Lagrange dual function (对偶问题)

$$g(\lambda, v) = \inf_{x \in D} L(x, \lambda, v) = \inf_{x \in D}\left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)\right)$$

SVM - Lagrange Multiplier:

Goal: find hyperplane with maximal margin $\left(\frac{1}{\|w\|}\right)$

$$\begin{cases} w^T x_n + b \geq 1 & \text{for } t_n = +1 \\ w^T x_n + b \leq -1 & \text{for } t_n = -1 \end{cases}$$

choose w, b such that $d = \frac{1}{\|w\|}$

$\Rightarrow$ minimizing $\|w\|^2$

$\underset{w,b}{\arg\min}\ \frac{1}{2}\|w\|^2$ under constraints $t_n(w^T x_n + b) \geq 1 \ \forall n$

$$\mathcal{L}_{(w,b,a)} = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} \underbrace{a_n}_{\text{Lagrange multiplier}} \boxed{\{t_n(w^T x_n + b) - 1\}}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} a_n t_n = 0$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow \boxed{w = \sum_{n=1}^{N} a_n t_n x_n}$$

if $t_n(w^T x_n + b) > 1 \ \forall n$ (consider KKT (Karush-Kuhn-Tucker condition))

$$\mathcal{L} = \|w\|^2 + \lambda \cdot f(x) \Rightarrow \lambda f(x) = 0$$

$\lambda \geq 0$ if $\lambda > 0$ $f(x) = 0$

$f(x) \geq 0$

$a_n \geq 0$

$t_n y(x_n) - 1 \geq 0$

$\Downarrow$

$$\boxed{a_n\{t_n y(x_n) - 1\} = 0}$$

$w^T x_n + b$

lagrange multiplier

$\uparrow$

if $a_n \neq 0$

$\Downarrow$

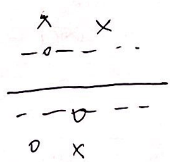
$t_n y(x_n) - 1 = 0$

vs.

$a_n = 0$

$\Downarrow$

$t_n y(x_n) - 1 \neq 0 \leftarrow$ not lay on the boundary.



$$t_n y(x_n) = t_n\Big(\sum_{m \in S \to \text{support vectors}} a_m t_m x_m^T \cdot x_n + b\Big) = 1, \quad t_n^2 = 1$$

$$b = t_n - \sum_{m \in S} a_m t_m x_m^T \cdot x_n$$

$$L_P = \frac{1}{2}\|W\|^2 - \sum_{n=1}^{N} a_n \{ t_n (w x_n^T + b) - 1 \} \quad \text{primal}$$

→ re-written as dual from

$$W = \sum_{n=1}^{N} a_n t_n x_n^T$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

$$L_{P_{rimal}} = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n t_n w^T x_n - b \sum_{n=1}^{N} \overset{0}{\cancel{a_n t_n}} + \sum_{n=1}^{N} a_n$$

$$= \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n t_n \left(\sum_{m=1}^{N} a_m t_m x_m^T \right) x_n + \sum_{n=1}^{N} a_n$$

$$\Rightarrow L_{dual} = -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m (x_m^T x_n) + \sum_{n=1}^{N} a_n \quad (\text{no } w, b \text{ anymore})$$

maximize L_{dual} under constraint $a_n \geq 0 \; \forall n$, $\sum_{n=1}^{N} a_n t_n = 0$

$$W = \sum_{n=1}^{N_{support\,vector}} a_n t_n x_n$$

↳ $N_{support\ vector}$ $O(\frac{N^3}{2})$

\# of data points for support vectors
not rely on # of dimensions.

In practice: $O(N) \sim O(N^2)$

$\boxed{\text{summary: Dual problem is good for high dimension problem}}$

↓

can apply kernel Trick.

Langrangelan multiplier

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} \underset{\neq 0}{a_n} \{ t_n (w^T \overset{\text{then S.V.}}{\boxed{x_n}} + b) - 1 \} = 0$$

only if ... $= 0$

suit for (low dimension data) $O(D^3)$ dimension

$$L_d = -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \underline{(x_m^T x_n)} + \sum_{n=1}^{N} a_n$$

$O(N^3)$: $O(N) \sim O(N^2)$ in practice b/c it only use S.V.

If we use kerneal trick, we need Ld (dual form)

$L_p(w, b, a)$

$L_d(a)$

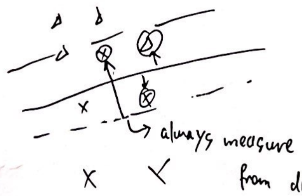
---

slack variable

$w^T x_n + b \geq +1 - \xi_n$ for $t_n = +1$

$w^T x_n + b \leq -1 - \xi_n$ for $t_n = -1$

$$\boxed{\xi_n \geq 0}$$



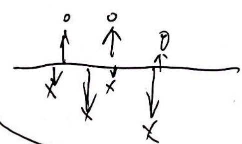
always measure from desicion boundary

goal : non-separable data, minimize $\frac{1}{2}\|w\|^2 + C \sum_{n=1}^{n} \xi_n$

→ Trade-off parameter

↳ trade-off btw slack & margin

$C \to \infty \Rightarrow$ linear discriminant (will measure every noise)



(New svm ~~principal~~ primal)

optimize $L_p = \frac{1}{2}\|w\|^2 + C \sum_{n=1}^{N} \boxed{\xi_n} - \sum_{n=1}^{N} \underbrace{a_n (t_n y(x_n) - 1 + \boxed{\xi_n})}_{①} - \sum_{n=1}^{N} \underbrace{\mu_n \xi_n}_{②}$

apply KKT:

① $a \geq 0$

$t_n y(x_n) - 1 + \xi_n \geq 0$

$a_n (t_n y(x_n) - 1 + \xi_n) = 0$

② $\mu_n \geq 0$

$\xi_n \geq 0$

$\mu_n \xi_n = 0$

$$\frac{\partial L}{\partial \xi_n} \overset{!}{=} 0$$

$C - a_n - \mu_n = 0$

$C - \mu_n = a_n$

$$\boxed{C \geq a_n \geq 0}$$

we can determine whether the data work with SVM using lagrange multiplier.

i) $a_n = 0$, it is not S.V.

it is clearly classified

ii) $0 < a_n < C \rightarrow \mu_n \neq 0 \rightarrow \xi_n = 0$, it is SV.

it doesn't have slack (it is on the desicion boundary)

iii) $a_n = C$

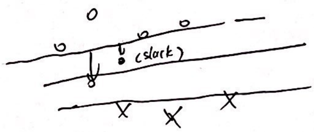
$C - a_n - \mu_n = 0$

$\mu_n = 0$

$\mu_n \xi_n = 0$, ~~we~~ It has slack variable (slack loss)

$\downarrow \quad \downarrow$

$0 \quad \neq 0$



so we can know where data located according to $\boxed{a_n}$.