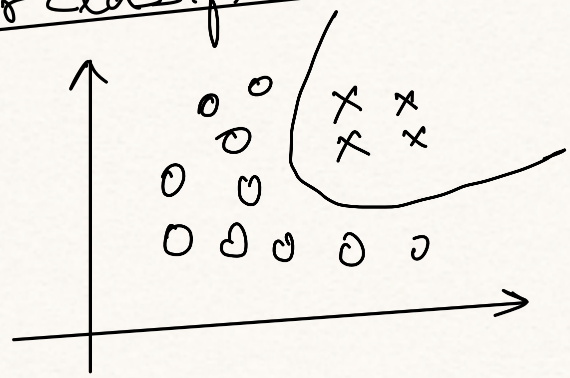
Kernel Methods

$$\underbrace{x}_{\text{attributes}} \to \underbrace{\phi(x)}_{\text{Features}} = \begin{bmatrix}1\\ x\\ x^2\\ \vdots\\ x^n\end{bmatrix}$$

For classification:



For L.R in G.D

$$\theta^{(t+1)} = \theta^{(t)} + \alpha \cdot \sum_{i=1}^{n} \left(y^{(i)} - \theta^T x^{(i)}\right) \cdot x^{(i)} \quad ; \theta \in \mathbb{R}^d$$

With a feature map,

$$\theta^{(t+1)} = \theta^{(t)} + \alpha \cdot \sum_{i=1}^{n} \left[y^{(i)} - \theta^T \phi(x^{(i)})\right] \cdot \phi(x^{(i)}) \quad ; \theta \in \mathbb{R}^p$$

$$\phi : \mathbb{R}^d \to \mathbb{R}^p$$

Eg: $\phi(x) = \begin{bmatrix}1\\ x_1\\ x_2\\ \vdots\\ x_1^2\\ x_1 x_2\\ x_1 x_3\\ \vdots\\ x_1^3\end{bmatrix}$ → monomial terms of order (≤ 3)

$p \approx O(d^3)$

Obs :

$$\theta^{(t)} = \sum_{i=1}^{n} \beta_i^{(t)} \cdot \phi(x^{(i)})$$

$$\theta^{(1)} = \sum_{i=1}^{n} \underbrace{\alpha y^{(i)}}_{\beta_i^{(1)}} \cdot \phi(x^{(i)})$$ → since $\theta^{(0)} = 0$.

$$\theta^{(t+1)} = \boxed{\theta^{(t)}} + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \boxed{\theta^{(t)}}^{T} \phi(x^{(i)}) \right) \cdot \phi(x^{(i)})$$

Induc^n

$$= \boxed{\sum_{i=1}^{n} \beta_i^{(t)} \cdot \phi(x^{(i)})} + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \left(\sum_{j=1}^{n} \beta_j^{(t)} \cdot \phi(x^{(j)}) \right)^{T} \phi(x^{(i)}) \right) \cdot \phi(x^{(i)})$$

$$= \sum_{i=1}^{n} \underbrace{\left[\beta_i^{(t)} + \alpha \left(y^{(i)} - \left(\sum_{j=1}^{n} \beta_j^{(t)} \phi(x^{(j)}) \right)^{T} \phi(x^{(i)}) \right) \right]}_{\beta_i^{(t+1)}} \cdot \phi(x^{(i)})$$

For $i = 1, 2, \ldots, n$

$$\beta_i^{(t+1)} = \beta_i^{(t)} + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j^{(t)} \underbrace{\phi(x^{(j)})^{T} \phi(x^{(i)})}_{\mathbb{R}} \right)$$

$x \quad d \quad x \in \mathbb{R}^d$

$x \in \mathcal{X} \qquad x \in \mathbb{R}^d \qquad \mathcal{X} = \mathbb{R}$

Kernel $\triangleq K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$

$$K(x, z) = \langle \phi(x), \phi(z) \rangle$$
$$= \phi(x)^T \phi(z)$$

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_1^3 \\ \vdots \end{bmatrix} \qquad p \approx O(d^3)$$

$$K(x, z) = 1 + \overbrace{\langle x, z \rangle}^{O(d)} + \overbrace{\langle x, z \rangle^2}^{O(d)} + \underbrace{\langle x, z \rangle^3}_{O(d)}$$

$$= \phi(x)^T \phi(z)$$

$O(d)$

Linear Regression (Kernelized)

1. Precompute

$$K_{ij} = K(x^{(i)}, x^{(j)})$$
$$= \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$$

2. Loop:

$$\forall i \in \{1, 2, \dots, n\}$$
$$\beta_i^{(t+1)} = \beta_i^{(t)} + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j^{(t)} K_{ij} \right)$$

$$\beta^{(t+1)} = \beta^{(t)} + \alpha \left(\vec{y} - K \beta^{(t)} \right)$$

Prediction

$$h_\theta(x) = \theta^T \phi(x)$$
$$= \sum_{i=1}^{n} \beta_i \phi(x^{(i)})^T \phi(x)$$
$$= \sum_{i=1}^{n} \beta_i K(x^{(i)}, x)$$

x – test example.

Obs:

① Train: $\beta := \beta + \alpha(\vec{y} - K\beta)$

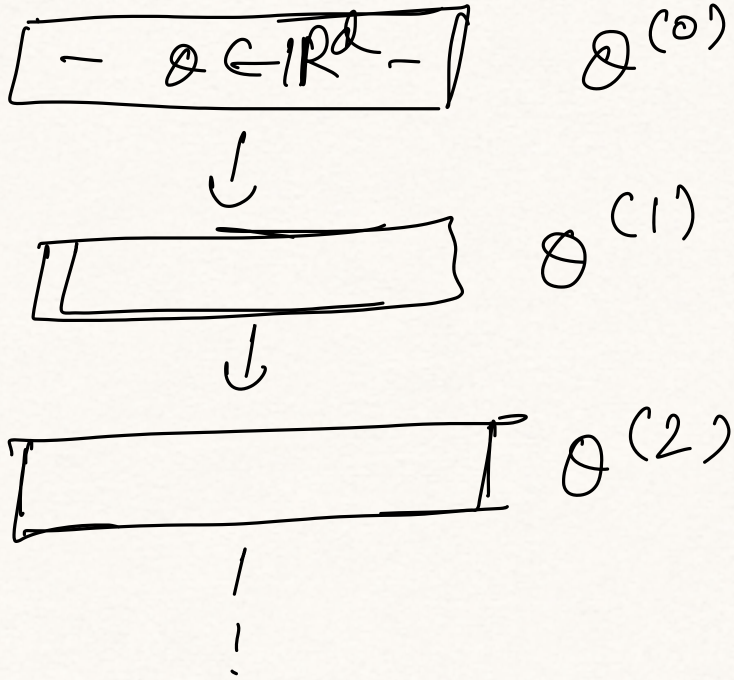
Test: $\hat{y} = \sum_{i=1}^{n} \beta_i K(x^{(i)}, x)$

$\Big\}$ $\phi(x)$ does not appear.

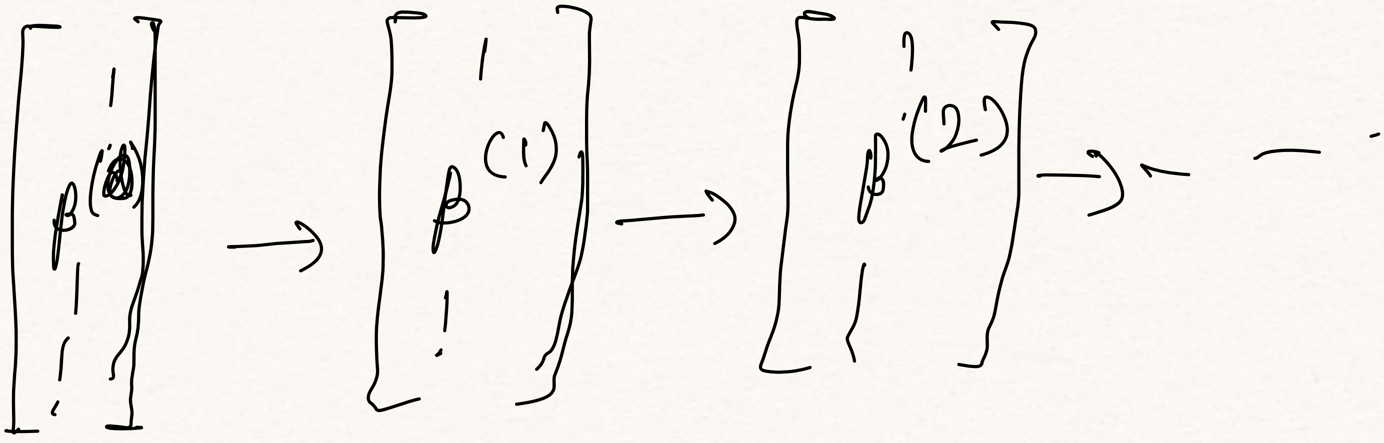
② For prediction we need training examples to be stored in memory.

$x \in \mathbb{R}^d$

$$\begin{bmatrix} \text{---}\, x^{(1)T} \,\text{---} \\ \text{------} \\ X \\ \text{------} \\ \text{---}\, x^{(n)T} \,\text{---} \end{bmatrix} \qquad \begin{bmatrix} \\ \\ \vec{y} \\ \\ \\ \end{bmatrix}$$



$\theta \in \mathbb{R}^d$



$\beta \in \mathbb{R}^n$

$x \in \mathbb{R}^d \Rightarrow \varphi(x) \in \mathbb{R}^p$

Kernel Examples

Eg: A) $K(x, z) = \langle x, z \rangle^2 \quad x \in \mathbb{R}^d$

$$\phi(x) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ \vdots \\ \vdots \\ x_d x_d \end{bmatrix}$$

B) $K(x, z) = (x^T z + c)^2$

$$\phi(x) = \begin{bmatrix} x_1^2 \\ x_1 x_2 \\ \vdots \\ \sqrt{2c} \cdot x_1 \\ \sqrt{2c} \cdot x_2 \\ \vdots \end{bmatrix}$$

$K(x, z) \uparrow$ (for similar x, z)

$\downarrow$ (for not similar x, z)

$K(x, z) = \phi(x)^T \phi(z)$

$$\boxed{K(x,z) = \exp\left(\frac{-\|x-z\|^2}{2\sigma^2}\right)}$$

↑ Gaussian Kernel

Necessary conditions for K to be Kernel

- K should be symmetric

$$K(x,z) = K(z,x) \qquad K(x,z) \triangleq \phi(x)^T \phi(z)$$

- $\{x^{(1)}, \dots, x^{(m)}\}$

$$K_{ij} = K(x^{(i)}, x^{(j)})$$

K is symmetric & P.S.D

$$z^T K z = \sum_i \sum_j z_i K_{ij} z_j$$

$$= \dots = \sum_k \left(\sum_i z_i \phi_k(x^{(i)})\right)^2$$

$$\geq 0$$

Mercer's Theorem

Let $K : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ be given

For K to be a Kernel, it is necessary & sufficient for any $\{x^{(1)}, \dots, x^{(m)}\}$ the corresponding Kernel matrix $K_{ij} = K(x^{(i)}, x^{(j)})$ is P.S.D & symm.

- To prove K is Kernel

① construct ϕ s.t. $K(\cdot) = \phi^T \phi$.

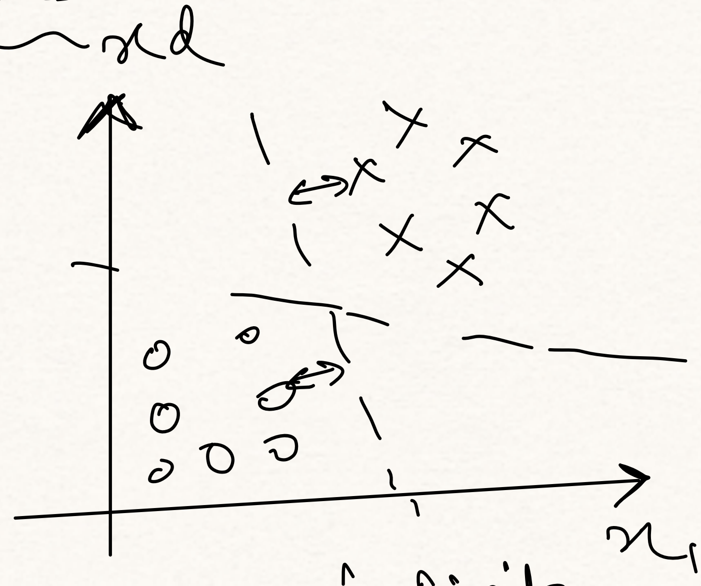
② Mercer's Thm.

$$\{x^{(i)}, \dots, x^{(m)}\}$$

$$K_{ij} = K(x^{(i)}, x^{(j)}) \text{ is P.S.D}$$

Support Vector Machines

- Discriminative
- Classification



infinite seperating hyperplanes possible

Notation

$y^{(i)} \in \{+1, -1\}$

parameters : w, b

$w \in \mathbb{R}^d \quad b \in \mathbb{R}$

$x \in \mathbb{R}^d$

$w^T x + b$

Functional Margin $= y^{(i)} (w^T x^{(i)} + b) > 0$

$w^T x + b > 0$ for $y = +1$

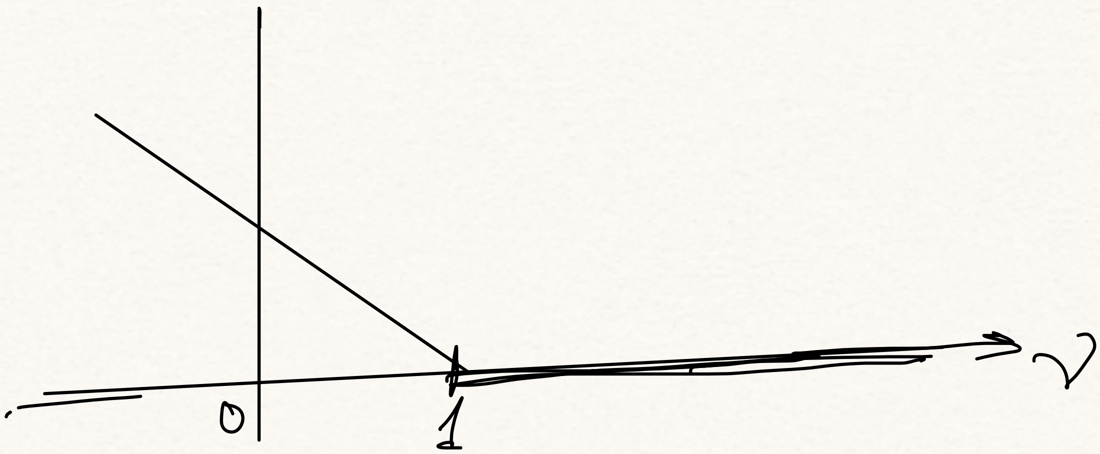
< 0 for $y = -1$

Desire : Margin to be large

→ Maximize the smallest margin.

$$\min_{w,b} \left(\underbrace{\sum_{i=1}^{n} \max\left[0,\ 1 - \underbrace{y^{(i)}\left(w^T x^{(i)} + b\right)}_{\text{margin} = \gamma^{(i)}}\right]} \right) + \frac{\|w\|^2}{C}$$

Hinge loss / SVM loss



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

$$\text{s.t.} \quad y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i \quad \forall\, i \in \{1, 2, \ldots, n\}$$

$$\xi_i \geq 0 \ ; \quad i = 1, 2, \ldots, n$$

Primal Convex Problem

$$\max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

s.t. $0 \le \alpha_i \le C$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

Dual convex problem