Primal Dual Relationship.

1. If primal has a solⁿ, so does dual & optimal values are equal.

$$q(y^*) = f(x^*)$$

2. x^* is optimal for primal & y^* is optimal for dual if & only if

   a) x^* is feasible for primal & y^* is feasible for dual.

   b) $f(x^*) = L(x^*, y^*) = \min_x L(x, y^*)$

In the context of SVM,

Primal: minimize $\frac{1}{2} w^T w$

s.t $1 - y_i(w^T x_i + b) \leq 0, \quad i = 1, \ldots n.$

$$L(w, b, \mu) = \frac{1}{2} w^T w + \sum_{i=1}^{n} \mu_i \left[1 - y_i (w^T x_i + b)\right]$$

K.T Cond:

$$\nabla_w L = 0 \Rightarrow w^* = \sum_{i=1}^{n} \mu_i^* y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \mu_i^* y_i = 0$$

$$1 - y_i (x_i^T w^* + b^*) \leq 0, \ \forall i.$$

$$\mu_i^* \geq 0, \quad \mu_i^* \left[1 - y_i (x_i^T w^* + b^*)\right] = 0 \ \forall i$$

Now, if $S = \{i : \mu_i^* > 0\}$

we have, $i \in S \Rightarrow y_i (x_i^T w^* + b^*) = 1$

$\Rightarrow$ x_i are the closest to the separating hyperplane.

Thus, $\{x_i \mid i \in S\}$ are called the "support vectors"

$$\therefore \; W^* = \sum_{i=1}^{n} \mu_i^* y_i x_i = \sum_{i \in S} \mu_i^* y_i x_i$$

separating hyperplane is a Linear Comb of support vectors

$$W^* = \sum_i \mu_i^* y_i x_i = \sum_{i \in S} \mu_i^* y_i x_i$$

$$b^* = y_j - x_j^T W^*, \quad j \text{ s.t } \mu_j^* > 0.$$

We want $\underline{\mu^*}$.

Dual :

$$q(\mu) = \inf_{w,b} \left\{ \frac{1}{2} w^T w + \sum_{i=1}^{n} \mu_i \left[1 - y_i (w^T x_i + b) \right] \right\}$$

we have if $\sum \mu_i y_i \neq 0$, then $q(\mu) = -\infty$

Thus, have another Constraint, $\sum \mu_i y_i = 0$

Also, min over w is attained at

$$w = \sum \mu_i y_i x_i$$

$\therefore$ $w = \sum \mu_i y_i x_i$ & $\sum \mu_i y_i = 0$ in $q(\mu)$

$$q(\mu) = \frac{1}{2} w^T w + \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \mu_i y_i (w^T x_i + b)$$

$$= \sum_i \mu_i - \frac{1}{2} \sum_i \sum_j \mu_i y_i \mu_j y_j x_i^T x_j$$

$\therefore$ The dual is

$$\max_{\mu} \quad q(\mu) = \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j=1}^{n} \mu_i \mu_j y_i y_j x_i^T x_j$$

$$\text{s.t} \quad \mu_i \geq 0, \quad i = 1 \dots, n$$

$$\sum_{i=1}^{n} y_i \mu_i = 0$$

Observe that in Dual training data only appears as inner products.

Also, optimization is over R^n, irrespective of data dimension (x_i)

$$\varphi(\mu) = \frac{1}{2} w^T w + \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \mu_i y_i (w^T x_i + b)$$

Substitute $w = \sum \mu_i y_i x_i$, $\sum \mu_i y_i = 0$,

$$\varphi(\mu) = \frac{1}{2} \Big(\sum_i \mu_i y_i x_i \Big)^T \sum_j \mu_j y_j x_j +$$

$$\sum_i \mu_i - \sum_i \mu_i y_i x_i^T \Big(\sum_j \mu_j y_j x_j \Big).$$

$$= \sum_i \mu_i - \frac{1}{2} \sum_i \sum_j \mu_i y_i \mu_j y_j x_i^T x_j$$

Dual: $\max_{\mu} \varphi(\mu) = \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j=1}^{n} \mu_i \mu_j y_i y_j x_i^T x_j$

s.t $\mu_i \geq 0$, $i = 1, \ldots, n$, $\sum_{i=1}^{n} y_i \mu_i = 0$.

Final soln for the SVM problem:

$$W^* = \sum \mu_j^* y_j x_j, \quad b^* = y_j - x_j^T W^*, \quad j : \mu_j > 0$$

SVM for non-separable case.

i) optimization problem has no feasible point if data is not linearly separable

$$\therefore \nexists \; W, \text{ s.t } y_i (W^T x_i + b) \geq 1 \quad \forall i$$

However, one can choose a set of slack variables $\xi_i \geq 0$ (loose) s.t

$$y_i (W^T x_i + b) \geq 1 - \xi_i \quad \forall i$$

But the problem now is, every W is feasible & thus $\frac{1}{2} W^T W$ will be minimized by $W = 0$.

To avoid such a degenerative soln, let us modify the cost as follows:

$$\min \frac{1}{2} W^T W + C \sum_{i=1}^{n} \xi_i$$

C - user defined

$$\text{s.t } y_i (W^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \quad \forall i$$

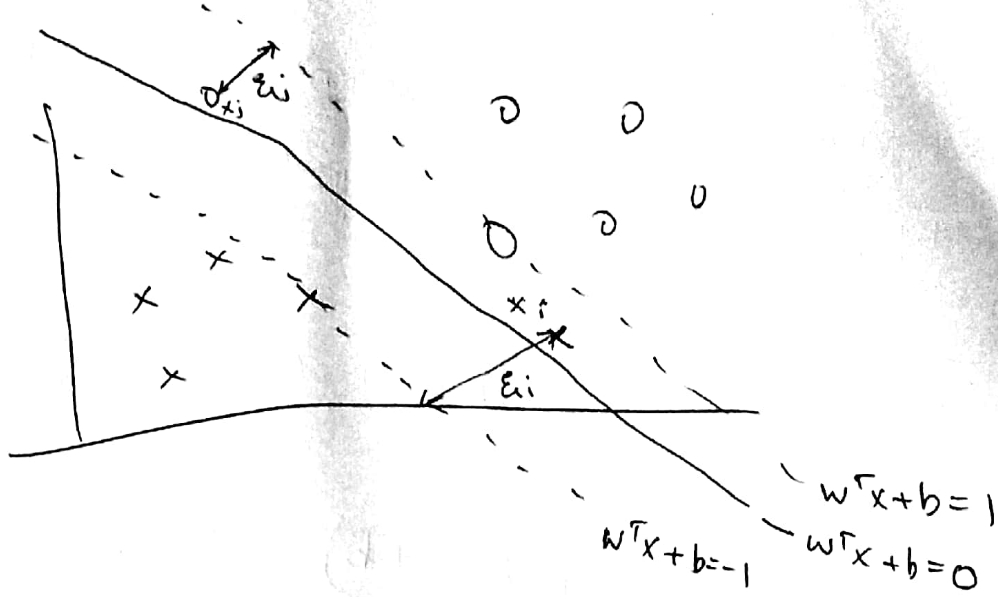
We not only want feasibility but ξ_is should be as small as possible.

Now, the optimization is over W, b & ξ_i

Geometrically, ξ_i measures the extent of violation of optimal separation

i.e, if $0 < \xi_i < 1$, then it is a margin error.

$\xi_i > 1$, x_i is misclassified.



New opti problem:

$$\min_{w,b,\xi_i} \frac{1}{2}w^Tw + C\sum_{i=1}^{n}\xi_i$$

s.t $\quad 1-\xi_i - y_i\left(w^Tx_i+b\right) \le 0, \quad i=1,\dots,n$

$\quad -\xi_i \le 0, \quad i=1,\dots n.$

Lagrangian: $L(w,b,\xi,\mu,\alpha) = \frac{1}{2}w^Tw + C\sum_{i=1}^{n}\xi_i$

$+ \sum_{i=1}^{n}\mu_i\left(1-\xi_i-y_i\left(w^Tx_i+b\right)\right) - \sum_{i=1}^{n}\alpha_i\xi_i$

$\mu_i \rightarrow$ lagran mul for separability Const.

$\alpha_i \rightarrow$ —— " —— for $-\xi_i \le 0$.

K.K.T

1) $\nabla_w L = 0 \Rightarrow W^* = \sum \mu_i^* y_i x_i$

2) $\frac{\partial L}{\partial b} = 0 \Rightarrow \sum \mu_i^* y_i = 0$

3) $\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \mu_i^* + \alpha_i^* = c, \ \forall i$

4) $1 - \xi_i - y_i (w^T x_i + b) \leq 0; \ \xi_i \geq 0; \forall i$

5) $\mu_i \geq 0; \ \alpha_i \geq 0, \ \forall i$

6) $\mu_i \left(1 - \xi_i - y_i (w^T x_i + b)\right) = 0; \ \alpha_i \xi_i = 0, \forall i$

w^* is same as before.

Also, $0 \leq \mu_i + \alpha_i = c$.

Now if $0 < \mu_i < c$, then $\alpha_i > 0 \Rightarrow \xi_i = 0$.

$\Rightarrow$ from (6), $1 - y_i (w^T x_i + b) = 0$

$\Rightarrow b^* = y_i - x_i^T w^*$, i s.t $0 < \mu_i < c$.

To find μ, derive ~~Lagra~~ Dual.

$$\varphi(\mu, \lambda) = \inf_{w, b, \xi} L(w, b, \xi, \mu, \lambda)$$

$$L = \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \mu_i \left(1 - \xi_i - y_i (w^T x_i + b)\right) - \sum_{i=1}^{n} \lambda_i \xi_i$$

L has a term : $\sum_i (C - \mu_i - \lambda_i) \xi_i$

$\Rightarrow$ we need to impose, $C = \mu_i + \lambda_i \;\; \forall i$

If we do, then all terms with λ_i & ξ_i would vanish & the φ function will be same as before.

only ensure that $\lambda_i \geq 0$ & $\mu_i + \lambda_i = C$. in the Dual. which could be easily achieved by $0 < \mu_i < C$.

$\Rightarrow$ Dual is

$$\max_{\mu} \varphi(\mu) = \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j=1}^{n} \mu_i \mu_j y_i y_j x_i^T x_j$$

$$\text{s.t } 0 \le \mu_i \le C, \; i = 1 \ldots n, \; \sum_{i=1}^{n} y_i \mu_i = 0$$

which is exactly as the previous one except for an upper bound on μ

This is another advantage of duality.

$\therefore$ SVM in the linear in-seperable case:

$$W^* = \sum \mu_i^* y_i x_i, \quad b^* = y_j - x_j^T W^*,$$

$$j \text{ s.t } 0 < \mu_j < C.$$

$\therefore$ Incorporating slack in the formulation ensures that one can find the 'best' hyperplane.