支持向量机的理论推导

1、线性可分情况

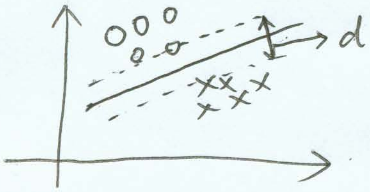

图1.

如图1.自纳，找一个平面，向上与向下平行移动该平面，使之擦过一些向量，将距离 d 定义为此平面的优化量度，使 d 尽可能大。d 叫做间距(margin)，擦过的向量叫支持向量(support vectors)

此想法的数学表达如下：

~~设平面~~设空间有N个向量 $x_1, x_2 \ldots x_N$，它们要么属于 C_1 类，要么属于 C_2 类，定义

$$y_i = \begin{cases} 1, & \text{如果 } x_i \in C_1 \\ -1, & \text{如果 } x_i \in C_2 \end{cases}$$

该优化问题可写为，寻找 w 和 b，使

最小化(Minimize)：$\frac{1}{2}\|w\|^2$

限制条件(Subject to)：$y_i[w^T x_i + b] \geq 1 \quad (i=1,2,\ldots N)$

（请参阅课堂笔记）

注意：① 此问题是凸优化中的二次规划问题。

② 此问题只有在线性可分情况下，才有 (w, b) 满足所有限制条件

③ $y_i[w^T x_i + b] = 1 \Longleftrightarrow x_i$ 为支持向量。

2. 线性不可分状况

在线性不可分情况下，优化问题写为：

最小化：$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\delta_i$ 或 $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\delta_i^2$

限制条件：① $\delta_i \geq 0 \quad (i=1,2,\ldots,N)$

② $y_i[w^T x_i + b] \geq 1 - \delta_i \quad (i=1,2,\ldots,N)$

注意：① C为常数，w，b，$\delta_i (i=1\sim N)$为待求变量。

② 此问题对任意点集，无论是否线性可分，都有解。

③此问题也是凸优化中的二次规划问题，理论上保证了有唯一的局部最小值(即局部最小值也是全局最小值)

④ $y_i [w^T x_i + b] = 1 \iff x_i$为支持向量。

当求出 w与b后，对一个向量x，需要判断其属于C_1还是C_2，判断标准为

$$\begin{cases} x \in C_1, \text{如果 } w^T x + b \geq 0 \\ x \in C_2, \text{如果 } w^T x + b < 0 \end{cases}$$

3. 非线性状况

支持向量机处理非线性是通过将向量x映射至高维，再用线性方式去分开。

例子，考虑如下异或问题

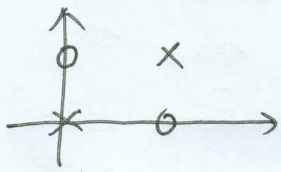

图2.

即 $x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \in C_1$　$x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in C_1$

$x_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in C_2$　$x_4 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \in C_2$

该例子中属于线性不可分，但如果我们定义由二维至五维的映射 $\varphi(x)$

$$\varphi(x): \quad x = \begin{bmatrix} a \\ b \end{bmatrix} \longrightarrow \varphi(x) = \begin{bmatrix} a^2 \\ b^2 \\ a \\ b \\ ab \end{bmatrix}$$

则 $\varphi(x_1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$　$\varphi(x_2) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

$\varphi(x_3) = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$　$\varphi(x_4) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$

则 $\varphi(x_1), \varphi(x_2), \varphi(x_3), \varphi(x_4)$ 线性可分。

设 $w = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ 6 \end{bmatrix}$ $\quad b = 1$

则
$$\begin{cases} w^T\varphi(x_1)+b = 1 \geq 0 \\ w^T\varphi(x_2)+b = 3 \geq 0 \\ w^T\varphi(x_3)+b = -1 < 0 \\ w^T\varphi(x_4)+b = -1 < 0 \end{cases}$$

所以线性可分。

对于此问题，只须修改SVM中优化问题 x 为 $\varphi(x)$ 即可。

最小化：$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\delta_i$ 或 $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\delta_i^2$

限制条件：① $\delta_i \geq 0 \quad (i=1\sim N)$

② $y_i[w^T\varphi(x_i)+b] \geq 1-\delta_i \quad (i=1\sim N)$

支持向量机创始人Vapnik在此问题上继续前进，他指出，我们可以不用知道 $\varphi(X)$ 的具体形式，取而代之，如果对空间任意向量，我们知道

$$K(x_1, x_2) = \varphi(x_1)^T\varphi(x_2),$$

则仍然能通过SVM，计算 $W^T\varphi(X)+b$ 的值，进而得出 X 的所属类别。定义 $K(x_1, x_2)$ 为核函数(Kernel function)

在讲如何通过核函数计算 $W^T\varphi(X)+b$ 之前，先研究核函数 $K(x_1, x_2)$ 满足什么性质，才能存在 $\varphi(x)$，使 $K(x_1, x_2) = \varphi(x_1)^T\varphi(x_2)$。Mercer's Theorem 给出了此问题的充要条件。

Mercer's Theorem：核函数 $K(X_1, X_2)$ 可拆为 $\varphi(X_1)^T\varphi(X_2)$ 的充要条件为：对于任意函数

$\psi(x)$ 满足 $\int_a^b \psi(x)^2 dx < +\infty$，

$$\int_a^b\int_a^b K(X_1, X_2)\,\varphi(X_1)\,\varphi(X_2)\,dX_1\,dX_2 \geq 0$$

其中，$[a, b]$ 为 X_1，X_2 的定义域。

举一个例子，例如可以证明高斯核 $K(X_1, X_2) = e^{-\frac{\|X_1 - X_2\|^2}{2\sigma^2}}$ 满足 Mercer's Theorem，那么可以将 $K(X_1, X_2)$ 表示为 $\varphi^T(X_1)\varphi(X_2)$ 的形式，但 $\varphi(X)$ 却不能写成显式表达式。虽然无法知道 $\varphi(X)$，但却可通过一些变换，知道 $W^T\varphi(X)+b$ 的值，进而获得 X 所属类别。

深入研究之前需要补充优化中的原问题与对偶问题的基础知识（详细请查阅 Stephen Boyd 等著 convex optimization）

一个优化问题的原问题与对偶问题定义如下：

原问题（Prime Problem）：

最小化（Minimize）：$f(w)$

限制条件（Subject to）：$g_i(w) \leq 0 \quad i = 1 \sim K$

$h_i(w) = 0 \quad i = 1 \sim M$

对偶问题（Dual Problem）：

定义
$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{K} \alpha_i g_i(w) + \sum_{i=1}^{M} \beta_i h_i(w)$$
$$= f(w) + \alpha^T g(w) + \beta^T h(w)$$

其中 $\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{bmatrix}$ $g(w) = \begin{bmatrix} g_1(w) \\ g_2(w) \\ \vdots \\ g_K(w) \end{bmatrix}$，$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix}$ $h(w) = \begin{bmatrix} h_1(w) \\ h_2(w) \\ \vdots \\ h_M(w) \end{bmatrix}$

对偶问题为：

最大化(Maximize)：

$$\theta(\alpha,\beta)=\inf_{\text{所有定义域内的}w} L(w,\alpha,\beta)$$

限制条件(Subject to) $\alpha_i \geq 0 \quad i=1\sim K$

定理1：如果 w^* 是原问题的解，α^* β^* 是对偶问题的解，则 $f(w^*) \geq \theta(\alpha^*,\beta^*)$

证明：

$$\theta(\alpha^*,\beta^*) = \inf_{w\in D} L(w,\alpha^*,\beta^*)$$

$$\leq L(w^*,\alpha^*,\beta^*)$$

$$= f(w^*) + \underbrace{\alpha^{*T}}_{\geq 0}\underbrace{g(w^*)}_{\leq 0} + \beta^{*T}\underbrace{h(w^*)}_{=0} \leq f(w^*)$$

得证。(注意：如果 $f(w^*)=\theta(\alpha^*,\beta^*)$，则必能推出对所有 $i=1\sim K$，$\alpha_i g_i(w^*)=0$，此条件叫KKT条件)

定义：把 $f(w^*)-\theta(\alpha^*,\beta^*)$ 定义为对偶差距(Duality Gap)

定理2(强对偶定理)：如果 $g(w)=Aw+b$，$h(w)=Cw+d$，$f(w)$ 为凸函数(凸函数定义为，对 $\forall w_1,w_2$，有 $f(\lambda w_1+(1-\lambda)w_2) \leq \lambda f(w_1)+(1-\lambda)f(w_2)$ $\lambda\in[0,1]$)

则 $f(w^*)=\theta(\alpha^*,\beta^*)$，即对偶差距为0

(详细证明见Convex Optimization一书)

支持向量机原问题：

最小化：$\frac{1}{2}\|w\|^2 - C\sum_{i=1}^{N}\xi_i$ 或 $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i^2$

情况1　　　　　　　　情况2

限制条件：① $\xi_i \le 0$ $(i=1\sim N)$

② $1+\xi_i - y_i w^T\varphi(x_i) - y_i b \le 0$ $(i=1\sim N)$

情况1的对偶问题：

最大化 $\theta(\alpha,\beta) = \inf\limits_{w,\xi_i,b} \frac{1}{2}\|w\|^2 - C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\beta_i\xi_i + \sum_{i=1}^{N}\alpha_i[1+\xi_i - y_i w^T\varphi(x_i) - y_i b]$ ①

$\frac{\partial\theta}{\partial w} = w - \sum_{i=1}^{N}\alpha_i\varphi(x_i)y_i = 0 \Rightarrow w = \sum_{i=1}^{N}\alpha_i y_i\varphi(x_i)$ ②

$\frac{\partial\theta}{\partial \xi_i} = -C + \beta_i + \alpha_i = 0 \Rightarrow \alpha_i + \beta_i = C$ ③

$\frac{\partial\theta}{\partial b} = -\sum_{i=1}^{N}\alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{N}\alpha_i y_i = 0$ ④

②③④代入①得：

最大化 $\theta(\alpha,\beta) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i\alpha_j \varphi(x_i)^T\varphi(x_j)$

限制条件：$\cancel{\alpha_i \ge 0, \beta_i \ge 0}$

① $0 \le \alpha_i \le C$（由于 $\beta_i \ge 0$ 且 $\beta_i = C-\alpha_i$，所以 $\alpha_i \le C$）$(i=1\sim N)$

② $\sum_{i=1}^{N}\alpha_i y_i = 0$ $(i=1\sim N)$

这也是一个二次规划问题，解此问题时，由于 $\varphi(x_i)^T\varphi(x_j) = K(x_i,x_j)$ 所以我们只须知道核函数，不需知道 $\varphi(x)$ 的显示表达。

解出 α 后，根据③得到 $w = \sum_{i=1}^{N}\alpha_i y_i\varphi(x_i)$，注意，由于 $\varphi(x)$ 不见得具有显式表达，所以 w 也不见得知道。但下面要说的是，不需要知道 w 的显式表达，我们也能计算 $w^T x + b$ 的值。

首先我们要求出 b。根据KKT条件，对所有 $i=1\sim N$，有：

$\alpha_i[1+\xi_i - y_i w^T\varphi(x_i) - y_i b] = 0$

$\beta_i\xi_i = 0 \Rightarrow (C-\alpha_i)\xi_i = 0$

由于 $w=\sum_{j=1}^{N}\alpha_j y_j \varphi(x_j)$

则 $w^T\varphi(x_i)=\sum_{j=1}^{N}\alpha_j y_j \varphi^T(x_j)\varphi(x_i)$

$$=\sum_{j=1}^{N}\alpha_j y_j K(x_j,x_i)$$

另一方面，如果对~~某~~个 i，$\alpha_i\neq 0$ 且 $\alpha_i\neq C$，则根据 KKT 条件，必有 $\delta_i=0$ 且

$$1+\underbrace{\delta_i}_{=0}-\underbrace{y_i w^T\varphi(x_i)}_{=\sum_{j=1}^{N}\alpha_j y_i y_j K(x_j,x_i)}-y_i b=0$$

所以，只须找一个 $0<\alpha_i<C$，则

$$b=\frac{1-\sum_{j=1}^{N}\alpha_j y_i y_j K(x_j,x_i)}{y_i}$$

所以，b 能够算出。

下面考虑，对于一个测试样本 X，我们需判断其所属类别，我们计算：

$w^T\varphi(x)+b$，将 $w=\sum_{i=1}^{N}\alpha_i y_i\varphi(x_i)$ 代入

$$w^T\varphi(x)+b=\sum_{i=1}^{N}\alpha_i y_i\varphi(x_i)^T\varphi(x)+b$$

$$=\sum_{i=1}^{N}\alpha_i y_i K(x_i,x)+b$$

所以，判决标准为

$$\begin{cases} X\in C_1, & \text{如果 } \sum_{i=1}^{N}\alpha_i y_i K(x_i,x)+b\geq 0 \\ X\in C_2, & \text{如果 } \sum_{i=1}^{N}\alpha_i y_i K(x_i,X)+b<0 \end{cases}$$

最终，我们只通过核函数，也能完成对 X 的类别判决。

情况 2：不详细讲，有兴趣的同学可自己推导一下。所有推导可见《支持向量机导论》一书。

一些常用核函数介绍

① 多项式核 $K(X_1, X_2) = (X_1^T X_2 + 1)^d$

② 高斯核 $K(X_1, X_2) = e^{-\frac{\|X_1 - X_2\|^2}{2\sigma^2}}$

③ tanh 核 $K(X_1, X_2) = \tanh(\beta X_1^T X_2 + b)$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

④