{y\in \Omega_i} (y-\tilde{m}_i)^2 \qquad (6-41)$$

3 一维Y空间总类内离散度 $\tilde{S}_w$

$$\tilde{S}_w = \tilde{S}_1^2 + \tilde{S}_2^2 \qquad (6-42)$$

有了上述基本的数学量，现在我们可以开始定义 Fisher 准则函数了。我们希望投影后在一维Y空间，两类样本尽可能分开，也就是说希望投影后两类均值之差 $(\tilde{m}_1 - \tilde{m}_2)^2$ 越大越好。同时我们还希望每类投影样本尽量密集，上已定义的类内离散度 $\tilde{S}_w = \tilde{S}_1^2 + \tilde{S}_2^2$ 显然反映了每类内样本密集程度，因此我们希望 $\tilde{S}_w$ 越小越好。这样我们可以定义 Fisher 准则函数为

$$J(w)=\frac{(\tilde{m}_1-\tilde{m}_2)^2}{\tilde{S}_1^2+\tilde{S}_2^2} \qquad (6-43)$$

我们的目的是寻找使 $J(w)$ 的分子尽可能地大，而分母尽可能地小，也就是使 $J(w)$ 尽可能大的 $\hat{w}$ 值做为投影方向。但(6-43)的 $J(w)$ 並不显含 w，所以我们必须根次法将(6-43)的 $J(w)$ 变为 w 的显函数，为此必须推导其中各号与 w 的关系。首先看 $\tilde{m}_i$ 与 w 的关系、

$$\tilde{m}_i=\frac{1}{N_i}\sum_{y\in\mathcal{Y}_i} y \qquad y=\frac{1}{N_i}\sum_{x\in\mathcal{X}_i} w^T x$$

$$=w^T\left(\frac{1}{N_i}\sum_{x\in\mathcal{X}_i} x\right)=w^T m_i \qquad (6-44)$$

推导中用了(6-40)和(6-35)的结果。这样(6-43)分子

$$(\tilde{m}_1-\tilde{m}_2)^2=(w^T m_1-w^T m_2)^2$$

$$=w^T(m_1-m_2)(m_1-m_2)^T w$$

$$=w^T S_B w \qquad (6-45)$$

这样(6-45)明显的表示出准则函数 $J(w)$ 的分子与 w 的关系。现在再来考察 $J(w)$ 的分母与 w 的关系

$$\tilde{S}_i^2=\sum_{y\in\mathcal{Y}_i}(y-\tilde{m}_i)^2$$

$$=\sum_{x\in\mathcal{X}_i}(w^T x-w^T m_i)$$

$$=w^T\left\{\sum_{x\in\mathcal{X}_i}(x-m_i)(x-m_i)^T\right\}w$$

$$=w^T S_i w \qquad (6-46)$$

且

$$\tilde{S}_1^2+\tilde{S}_2^2=w^T(S_1+S_2)w$$

$$=w^T S_w w \qquad (6-47)$$

以上推导利用了上述各定义的数学关系。(6-47)将 $J(w)$ 分母表成了 w 的函数。现在我们将(6-45)(6-47)代入Fisher准则函数(6-43)从而可得

$$J(W)=\frac{W^T S_B W}{W^T S_W W} \tag{6-48}$$

不同的任务是寻找 $J(W)$ 取最大值时的 W，(6-48) 的 $J(W)$ 是数学物理中著名的广义 Rayleigh 比，可以用 Lagrange 乘子法求极值解。令分母等于非零常数，即

$$W^T S_W W = C \neq 0 \tag{6-49}$$

到求 Lagrange 函数

$$L(W,\lambda)=W^T S_B W+(C-W^T S_W W)\lambda \tag{6-50}$$

其中 λ 为 Lagrange 乘子，将 (6-50) 对 W 求偏导数

$$\partial L(W,\lambda)/\partial W = S_B W-\lambda S_W W \tag{6-51}$$

使 (6-51) 等于零，W 就是 $J(W)$ 极值解，

$$S_B \hat{W}-\lambda S_W \hat{W}=0 \tag{6-52}$$

即

$$S_B \hat{W}=\lambda S_W \hat{W} \tag{6-53}$$

因 S_W 非异，(6-53) 两边左乘 S_W^{-1} 得

$$S_W^{-1} S_B \hat{W}=\lambda \hat{W} \tag{6-54}$$

(6-54) 为求一般矩阵 $S_W^{-1}S_B$ 特征值问题。但在我们这个特殊情况下，利用 (6-39) S_B 的定义，(6-54) 左边 $S_B\hat{W}$ 可写成

$$\begin{aligned} S_B \hat{W} &= (\mu_1-\mu_2)(\mu_1-\mu_2)^T \hat{W} \\ &= (\mu_1-\mu_2) R \end{aligned} \tag{6-55}$$

其中 $R=(\mu_1-\mu_2)^T W$ 为一常数

$S_B W$ 总是在向量 $(\mu_1-\mu_2)$ 方向上，因此没有必要求出 $S_W^{-1}S_B$ 的特征值和特征向量。由于我们的目的是寻找最佳投影向 $\hat{W}$，而 $\hat{W}$ 的比例因子对此并无影响，因此从 (6-54) 可得

$$\lambda \hat{W}=S_W^{-1}(S_B \hat{W})$$

$$-S_w^{-1}(m_1-m_2)R \qquad (6-56)$$

从而

$$\hat{w}=\frac{R}{\lambda}S_w^{-1}(m_1-m_2) \qquad (6-57)$$

忽略比例因子 R/λ，(6-57)可写成

$$\hat{w}=S_w^{-1}(m_1-m_2) \qquad (6-58)$$

$\hat{w}$就是使$J(w)$最大的Fisher解，也就是d维X空间向一维Y空间的最佳投影方向。有了$\hat{w}$，利用(6-35)式

$$y_i=\hat{w}^T x_i \qquad i=1,2,\cdots,N_i$$

就可以把各类d维样本x_i投影到一维Y空间，变成一维样本y_i。它们保证在Y空间类间尽可能分离，类内尽可能紧密。这样我们就把难以考虑的d维问题转化为容易考虑的一维问题。这实际上是一种多维到一维的映射，理论上它所产生的错误率一般不能达到最小。有人曾对两类正态非等协方差分布情况下Bayes分类器，最小错误率线性分类器和Fisher线性判别函数下做分类器的错误率做过理论计算，结果是

Bayes分类器　　$\varepsilon=1.9\%$

最小错误率线性分类器　$\varepsilon=5\%$

Fisher判别分类器　　$\varepsilon=6.5\%$

这就说明Fisher判别理论上不能产生最佳解，但一般来说，我们宁愿牺牲一些理论上的最佳性能去换得工作在一维下的好处。但可以证明当条件密度$P(x/\omega_i)$是多维正态，且协方差矩阵相同，均为Σ时，Fisher判别就是Bayes分类器，Fisher解就是使错误率为最小的最佳解，我们把它的证明留给读者。

最后，我们尚需再强调指出，Fisher线性判别函数本身并不能决策分类，它只是一种压缩维数的方法。使在原空间里行不通