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

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

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

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

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

On rearranging, we get (Dual formulation)

dual formulation of SVM

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

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

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

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

$d = 100000$

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

This dual formulation is not aware of d.

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

Gram Matrix

200×200

$(m \times m)$

$m \times m$ scalar matrix

$$\begin{array}{c|c} & \quad 1 \qquad\qquad 2 \; (j \rightarrow) \qquad\qquad\qquad m \\ \hline \begin{array}{c} i=1 \\ 2 \\ \vdots \\ m \end{array} & \begin{bmatrix} \langle \underline{x_1}, \underline{x_1} \rangle & \langle \underline{x_1}, \underline{x_2} \rangle & \cdots & \langle \underline{x_1}, \underline{x_m} \rangle \\ \langle \underline{x_2}, \underline{x_1} \rangle & \langle \underline{x_2}, \underline{x_2} \rangle & \cdots & \langle \underline{x_2}, \underline{x_m} \rangle \\ \vdots & & & \\ \langle \underline{x_m}, \underline{x_1} \rangle & \langle \underline{x_m}, \underline{x_2} \rangle & \cdots & \langle \underline{x_m}, \underline{x_m} \rangle \end{bmatrix} \end{array}$$

Inner product : distance between the vectors

$m \times m$ Gram

Input space. $\langle \underline{x}, \underline{x}' \rangle$ $\quad$ $\phi(\underline{x})$: feature space.

$\boxed{\mathbb{R}^d \quad \underline{x} \text{ to } \phi(\underline{x}) \quad \mathbb{R}^D}$ $\quad$ $\underline{D >> d}$

Gram matrix $\quad \langle \phi(\underline{x}), \phi(\underline{x}') \rangle$

Similarity in the feature space

$\phi(\underline{x})$ is in D dim space.

$\therefore$ inner product in D dim space

$$K(\underline{x}, \underline{x}') = \langle \phi(\underline{x}), \phi(\underline{x}') \rangle$$

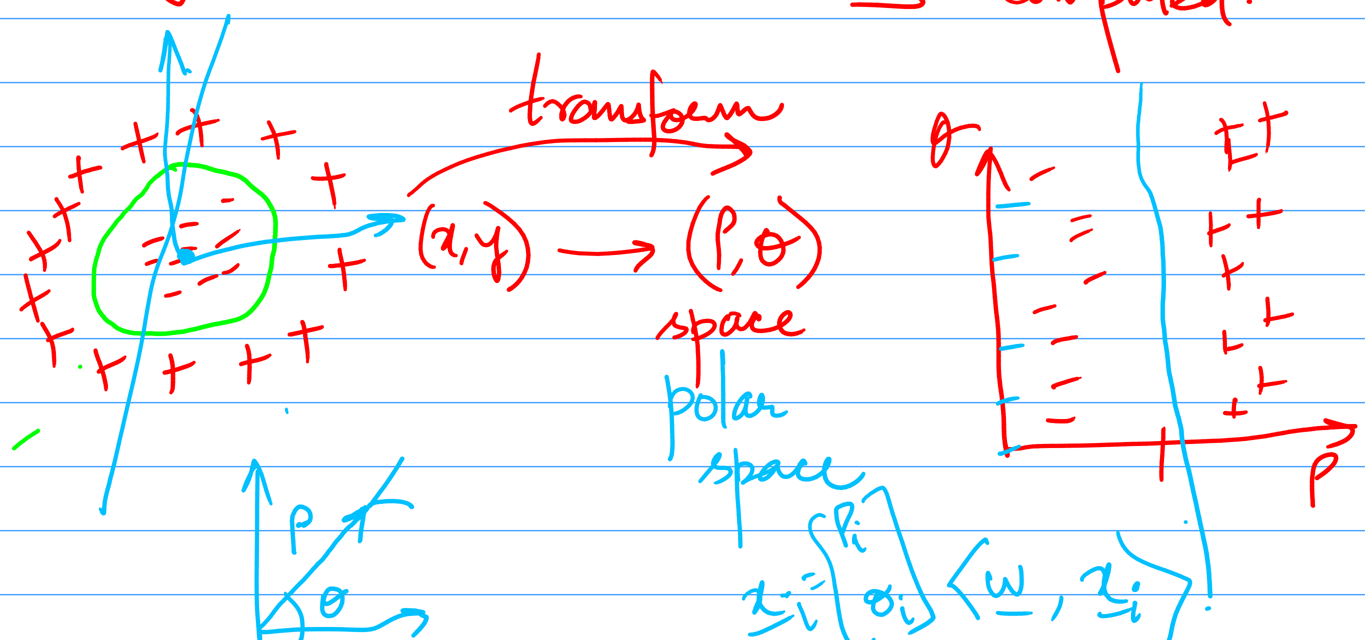
Kernel function.

gives a measure of similarity.

$\underline{x} \to \phi(\underline{x})$

$\underline{x}' \to \phi(\underline{x}')$

Explicit mapping need not be computed.



Once again consider the problem

$$\min \underline{f(\underline{w})} \quad \text{subject to} \quad \underline{g_i(\underline{w}) \le 0} \quad \forall i$$

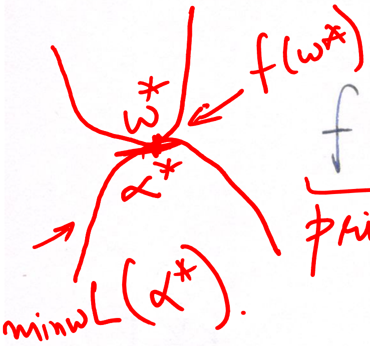
Using Lagrange parameters we reformulate it as

LHS $$f(\underline{w}^*) = \min f(\underline{w}) = \min_{\underline{w}} \max_{\underline{\alpha}} \left[f(\underline{w}) + \underline{\sum_i \alpha_i g_i(\underline{w})} \right]$$

strong dual. $$\equiv \max_{\underline{\alpha}} \underbrace{\min_{\underline{w}} \left[f(\underline{w}) + \sum_i \alpha_i g_i(\underline{w}) \right]}_{\min_w L(\underline{\alpha})}$$ when strong duality holds.

Let $\underline{w}^*$ be the optimal that minimizes LHS $\min_{\underline{w}} f(\underline{w})$

Let $\underline{\alpha}^*$ be the optimal solution of the RHS $\max_{\underline{\alpha}} \min_w L(\underline{\alpha})$



$$\underbrace{f(\underline{w}^*)}_{\text{primal.}} = \underline{\min_w L(\underline{\alpha}^*)} \ \text{dual.}$$

Lagrangian L

$$= \min_{\underline{w}} \boxed{\left(f(\underline{w}) + \sum_i \alpha_i^* g_i(\underline{w}) \right)}$$

$$\le f(\underline{w}^*) + \underbrace{\boxed{\sum_i \alpha_i^* g_i(\underline{w}^*)}}_{\le 0} = 0$$

$p^* = d^*$

saddle point

$$\le f(\underline{w}^*)$$

$\therefore$ the inequality $f(\underline{w}^*) \le f(\underline{w}^*)$ must be an equality.

and this can be ensured only if at the optimal solution.

(1) $$\boxed{\sum_i \underbrace{\alpha_i^*}_{\ge 0} \underbrace{g_i(\underline{w}^*)}_{\le 0} = 0}$$

$\sum(\text{terms}) = 0$

support vectors. examples $\underline{w} = \underline{\quad}$.

Because (2) $\alpha_i^* \geq 0 \quad \forall i$

Karush

Kuhn

Tucker

(KKT)

Conditions are valid at the optimal solution w^*

and (3) $g_i(\underline{w}^*) \leq 0 \quad \forall i$

we have (4) $\alpha_i^* g_i(\underline{w}^*) = 0 \quad \forall i$ examples.

Again consider

$$f(\underline{w}^*) = \min_w \mathcal{L}(\underline{\alpha}^*)$$

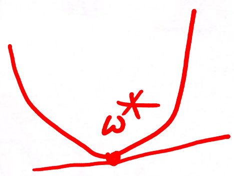
w^* is a minimizer of f.

$$= \min_{\underline{w}} \left(f(\underline{w}) + \sum_i \alpha_i^* g_i(\underline{w}) \right)$$

$$= f(\underline{w}^*) + \underbrace{\sum_i \alpha_i^* g_i(\underline{w}^*)}_{=0} = f(\underline{w}^*).$$

$\therefore$ $\underline{w}^*$ is the minimizer of

$$\boxed{f(\underline{w}) + \sum_i \alpha_i^* g_i(\underline{w})}$$

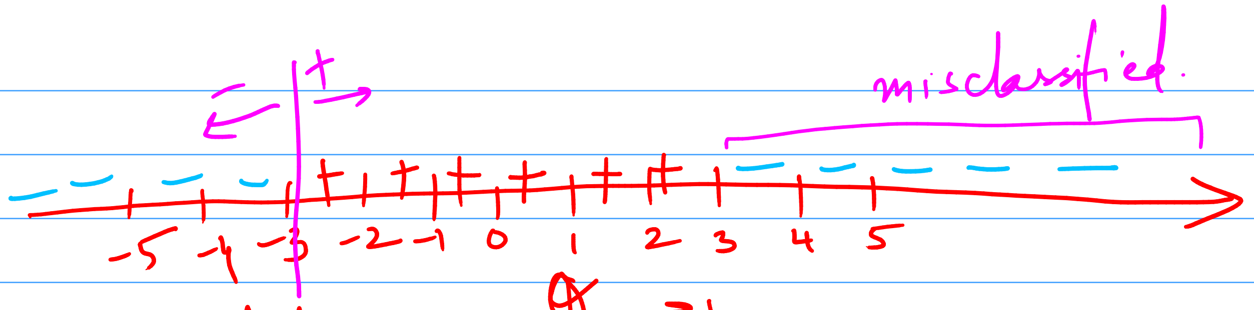


$\therefore$ the gradient will vanish at $\underline{w}^*$

(5) $$\boxed{\nabla f(\underline{w}^*) + \sum_i \alpha_i^* \nabla g_i(\underline{w}^*) = 0}$$

KKT Condition

(5) KKT conditions are necessary and sufficient to solve for the optimal $\underline{w}^*$ and $\underline{\alpha}^*$



1D data $- \leftarrow \theta \rightarrow +$

threshold θ: defines the separating hyperplane

Not linearly separable.

$$x \rightarrow \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

1D ⊕. 2D. ⊕

Transformed space.