第2-3节　支持向量机基础

—— Fisher线性分类器

判别分析问题，即：根据历史上划分类别的有关资料和某种最优准则，确定一种判别方法，判定一个新的样本归属哪一类。

例如：在天气预报中，我们有一段较长时间关于某地区每天气象的记录资料(晴阴雨、气温、气压、湿度等)，现在想建立一种用连续五天的气象资料来预报第六天是什么天气的方法。这些问题都可以应用判别分析方法予以解决。

四、Fisher 线性判别函数

4.1 判别函数定义

对一组具有d维特征的数据$X_1, X_2, \ldots, X_n$，找出根据某种原则制定的一个判别函数，将上述样本点尽可能地区别开来，即该评判函数在该原则下最优。

然后，利用该判别函数能对同样 d 维的一个新样本数据进行预测，判定这个新样本归属于哪一类。

对于上述具有d维特征的一个数据X_i（$i=1, 2, \ldots, n$），其向量表达形式为：

$$X_i = [x_{i1}, x_{i2}, \cdots, x_{id}]^T$$

假设判别的目的，就是要将X_i分入两类，其类别参数可以表示为：

两类情况，$\Omega = \{\omega_1, \omega_2\}$

例如：设某种产品的市场情况有“畅销”，“滞销”两种，我们要预测产品在一个时期是“畅销”还是“滞销”。

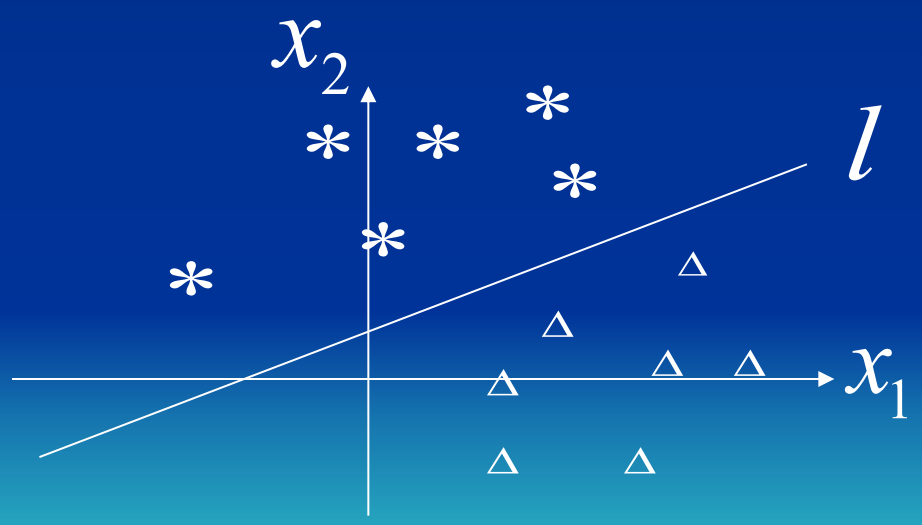
根据过去的销售情况可知，该产品销路好坏与价格有关，也和市民的收入有关，因此可以用产品的价格和市民的收入这两个量去预测该产品的销路的好坏。

以 x_1 代表产品的价格，以 x_2 表示市民的收入。现在假定调查了 n 个时期，得到 n 组数据。这 n 组数据反应的有畅销的也有滞销的，不妨设有 r 组畅销，l 组滞销 ($l = n - r$)，则可将 n 组数据分组如下：

畅销组 $\begin{cases} \left(x_{11}^0, x_{12}^0\right) \\ \left(x_{21}^0, x_{22}^0\right) \\ \vdots \\ \left(x_{r1}^0, x_{r2}^0\right) \end{cases}$

滞销组 $\begin{cases} \left(x_{11}^1, x_{12}^1\right) \\ \left(x_{21}^1, x_{22}^1\right) \\ \vdots \\ \left(x_{l1}^0, x_{l2}^0\right) \end{cases}$

将这 n 组数据标在平面上，以“Δ”表示畅销组所对应的点，以“*”表示滞销数据对应的点。若能得到如图所示的点聚图，说明产品畅销时期的数据和滞销时期的数据有较为明显的区别，可以线性区分，我们就可以某个线性判别函数做直观的判断。



例如，若某个预测时期的数据对应的点为“Δ”，则我们应判断这一时期为畅销期；若对应点为“*”，则应判断这一时期为滞销期。因此，在预测时，重要的问题是要找出分界线 l，其方程为：

$$c_0 + c_1 x + c_2 x = 0$$

使得当第 i 个时期的数据 (x_{i1}, x_{i2}) 代入上式左端，若有 $c_0+c_1 x+c_2 x > 0$，即 $c_1 x+c_2 x > -c_0$，则预测该时期为畅销期；

若有 $c_0+ c_1 x+c_2 x < 0$，即 $c_1 x+c_2 x < -c_0$，则预测该时期为滞销期。

这种预测分析的方法就是线性判别分析法。在利用这种方法时必须要求畅销期的数据和滞销期的数据之间有一条较明显的分界线。

令： $y = c_1 x_1 + c_2 x_2$

称此函数为线性判别函数，称 $y_0 = - c_0$ 为临界值。

进行判别分析就是要在某种最优准则下，确定线性判别函数的系数 c_1，c_2 以及临界值 $- c_0$ 的值。

多维样本的线性判别函数：

假设预测因子有 d 个指标，即 d 个维度，有 n 组观察或调查得到的数据。判别分析就是要根据这些数据，在适当的判别准则下，确定判别函数：

$$y = c_1 x_1 + c_2 x_2 + \ldots\ldots + c_d x_d$$

并找出临界值 $y_0 = -c_0$。

我们将要判别的两组分别标记为 A 和 B (如 A 代表畅销，B 代表滞销)。对于 d 个判别指标，不妨设组 A 有 s 组数据，组 B 有 t 组数据，$n = s + t$，现将数据分组如下：

$$\begin{cases} \left(x_{11}^{0}, x_{12}^{0}, \cdots, x_{1d}^{0}\right) \\ \left(x_{21}^{0}, x_{22}^{0}, \cdots, x_{2d}^{0}\right) \\ \vdots \\ \left(x_{s1}^{0}, x_{s2}^{0}, \cdots, x_{sd}^{0}\right) \end{cases}$$

组A的数据

$$\begin{cases} \left(x_{11}^{1}, x_{12}^{1}, \cdots, x_{1d}^{1}\right) \\ \left(x_{21}^{1}, x_{22}^{1}, \cdots, x_{2d}^{1}\right) \\ \vdots \\ \left(x_{t1}^{1}, x_{t2}^{1}, \cdots, x_{td}^{1}\right) \end{cases}$$

组B的数据

下面反过来思考整个问题，假定用：

$$y = c_1 x_1 + c_2 x_2 + \cdots\cdots + c_d x_d$$

作为判别函数，则组 A 的数值对应的判别值为：

$$\begin{cases} y_1^0 = c_1 x_{11}^0 + c_2 x_{12}^0 + \ldots + c_d x_{1d}^0 \\ y_2^0 = c_1 x_{21}^0 + c_2 x_{22}^0 + \ldots + c_d x_{2d}^0 \\ \vdots \\ y_s^0 = c_1 x_{s1}^0 + c_2 x_{s2}^0 + \ldots + c_d x_{sd}^0 \end{cases}$$

组B的数值对应的判别值为：

$$\begin{cases} y_1^1 = c_1 x_{11}^1 + c_2 x_{12}^1 + \ldots + c_d x_{1d}^1 \\ y_2^1 = c_1 x_{21}^1 + c_2 x_{22}^1 + \ldots + c_d x_{2d}^1 \\ \vdots \\ y_t^1 = c_1 x_{t1}^1 + c_2 x_{t2}^1 + \ldots + c_d x_{td}^1 \end{cases}$$

计算：　$\overline{y}^0 = \frac{1}{s}\sum_{i=1}^{s} y_i^0$　　$\overline{y}^1 = \frac{1}{t}\sum_{i=1}^{t} y_i^1$

即 $\overline{y}^0$ 为组 A 的代表，　$\overline{y}^1$ 为组 B 的代表。

我们通过判别值 y 来进行判别，为使组 A 同组 B 之间有明显的区别，自然希望它们的代表值之间的差距越大越好。即：

(1) $(\bar{y}^0-\bar{y}^1)^2$ 越大越好；

又 $y_1^0, y_2^0, \cdots, y_s^0$ 同属于组 A，$y_1^1, y_2^1, \cdots, y_t^1$ 同属于组 B，它们之间的差距越小越好，即：

(2) $\sum_{i=1}^{s}(y_i^0-\bar{y}^0)^2+\sum_{i=1}^{t}(y_i^1-\bar{y}^1)^2$ 越小越好。

综合上述(1)，(2)就是Fisher提出的最优判别准则。

4.2 Fisher最优判别准则：

$$\max L(c_1, c_2, \ldots, c_d) = \frac{(\bar{y}^0 - \bar{y}^1)^2}{\sum_{i=1}^{s}(y_i^0 - \bar{y}^0)^2 + \sum_{i=1}^{t}(y_i^1 - \bar{y}^1)^2}$$

最优判别函数的系数 $c_1, c_2, \ldots, c_d$ 为函数 $L(c_1, c_2, \ldots, c_d)$ 的极大值点。

由微分学方程可知，

$$\frac{\partial L(c_1, c_2, \quad c_d)}{\partial c_i} = 0, \qquad j = 1, \ 2, \qquad , \ d$$

为函数 $L\,(\,c_1, c_2, \ldots, c_d\,)$ 的极值解。

定义：S_1，S_2代表两类的类内差矩阵；

m_1，m_2分别代表两类的均值向量。

则总的类内差 (类内散度) 矩阵 S_w 为：

$$S_w = S_1 + S_2$$

以及类间差 (类间散度) 矩阵 S_b 为：

$$S_b = (m_1 - m_2) \cdot (m_1 - m_2)^{\mathrm{T}}$$

代入Fisher最优判别式，可得：

$$L = \frac{c^T S_b c}{c^T S_w c}$$

这就是Fisher线性分类器欲最大化的目标，即 S_b 与 S_w 的“广义瑞利商 (generalized Rayleigh quotient)”。

定义 Lagrange函数为：

$$L(c, \lambda) = c^{T} S_b c - \lambda (c^{T} S_w c - \text{const})$$

其中λ为Lagrange乘子，const表示分母不为0。

将上式对c求偏导，可得：

$$\frac{\partial L(c, \lambda)}{\partial c} = S_b c - \lambda S_w c$$

令偏导数为0，有：

$$S_b c^* - \lambda S_w \cdot c^* = 0$$

即：

$$S_b c^* = \lambda S_w \cdot c^*$$

将上式两边左乘 S_w^{-1}，可得：

$$S_w^{-1} \cdot (S_b c^*) = \lambda c^*$$

由于 $S_b = (m_1 - m_2) \cdot (m_1 - m_2)^T$，

$$S_b c^* = (m_1 - m_2) \cdot (m_1 - m_2)^T c^*$$

其中，$R = (m_1 - m_2)^T c^*$ 是一个标量，因此上式可以改写为：

$$S_b c^* = (m_1 - m_2) \cdot R$$

可得到：

$$\lambda c^* = S_w^{-1} \cdot (m_1 - m_2) \cdot R$$

即：

$$c^* = \frac{R}{\lambda} S_w^{-1} \cdot (m_1 - m_2)$$

由于 R / λ 为比例因子，可忽略。从而可得：

$$c^* = S_w^{-1} \cdot (m_1 - m_2)$$

其中：

$$S_w = S_1 + S_2$$

为两类内差矩阵之和；

$$m_1 - m_2$$

为两类的类间差向量。

4.3 Fisher 判别函数求取步骤：

(1) 先将原始数据写成矩阵形式。

组 *A* 的数据矩阵：$w^0 = \begin{bmatrix} x_{11}^0 & x_{12}^0 & \cdots & x_{1d}^0 \\ x_{21}^0 & x_{22}^0 & \cdots & x_{2d}^0 \\ & & & \\ x_{s1}^0 & x_{s2}^0 & \cdots & x_{sd}^0 \end{bmatrix}$

组 *B* 的数据矩阵：$w^1 = \begin{bmatrix} x_{11}^1 & x_{12}^1 & \cdots & x_{1d}^1 \\ x_{21}^1 & x_{22}^1 & \cdots & x_{2d}^1 \\ & & & \\ x_{t1}^1 & x_{t2}^1 & \cdots & x_{td}^1 \end{bmatrix}$

矩阵 w^0 和 w^1 的列平均数分别为 $(\overline{x}_1^0, \overline{x}_2^0, \ldots, \overline{x}_d^0)$ 和 $(\overline{x}_1^1, \overline{x}_2^1, \ldots, \overline{x}_d^1)$

(2) 算出各组数据的代表，即平均值

$$\bar{x}_j^0 = \frac{1}{s}\sum_{i=1}^{s} x_{ij}^0 \qquad j = 1, 2, \ldots, d$$

$$\bar{x}_j^1 = \frac{1}{s}\sum_{i=1}^{s} x_{ij}^1 \qquad j = 1, 2, \ldots, d$$

(3) 作新的矩阵*A*，*B*及两组的类内差矩阵 S_1，S_2：

$$A = \begin{bmatrix} x_{11}^0 - \bar{x}_1^0 & x_{12}^0 - \bar{x}_2^0 & \cdots & x_{1d}^0 - \bar{x}_d^0 \\ x_{21}^0 - \bar{x}_1^0 & x_{22}^0 - \bar{x}_2^0 & \cdots & x_{2d}^0 - \bar{x}_d^0 \\ & & & \\ x_{s1}^0 - \bar{x}_1^0 & x_{s2}^0 - \bar{x}_2^0 & \cdots & x_{sd}^0 - \bar{x}_d^0 \end{bmatrix} \qquad B = \begin{bmatrix} x_{11}^1 - \bar{x}_1^1 & x_{12}^1 - \bar{x}_2^1 & \cdots & x_{1d}^1 - \bar{x}_d^1 \\ x_{21}^1 - \bar{x}_1^1 & x_{22}^1 - \bar{x}_2^1 & \cdots & x_{2d}^1 - \bar{x}_d^1 \\ & & & \\ x_{t1}^1 - \bar{x}_1^1 & x_{t2}^1 - \bar{x}_2^1 & \cdots & x_{td}^1 - \bar{x}_d^1 \end{bmatrix}$$

$$S_1 = A'A, \qquad S_2 = B'B, \qquad S = S_1 + S_2$$

(4) 根据极值点推导，最优判别函数系数 $c_1, c_2, \ldots, c_d$ 为下述方程的解；

$$S\begin{pmatrix} c_1 \\ c_2 \\ \\ c_d \end{pmatrix} = \begin{pmatrix} m_{11}-m_{21} \\ m_{12}-m_{22} \\ \\ m_{1d}-m_{2d} \end{pmatrix} = \begin{pmatrix} \bar{x}_1^0-\bar{x}_1^1 \\ \bar{x}_2^0-\bar{x}_2^1 \\ \\ \bar{x}_d^0-\bar{x}_d^1 \end{pmatrix} \quad \text{即} \quad \begin{pmatrix} c_1 \\ c_2 \\ \\ c_d \end{pmatrix} = S^{-1}\begin{pmatrix} \bar{x}_1^0-\bar{x}_1^1 \\ \bar{x}_2^0-\bar{x}_2^1 \\ \\ \bar{x}_d^0-\bar{x}_d^1 \end{pmatrix}$$

(5) 解上述方程，得到判别函数：

$$y = c_1 x_1 + c_2 x_2 + \ldots\ldots + c_d x_d$$

(6) 算出组 *A*，组 *B* 的代表的判别值及临界判别值；

$$\bar{y}_A = c_1\bar{x}_1^0 + c_2\bar{x}_2^0 + \ldots + c_d\bar{x}_d^0 \qquad \bar{y}_B = c_1\bar{x}_1^1 + c_2\bar{x}_2^1 + \ldots + c_d\bar{x}_d^1$$

$$y_0 = \frac{s\bar{y}_A + t\bar{y}_B}{s+t}$$

(7) 对新数据作预测判别。

假设有一新数据为 $(x_{01}, x_{02}, \ldots, x_{0d})$，则其判别值为

$$y = c_1x_{01} + c_2x_{02} + \ldots + c_dx_{0d}$$

a)当 $\bar{y}_A > y_0$ 时，若 $y > y_0$，则判别该对象属于组 *A*，若 $y < y_0$，则判别该对象属于组 *B* 。

b)当 $\bar{y}_B > y_0$ 时，若 $y > y_0$，则判别该对象属于组 *B*，若 $y < y_0$，则判别该对象属于组 *A* 。

4.4 应用举例：

设某外贸公司生产一种产品，为正式上市之前，将样品寄往12个国家的进口代理商，并附意见调查表，要求对该产品进行评估。评估的内容有式样，包装，耐久性三个方面。评估的结果采用10分制计分，评估后并被要求说明是否愿意购买，调查结果列入表 2.3 中，表中的分数，高者表示代理商认为其特性良好，否则即较差。

表2.3 Fisher线性分类器训练样本

	编号	式样 x_1	包装 x_2	耐久性 x_3		编号	式样 x_1	包装 x_2	耐久性 x_3
购买者	1	9	8	7	非购买者	8	8	4	4
	2	7	6	6		9	3	6	6
	3	10	7	8		10	6	3	3
	4	8	4	5		11	6	4	5
	5	9	9	3		12	8	2	2
	6	8	6	7					
	7	7	5	6					

问：

今有第13个国家的进口代理商对该产品的评分分别是：式样9分，包装5分，耐久性4分，要预测该国是否愿意购买该产品。

(1) 计算两组的平均值：

购买者平均得分为 (8.29, 6.43, 6.00) ；

非购买者平均得分为 (6.20, 3.80, 4.00) 。

(2) 计算两组资料的离差矩阵：

$$A=\begin{pmatrix} 0.71 & 1.57 & 1 \\ -1.29 & -0.43 & 0 \\ 1.71 & 0.57 & 2 \\ -0.29 & -2.43 & -1 \\ 0.71 & 2.57 & -3 \\ -0.29 & -0.43 & 1 \\ -1.29 & -1.43 & 0 \end{pmatrix} \qquad B=\begin{pmatrix} 1.8 & 0.2 & 0 \\ -3.2 & 2.2 & 2 \\ -0.2 & -0.8 & -1 \\ -0.2 & 0.2 & 1 \\ 1.8 & -1.8 & -2 \end{pmatrix}$$

两组的离差矩阵分别为

$$S_1 = A'A = \begin{pmatrix} 7.43 & 7.14 & 2 \\ 7.14 & 17.71 & -3 \\ 2 & -3 & 16 \end{pmatrix} \qquad S_2 = B'B = \begin{pmatrix} 16.8 & -9.8 & -10 \\ -9.8 & 8.8 & 9 \\ -10 & 9 & 10 \end{pmatrix}$$

$$S = S_1 + S_2 = \begin{pmatrix} 24.23 & -2.66 & -8 \\ -2.66 & 26.51 & 6 \\ -8 & 6 & 26 \end{pmatrix}$$

(3) 解线性方程组：

$$S\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} \overline{x}_1^0 - \overline{x}_1^1 \\ \overline{x}_2^0 - \overline{x}_2^1 \\ \overline{x}_3^0 - \overline{x}_3^1 \end{pmatrix} = \begin{pmatrix} 8.29 - 6.20 \\ 6.43 - 3.80 \\ 6.00 - 4.00 \end{pmatrix}$$

即

$$\begin{cases} 24.03c_1 - 2.66c_2 - 8c_3 = 2.09 \\ -2.66c_1 + 26.51c_2 + 6c_3 = 2.63 \\ -8c_1 + 6c_2 + 26c_3 = 2 \end{cases}$$

得判别系数：

$$c_1 = 0.128 \qquad c_2 = 0.090 \qquad c_3 = 0.095$$

(4) 根据计算结果，得判别函数为：

$$y = 0.128x_1 + 0.090x_2 + 0.095x_3$$

(5) 求出判别临界值：

购买组的平均值对应的判别值：

$$\overline{y}_A = 0.128 \times 8.29 + 0.090 \times 6.43 + 0.095 \times 6.00 = 2.210$$

非购买组的平均值对应的判别值为：

$$\overline{y}_B = 0.128 \times 6.20 + 0.090 \times 3.80 + 0.095 \times 4.00 = 1.516$$

从而临界值为：$y_0 = \dfrac{2.210 \times 7 + 1.516 \times 5}{7 + 5} = 1.921$

(6) 对第13国的新数据作判别预测：

按判别规则，由于 $\bar{y}_A > y_0$ ，故凡判别值大于临界值 y_0=1.921 者，即判别其属于购买组。今第13个国家的判别值为

$$y = 0.128 \times 9 + 0.090 \times 5 + 0.095 \times 4 = 1.982$$

因1.982>1.921，故预测该国属于购买者范围。