第3节 支持向量机理论

——最优决策面

3.1 分类间隔最大

SVM算法从寻找线性可分情况下的最优分类面开始，要求该分类面不但能将两类无错误地分开，而且要使两类的分类间隔最大。前者是保证经验风险最小 (如使训练误差为0)，而使分类间隔最大实际上就是使推广性的界中的置信范围最小，从而使真实风险最小。

VC维与类间最大间隔的关系：

对于线性可分的样本 $(x_1, y_1), \ldots, (x_l, y_l) \in \{-1,+1\}$。设 R 为包含所有样本的球半径，F是以 Δ 间隔把 l 个样本分开的超平面集合，则函数集 F 的VC维 h 满足：

$$h \le min\left(\left[\frac{R^2}{\Delta^2}\right], n\right) + 1$$

式中， n 为样本的维数。

当 $\frac{R^2}{\Delta^2} < n$ 时，若超球半径 R 越小，且间隔 Δ 越大，则超平面集合F 的 VC 维 h 越小。

3.2 基于最大间隔的决策面

3.2.1 线性决策面定义：

对于 C 类分类问题，按分类决策规则可把 d 维特征空间分成 C 个决策域。划分决策域的边界面称为决策面。

判别函数和决策面是密切相关的，都由相应的决策规则确定。

在两类情况下，可以定义一个判别函数：

$$g(X) = g_1(X) - g_2(X)$$

可以线性分割的数据，其决策面可以表示为如下的线性组合函数：

$$g(X) = w^{\mathrm{T}} \cdot X + w_0$$

一般而言，X 为一维时，决策面为一个点；X 为二维时，决策面为一条直线；X 为三维时，决策面为一平面；X 大于三维时，决策面为一超平面。

对于两类问题的决策规则为：

- 如果 $g(X)>0$，则判定 X 属于 C_1；
- 如果 $g(X)<0$，则判定 X 属于 C_2；
- 如果 $g(X)=0$，则判定 X 在决策面上。如图3.1所示：

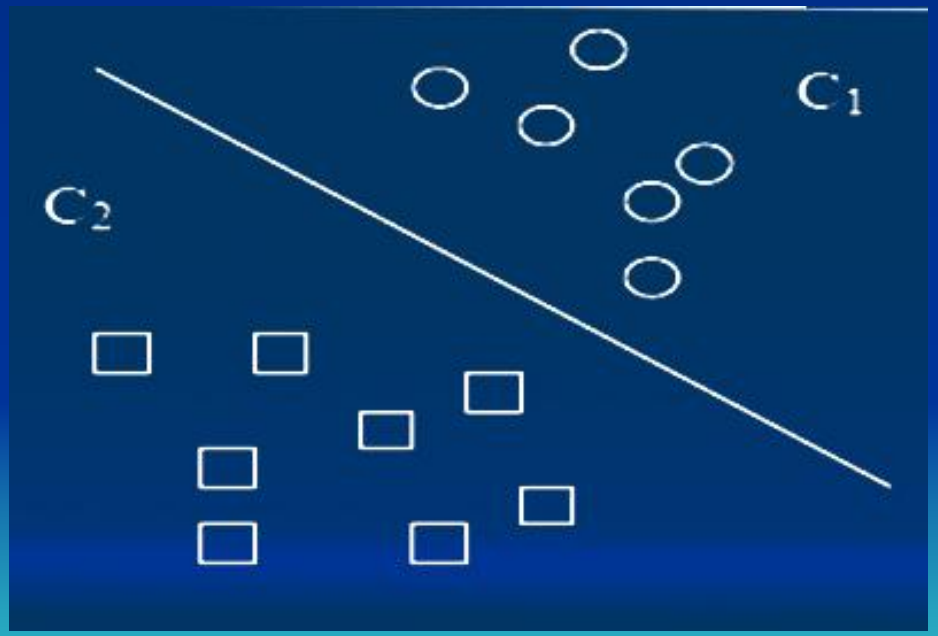

图3.1 两类线性数据的决策面

当两个数据点 X_1 和 X_2 都在决策面上时，有：

$$w^{\mathrm{T}} \cdot X_1 + w_0 = w^{\mathrm{T}} \cdot X_2 + w_0$$

移项得到：　$w^{\mathrm{T}} \cdot (X_1 - X_2) = 0$

这表明权系数矩阵 w 和决策面上的任意向量正交，则 w 为决策面的法向量。

注意到：$X_1 - X_2$ 表示决策面上的一个任意向量。

X

3.2.2 线性决策面的函数特性

线性决策面的函数中，X 是 d 维特征向量，又称样本向量，可以表示为：

$$X = [x_1, x_2, \ldots, x_d]^T = \begin{bmatrix} x_1 \\ x_2 \\ \\ \\ x_d \end{bmatrix}$$

w 称为权向量，可以表示为：

$$w = [w_1, w_2, \ldots, w_d]^T = \begin{bmatrix} w_1 \\ w_2 \\ \\ \\ w_d \end{bmatrix}$$

而 w_0 是一个常数，称为阀值。

方程 $g(X) = 0$ 则定义了一个分类面 H，它把归类于类 C_1 的点和归类于类 C_2 的点分隔开来。

把样本向量 X 表示为其在决策面 H 上的投影，得到：

$$X = X_P + r \cdot \frac{w}{\|w\|}$$

其中：X_P 是 X 在决策面 H 上的投影向量；

r 是 X 到决策面 H 的垂直距离，是一个标量。

上述分析从图 3.2 容易看出：

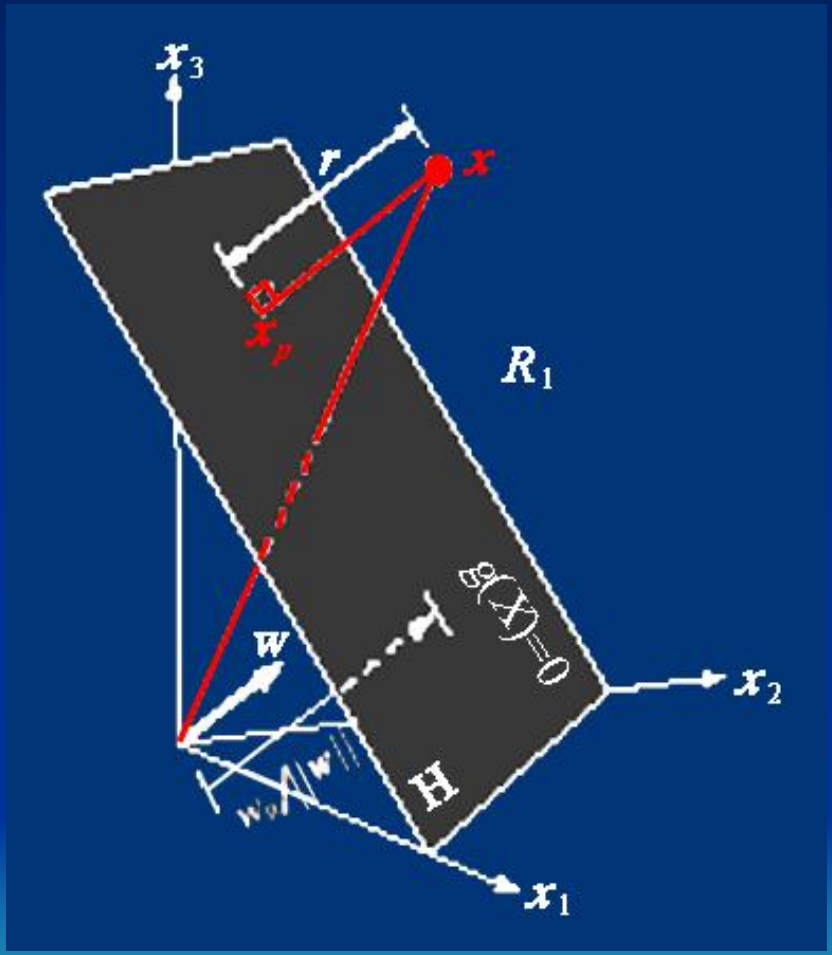

图3.3 X 分解为 H 上的投影 X_p 及垂直于 H 的向量

将 $X = X_P + r \cdot \frac{w}{\|w\|}$ 代入 $g(X) = w^T \cdot X + w_0$ 中，有：

$$g(X) = w^T X + w_0$$

$$= w^T (X_P + r \cdot \frac{w}{\|w\|}) + w_0$$

$$= w^T X_P + w_0 + w^T \cdot r \cdot \frac{w}{\|w\|}$$

$$= r \cdot \|w\|$$

$\frac{w}{\|w\|}$ 是决策面 H 法线 方向上的单位向量；

$\|w\| = \sqrt{w_1^2 + w_2^2 + \ldots + w_d^2}$ 是权向量 w 的欧几里德模。

从而有：$$r = \frac{g(X)}{\|w\|}$$

若 X 是原点，则有：

$$g(X) = w^{\mathrm{T}}X + w_0 = 0 + w_0$$

即：$$g(X) = w_0$$

将 $g(X)$的表达式代入 r 的表达式中，则有原点到法平面 H 的距离 r_0 的表达：

$$r_0 = \frac{w_0}{\|w\|}$$

综合上述式子，可知：

1) 若 $w_0 = 0$，则 $g(X)$具有齐次形式：

$$g(X) = w^{\mathrm{T}} \cdot X$$

这说明决策面 H 通过原点；

2) 若 $w_0 > 0$，则原点在 H 的正侧；(法向量的方向为正向，反向为负向。H的正侧、负侧也可以此为由。)

3) 若 $w_0 < 0$，则原点在 H 的负侧。

结论：利用线性决策面函数进行决策，就是用一个决策面把特征空间分割成两个决策区域，决策面的方向由权向量 w 确定，位置由阀值权 w_0 确定。判别函数 $g(X)$ 正比于点到决策面的代数距离 (带正负号).在正侧时 $g(X)>0$；在负侧时 $g(X)<0$。

3.2.3 最优分类面

支持向量机SVM是以线性可分情况下，提出最优分类面 (Optimal Hyperplane) 的。

1) 最优分类面概念

考虑二维两类线性可分情况，如图3.4所示：

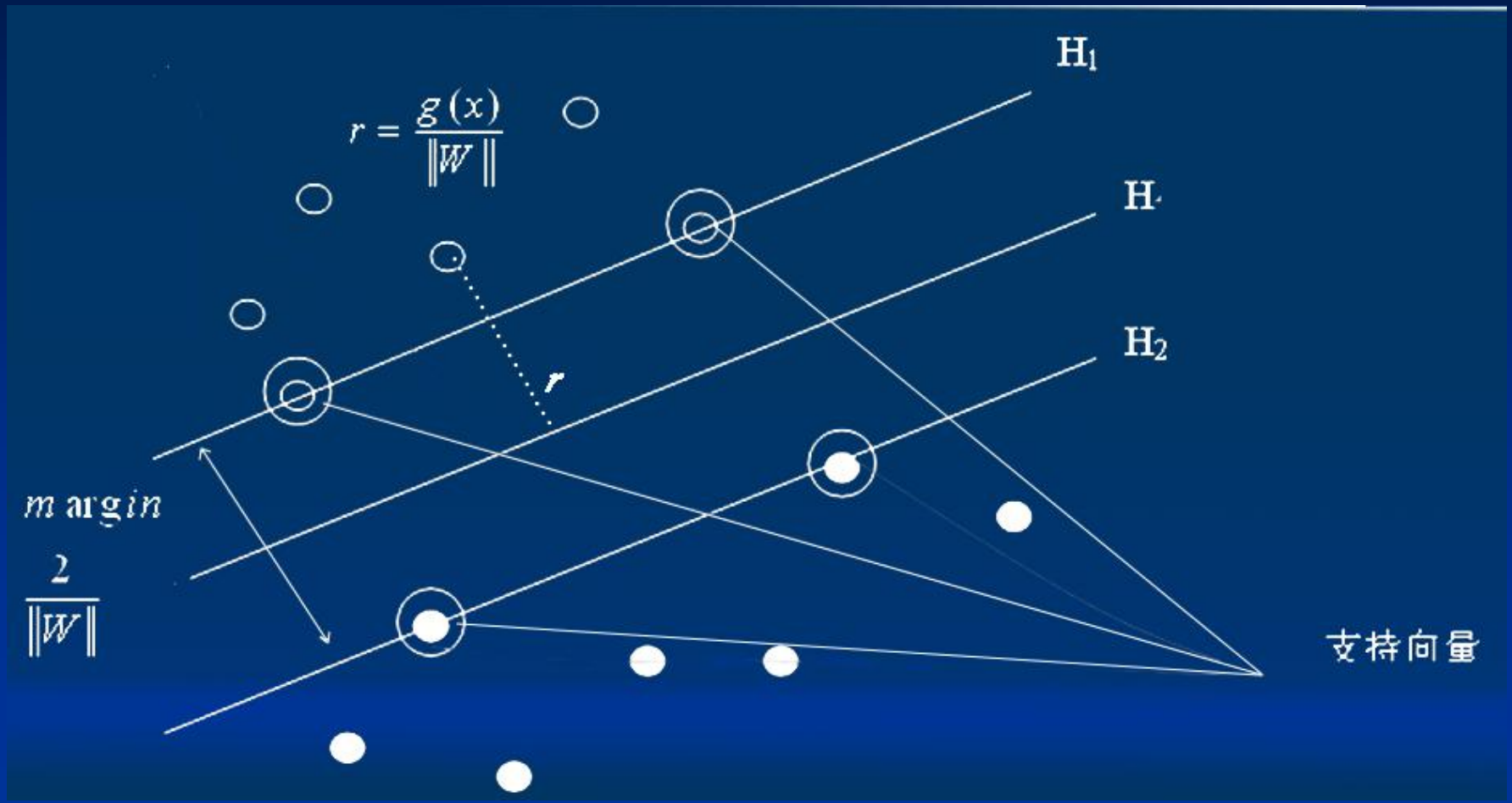

图3.4 二维数据点两类线性可分

图中，H 是分类线，H_1，H_2 分别为过各类样本中离分类线 H 最近的点且平行于分类线 H 的直线。H_1 和 H_2 之间的距离叫做两类的分类间隔 (**Margin**)。

最优分类线 —— 不但能将两类无错误地分开，而且使两类的分类间隔最大的分类线。

无错误分类是保证经验风险最小。分类间隔最大是使推广性的界中的置信范围最小，从而使真实风险最小。推广到高维空间，最优分类线就是最优分类面。

2) 最优分类面的形式化描述

设存在线性可分样本集：

$$(x_i, y_i) \quad i = 1, 2, \ldots, n$$

$y_i \in \{+1, -1\}$是类别标号。

d 维空间线性判别函数一般形式为：

$$g(X) = W \cdot X + b$$

则分类面方程为：

$$g(X) = W \cdot X + b = 0$$

对判别函数进行归一化，即使两类的所有样本都满足：

$$|g(X)| \geq 1$$

得到：$W \cdot X + b \geq 1$

或 $W \cdot X + b \leq -1$

则离分类面最近的样本满足：

$$|g(X)| = 1$$

由于样本到分类面的距离都为：

$$r = \frac{g(X)}{\|w\|}$$

则分类间隔 M 为：

$$M = 2r = \frac{2g(X)}{\|w\|} = \frac{2}{\|w\|}$$

要使分类间隔 M 取得最大值，则等价于使 $\|w\|$ 或 $\|w\|^2$ 最小。

要使分类面对所有样本正确分类，还要求：

$$y_i \cdot (W \cdot X_i + b) - 1 \geq 0$$

满足上述约束条件，并且使 $\|w\|^2$ 最小的分类面就是最优分类面。

3) 支持向量（Support Vectors）

两类样本中，离分类面最近且平行于最优分类面的超平面 H_1、H_2 上的样本，称为支持向量。

很明显，支持向量满足：

$$y_i \cdot (W \cdot X_i + b) - 1 = 0$$

从上式看出，支持向量支撑了最优分类面。

4) 最优分类面的求取

根据最优分类面的定义，则最优分类面的求取可以表示成如下约束优化问题：

即在约束 $y_i \cdot (W \cdot X_i + b) - 1 \geq 0$ 下，求函数

$$\Phi(W) = \frac{1}{2}\|W\|^2 = \frac{1}{2}(W \cdot W)$$

的最小值。

为了带约束项的最优化问题，可采用Lagrange乘子法。

Lagrange乘子法是一种在不等式约束条件下的优化算法。其基本思想是把不等式约束问题转化为无约束问题。

(*a*) Lagrange乘子法

以二维情况对Lagrange乘子法进行介绍。

SVM的优化问题可表示为：

$$\min_{W,b} \frac{1}{2}\|W^2\|$$

$$s.t. \quad y_i \cdot (W \cdot X_i + b) \geq 1, \qquad i = 1,2,...,n$$

将约束条件改写为：

$$g_i(W) = -y_i \cdot (W \cdot X_i + b) + 1 \leq 0$$

则可以构造SVM的拉格朗日函数：

$$L(W,b,\alpha) = \frac{1}{2}\|W\|^2 - \sum_{i=1}^{n} \alpha_i [y_i (W \cdot X_i + b) - 1], \qquad \alpha_i > 0$$

(*b*) Lagrange不等式约束的优化问题

对于更一般化的不等式约束优化问题如下：

$$\min_{\omega} \quad f(\omega)$$

$$s.t. \quad g_i(\omega) \le 0, \qquad i = 1,2,...,n$$

定义其最一般化的拉格朗日公式：

$$L(\omega,\alpha) = f(\omega) + \sum_{i=1}^{n} \alpha_i \cdot g_i(\omega) \ , \qquad \alpha_i > 0$$

如果按照最一般化的拉格朗日公式求解，当 α_i 取很大的正值时，$L(\omega,\alpha)$优化得到的最后结果会达到负无穷。

为避免出现上述问题，定义一个新的函数：

$$\theta_P(\omega)=\max_{\alpha,\alpha_i\geq 0} L(\omega,\alpha)$$

对违反约束的 $g_i(\omega)>0$，取 α_i 使$\theta_P(\omega)$ 结果为正无穷，从而阻止原优化函数 $L(\omega,\alpha)$ 求取极小值，作为对违反约束项的惩罚。

相反地，如果样本 X_i 满足约束 $g_i(\omega) \le 0$，则可以取 α_i 使 $\sum_{i=1}^{n} \alpha_i \cdot g_i(\omega) = 0$ 。

此时的 $f(\omega) = \theta_P(\omega)$，$\min_{\omega} f(\omega) = \min_{\omega} \theta_P(\omega)$ 。

则最原始的优化问题可以改写为如下形式：

$$\min_{\omega} f(\omega) = \min_{\omega} \max_{\alpha, \alpha_i \ge 0} L(\omega, \alpha)$$

上式称为广义拉格朗日函数的极小极大问题。定义原始问题的最优解：

$$p^* = \min_{\omega} \theta_P(\omega)$$

(c) Lagrange优化的对偶表示

定义：

$$\theta_D(\alpha)=\min_{\omega} L(\omega,\alpha)$$

考虑极大化 $\theta_D(\omega)$ 函数，得到：

$$\max_{\alpha,\alpha_i\geq 0}\theta_D(\alpha)=\max_{\alpha,\alpha_i\geq 0}\min_{\omega} L(\omega,\alpha)\ ,$$

$$s.t. \qquad \alpha_i \geq 0\ , \qquad i=1,2,...,n$$

称为广义拉格朗日函数的极大极小问题，是上述原始问题的对偶问题，其最优值为：$d^*=\min_{\alpha,\alpha_i\geq 0}\theta_D(\alpha)$

(d) SVM中原始问题和对偶问题互换

由于原始问题中，需要先对 α 参数进行优化，已求得使 $\sum_{i=1}^{n}\alpha_i \cdot g_i(\omega)=0$ 的 α 参数组合，但不好求解。

另一方面，原始问题的对偶问题，即将参数α看成固定值，先对ω，b参数进行优化，之后再求 $\theta_D(\omega)$ 的极大值要容易。

一般来说，这两种做法的结果并不一致，关系如下：

$$d^* = \max_{\alpha,\alpha_i\geq 0} \min_{\omega} L(\omega,\alpha) \leq \min_{\omega} \max_{\alpha,\alpha_i\geq 0} L(\omega,\alpha) = p^*$$

但是，在SVM的最优分界面问题中，这两者一致。因此，求解可以通过对偶问题实现。

首先，固定参数α的值，此时$L(\omega, \alpha)$的最小值只与ω(即W，b)参数有关。对W和b分别求偏导，得到：

$$\begin{cases} \dfrac{\partial}{\partial W} L(W, b, \alpha) = W - \sum_{i=1}^{n} \alpha_i y_i X_i = 0 \\ \dfrac{\partial}{\partial b} L(W, b, \alpha) = \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

然后，根据得到的α表达式再对α求偏导，求解相应的α最优值。

3.3 最优分类面的求解过程

针对目标函数：$\Phi(W)=\frac{1}{2}(W\cdot W)$ 及约束条件：

$$y_i[(W\cdot X_i)+b]-1=0,\qquad i=1,2,\ldots,n$$

定义Langrange函数如下：

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

在最优分类面求解中，就是对 W 和 b 求Lagrange函数的极小值。

对上式分别求对 W 和 b 的偏微分，并令其为 0，有：

$$\begin{cases} \dfrac{\partial L}{\partial W} = W - \sum_{i=1}^{n} \alpha_i y_i X_i = 0 \\ \dfrac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

3.3.1 求 W 的极值 W^*

$$W^* = \sum_{i=1}^{n} \alpha_i y_i X_i$$

只有当 $\alpha_i \neq 0$ 时，样本 X_i 才能对权向量 W 起作用，形成最优分类面。故而 $\alpha_i \neq 0$ 对应的样本 X_i 称为支持向量。

将对 b 求导的结果代入目标拉格朗日函数中，有：

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

$$=\frac{1}{2}(W\bullet W)-\sum_{i=1}^{n}\alpha_i\{y_i(W\cdot X_i)-1\}$$

再把 $W=\sum_{i=1}^{n}\alpha_i y_i X_i$ 代入上式，则有：

$$L(W,b,\alpha)=\frac{1}{2}(\sum_{i=1}^{n}\alpha_i y_i X_i)(\sum_{i=1}^{n}\alpha_i y_i X_i)-\sum_{i=1}^{n}\alpha_i y_i(\sum_{j=1}^{n}\alpha_j y_j X_j X_i)+\sum_{i=1}^{n}\alpha_i$$

$$=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\alpha_i y_i X_i\sum_{j=1}^{n}\alpha_j y_j X_j=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j X_i X_j$$

$$=A(\alpha)$$

3.3.2 原始问题转换为对偶问题的求解

即在约束条件

$$\sum_{i=1}^{n}\alpha_i y_i = 0(\alpha_i \geq 0, i = 1,2, \quad n)$$

下求：

$$A(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j X_i X_j$$

的极值问题。

上述问题可用下列方程组求解：

$$\begin{cases} \dfrac{\partial A(L)}{\partial \alpha_i} = 0 \\ \displaystyle\sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

根据**Kiilm-Tucker**条件，这个问题的解必须满足：

$$\alpha_i [y_i (W \cdot X_i + b) - 1] = 0, \qquad i = 1, 2, ..., n$$

3.3.3 支持向量和非支持向量

1) 当 $\alpha_i \neq 0$ 时，满足 $y_i(W \cdot X_i + b) - 1 = 0$ 的样本 X_i 在超平面 H_1、H_2 上，是支持向量。

2) 当 $\alpha_i = 0$ 时， 可能存在 $y_i(W \cdot X_i + b) - 1 \neq 0$，此时样本 X_i 在超平面 H_1、H_2 之外，是非支持向量。

3.3.4 最优分界面参数 W^* 和 b^* 求取

1) W^* 的求取：

$A(\alpha)$ 函数求导后得到 n 个方程，再加上参数 b 的约束项，一共有 $n+1$个方程。而要优化的 α 参数只有 n 个，故方程有解，表示为 α_i^*。

把优化得到的 α_i^* 的值代入W^* 的表达式中，求出 W^* 的值：

$$W^* = \sum_{i=1}^{n} \alpha_i^* y_i X_i$$

2) b^* 的求取：

当 $\alpha_i \neq 0$ 时，对应满足 $y_i(W \cdot X_i + b) - 1 = 0$ 的样本 X_i 就是支持向量。故而从 $\alpha_i \neq 0$ 的下标 i 去找出对应下标的样本 X_i，可以得到支持向量。

根据支持向量的定义，有

$$y_i(W^* \cdot X_i + b) - 1 = 0$$

从而有最优 b^*：

$$b^* = \frac{1}{y_i} - W^* \cdot X_i$$

其中，X_i 是支持向量。

3) 最优分类面的求取：

在求出最优权向量 W^*，最优偏置量 b^*之后，结合支持向量 X_i，可以得到最优分类面函数：

$$d(X) = W^* \cdot X + b^* = \sum_{i=1}^{n} \alpha_i^* y_i X_i \cdot X + b^*$$

当有一个样本 X 输入时，则可用最优分类面函数的符号来判别：

$d(X) > 0$，样本 X 属于 +1类 ($y_i = +1$)；

$d(X) < 0$，样本 X 属于 -1类 ($y_i = -1$)。