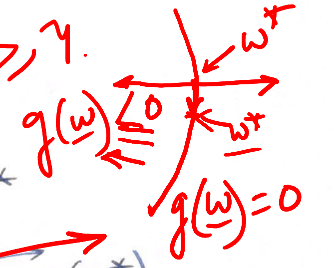
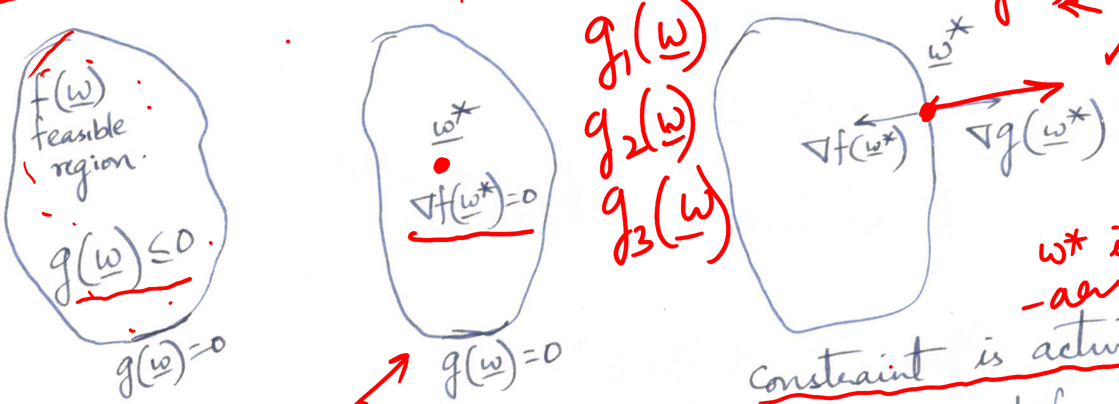
Dual Formulation of SVM

$\underbrace{\min f(\underline{w})}$ subject to $\underbrace{g(\underline{w}) \leq 0}$ $h(w) \geq 4$.

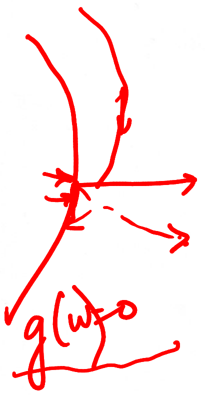




w^* is on the boundary of the feasible region

w^* is inside the feasible region

Constraint is active because $\underline{w}^*$ satisfies $g(\underline{w}^*) = 0$



α : scalar $\quad \nabla f(\underline{w}^*) = -\alpha \nabla g(\underline{w}^*)$

Hold at $\underline{w}^*$ $\rightarrow$ $\therefore$ $\boxed{\nabla f(\underline{w}^*) + \alpha \nabla g(\underline{w}^*) = 0}$

When there are several constraints $g_i(\underline{w}) \leq 0$, the feasible region is the intersection of all the regions $g_i(\underline{w}) \leq 0$. $\forall i$

Assuming that $\underline{f}$ and $\underline{g_i}$ are differentiable,

$$\rightarrow \quad \nabla f(\underline{w}^*) = -\underbrace{\sum_{i \in I} \alpha_i \nabla g_i(\underline{w}^*)}$$

At the optimal soln w^*

α_i : scalars.

For active constraints $\alpha_i \neq 0$

where I is the set of constraints which are active at $\underline{w}^*$

constraints satisfied with equality $g_i(\underline{w}^*) = 0$

$$\boxed{\nabla f(\underline{w}^*) + \sum_{i \in I} \alpha_i \nabla g_i(\underline{w}^*) = \underline{0}}$$

The Hard SVM learning rule specifies

$f(\underline{w}) = \underbrace{\frac{1}{2} \|\underline{w}\|^2}$ and $\underbrace{g_i(\underline{w}) \leq 0}$, $g_i = 1 - y_i \langle w, x_i \rangle$

The constraints g_i take the form $1 - y_i \langle \underline{w}, \underline{x}_i \rangle \le 0$

$$\nabla_{\underline{w}} f(\underline{w}^*) = \underline{w}$$

$$\nabla_{\underline{w}}\, g_i(\underline{w}^*) = -y_i\, \underline{x}_i$$

At the optimal solution we have

$$\nabla_{\underline{w}} f(\underline{w}^*) + \sum_{i \in I} \alpha_i \nabla g_i(\underline{w}^*) = \underline{0}$$

$$\underline{w}^* - \sum_{i \in I} \alpha_i\, y_i\, x_i = \underline{0}$$

$$\underline{w}^* = \sum_{i \in I} \alpha_i\, y_i\, \underline{x}_i$$

Optimal $\underline{w}^*$ can be specified using few of $\underline{x}_i$'s $\equiv$ supporting the hyperplane

Here I is the set of indices of the constraints which are active at $\underline{w}^*$.

$$I = \{ i : g_i(\underline{w}^*) = 0 \}$$

$$g_i = 1 - y_i \langle \underline{w}^*, \underline{x}_i \rangle \le 0 \quad = 0$$

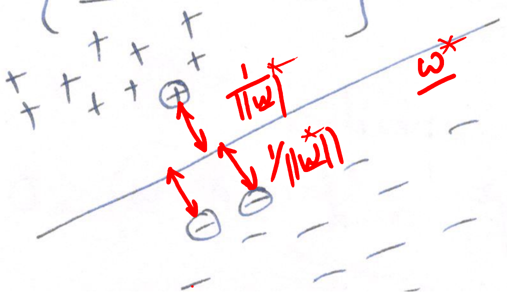
$$y_i \langle \underline{w}^*, \underline{x}_i \rangle = 1$$

Homogeneous representation

$$I = \{ i : |\langle \underline{w}^*, x_i \rangle| = 1 \}$$

examples $\underline{x}_i$ exactly at distance $\frac{1}{\|\underline{w}\|}$ from the hyperplane

Examples $\{ \underline{x}_i : i \in I \}$ are support vectors.



The circled examples are the support vectors.

How to solve for the α_i support vectors.

The hyper-plane can be specified in terms of the support vectors

$$\underline{w}^* = \sum_{i \in I} \alpha_i \, \underline{x}_i \, y_i \qquad \underline{x}_i \in \mathbb{R}^d$$

??

d: very large

Even if the data exists in a very high dimensional space, we need very few coefficients $\{\alpha_i\}$ to specify ~~it~~. the hyperplane.

Primal

The Dual Representation

The dual representation is suited when $m << d$.

m: # examples

d: dimensionality

To derive the dual representation, we start with the primal representation

Primal

$$\min_{\underline{w}} \underbrace{\frac{1}{2} \|\underline{w}\|^2}_{f(\underline{w})} \quad \text{such that} \quad \forall i \quad \boxed{y_i \langle \underline{w}, \underline{x}_i \rangle \geq 1}$$

$\underline{w} \in \mathbb{R}^d$.

OR.

$$\underbrace{1 - y_i \langle \underline{w}, \underline{x}_i \rangle}_{g_i(\underline{w})} \leq 0$$

$$\sum_{i=1}^{m} \alpha_i \, g_i(\underline{w})$$

Consider a function

$$\underline{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix}$$

m: #examples

$$g(\underline{w}) = \max_{\substack{\underline{\alpha} \in \mathbb{R}^m \\ \underline{\alpha} \geq \underline{0}}} \sum_{i=1}^{m} \alpha_i \left(1 - y_i \langle \underline{w}, \underline{x}_i \rangle\right)$$

$$= \begin{cases} 0 & \text{if } \forall i \quad 1 - y_i \langle \underline{w}, \underline{x}_i \rangle \leq 0 \\ \infty & \text{otherwise.} \end{cases}$$

$g_i(\underline{w}) \leq 0$

If any $g_i(\underline{w}) > 0$

$\sum_i \alpha_i \, g_i(\underline{w})$

switch: off 0. on ∞

We can rewrite the <u>primal</u> representation as:

$$\min_{\underline{w}} \left(f(\underline{w}) + \underline{g(\underline{w})} \right) \equiv \min f(\underline{w})$$

Minimization will try to adjust $\underline{w}$ so that $g(\underline{w})$ does not go to infinity.

$$\min_{\underline{w}} \left(f(\underline{w}) + \underbrace{\max_{\underline{\alpha}} \sum_{i=1}^{m} \alpha_i g_i(\underline{w}) \Big)}_{\text{switch}}$$

$$\min_{\underline{w}} \max_{\underline{\alpha}} \underbrace{\left(f(\underline{w}) + \sum_{i=1}^{m} \alpha_i g_i(\underline{w}) \right)}_{h(\underline{w}, \underline{\alpha})}$$

Max-min duality

Claim $\min_{\underline{w}} \max_{\underline{\alpha}} h(\underline{w}, \underline{\alpha}) \geq \max_{\underline{\alpha}} \min_{\underline{w}} h(\underline{w}, \underline{\alpha})$

Proof: Define a function Ω

$$\Omega(\underline{w}) = \max_{\underline{\alpha} \in \mathbb{R}^m} h(\underline{w}, \underline{\alpha})$$

For every $\underline{w}$ and every $\underline{\alpha}$ $\quad \Omega(\underline{w}) \geq h(\underline{w}, \underline{\alpha})$

$$\Rightarrow \min_{\underline{w}} \Omega(\underline{w}) \geq \min_{\underline{w}} h(\underline{w}, \underline{\alpha})$$

$$\min_{\underline{w}} \max_{\underline{\alpha}} h(\underline{w}, \underline{\alpha}) \geq \min_{\underline{w}} h(\underline{w}, \underline{\alpha})$$

Since this inequality holds for every value substituted for $\underline{\alpha}$ on the RHS, We can raise the RHS and still be assured of the inequality being satisfied

$$\therefore \quad \min_{\underline{w}} \max_{\underline{\alpha}} h(\underline{w}, \underline{\alpha}) \geq \max_{\underline{\alpha}} \min_{\underline{w}} h(\underline{w}, \underline{\alpha})$$

$$\min_{\underline{w}} \max_{\underline{\alpha}} f(\underline{w}) + \sum_{i=1}^{m} \alpha_i g_i(\underline{w}) \geq \max_{\underline{\alpha}} \min_{\underline{w}} f(\underline{w}) + \sum_{i=1}^{m} \alpha_i g_i(\underline{w})$$

For the hard SVM formulation

$$\min_{\underline{w}} \max_{\underline{\alpha}} \Big(\underbrace{\tfrac{1}{2} \|\underline{w}\|^2}_{f(w)} + \sum_{i=1}^{m} \alpha_i \underbrace{\big(1 - y_i \langle \underline{w}, \underline{x_i} \rangle\big)}_{g_i(w)} \Big) \geq \quad \text{duality}$$

$p^* \geq d^*$

p: primal

d: dual

$$\max_{\underline{\alpha}} \min_{\underline{w}} \Big(\tfrac{1}{2} \|\underline{w}\|^2 + \sum_{i=1}^{m} \alpha_i \big(1 - y_i \langle \underline{w}, \underline{x_i} \rangle\big) \Big)$$

It so turns out that in this case, even the strong duality exists.

If p^* is the optimal value returned by the primal formulation and d^* is the optimal value returned by the dual formulation

Then $\boxed{p^* = d^*}$

$\therefore$ we can solve for the optimization in the dual formulation

$$\max_{\substack{\underline{\alpha} \\ \underline{\alpha} \geq 0}} \underbrace{\min_{\underline{w}} \Big(\tfrac{1}{2} \|\underline{w}\|^2 + \sum_{i=1}^{m} \alpha_i \big(1 - y_i \langle \underline{w}, \underline{x_i} \rangle\big) \Big)}_{\text{unconstrained}}$$

Since the maximization is outside the minimization, we note that for a given $\underline{\alpha}$, the optimization w.r.t. $\underline{w}$ is unconstrained and the objective is differentiable.

For the optimum value of $\underline{w}$, the gradient vanishes

$$\underline{w} - \sum_{i=1}^{m} \alpha_i y_i \underline{x_i} = 0 \Rightarrow \boxed{\underline{w} = \sum_{i=1}^{m} \alpha_i y_i \underline{x_i}}$$

Most of the α_i terms are zero.

Substituting this optimum value of $\underline{w}$ in

$$\max_{\substack{\alpha \\ \alpha \geq 0}} \frac{1}{2} \|\underline{w}\|^2 + \sum_i \alpha_i \left(1 - y_i \langle \underline{w}, \underline{x_i} \rangle\right)$$

Please derive.

gives

$$\underline{w} = \sum_j \alpha_j y_j x_j$$

$$\max_{\substack{\alpha \\ \alpha \geq 0}} \left(\frac{1}{2} \underbrace{\Big\| \sum_{i=1}^{m} \alpha_i y_i \underline{x_i} \Big\|^2}_{\|\underline{w}\|^2} + \sum_{i=1}^{m} \alpha_i \Big(1 - y_i \big\langle \underbrace{\sum_j \alpha_j y_j \underline{x_j}}_{\underline{w}}, \underline{x_i} \big\rangle \Big) \right)$$

On rearranging, we get

$$\max_{\substack{\alpha \\ \alpha \geq 0}} \left[\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \underline{x_i}, \underline{x_j} \rangle \right]$$

dual formulation of SVM

$\underline{\alpha} \in \mathbb{R}^m$

Vector $\underline{x}$ occurs inside inner product

$\|\underline{v}\|^2 = \langle \underline{v}, \underline{v} \rangle$

This dual formulation is not aware of d.

$d = 100000$

$m = 200$ $\quad [\underline{x_1}, \underline{x_2} \dots \underline{x_m}]$

Input to this optimization is a matrix of size $m \times m$.

Gram Matrix 200×200 $(m \times m)$

$$\begin{array}{c|c} & j \rightarrow \\ & \begin{array}{cccc} 1 & 2 & \cdots & m \end{array} \\ \begin{array}{c} i=1 \\ 2 \\ \vdots \\ m \end{array} & \begin{bmatrix} \langle \underline{x_1}, \underline{x_1} \rangle & \langle \underline{x_1}, \underline{x_2} \rangle & \cdots & \langle \underline{x_1}, \underline{x_m} \rangle \\ \langle \underline{x_2}, \underline{x_1} \rangle & \langle \underline{x_2}, \underline{x_2} \rangle & \cdots & \langle \underline{x_2}, \underline{x_m} \rangle \\ \vdots & & & \\ \langle \underline{x_m}, \underline{x_1} \rangle & \langle \underline{x_m}, \underline{x_2} \rangle & \cdots & \langle \underline{x_m}, \underline{x_m} \rangle \end{bmatrix} \end{array}$$

$m \times m$ scalar matrix

Inner product : distance between the vectors