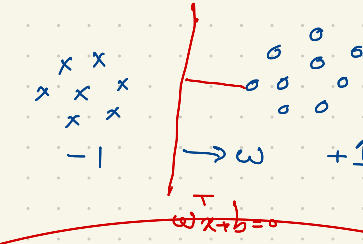
Lecture : 03/24



$$\gamma \triangleq \min_{i=1,\dots,N} \text{dist}\left(x^i, \omega^T x + b = 0\right)$$

$$\begin{cases} \max \ \gamma \\ \text{s.t.} \quad y^i\left(\omega^T x^i + b\right) \geq \gamma, \quad i = 1, \dots, N \end{cases} \equiv \begin{cases} \min\limits_{\omega, b} \ \frac{1}{2}\|\omega\|_2^2 \\ \text{s.t.} \quad y^i\left(\omega^T x^i + b\right) \geq 1 \quad \forall i = 1, \dots, N \end{cases}$$

Vanilla SVM

Training → Convex opt! ⇒ Solve using convex solvers!

Testing : x^{new} $\quad y(x^{new}) = \text{sgn}\left(\omega^{*T} x^{new} + b^*\right)$

Challenge : Could be costly to solve when d $\left(x^i \in \mathbb{R}^d\right)$ is large!

$\omega \in \mathbb{R}^d, \ b \in \mathbb{R} \Rightarrow$ # pars $= O(d)$

complexity of solver $O(d^3)$:(

How to overcome the computational bottleneck?

Build the Lagrangian function: [Reminder:

$$\min_{z} f(z) \quad \text{s.t. } g_1(z) \leq 0,\ g_2(z) \leq 0 \Rightarrow L(z,\alpha_1,\alpha_2) = f(z) + \underset{\geq 0}{\alpha_1} g_1(z) + \underset{\geq 0}{\alpha_2} g_2(z)$$

$$\frac{\partial L}{\partial z} = 0, \quad \frac{\partial L}{\partial \alpha_1} \leq 0, \quad \frac{\partial L}{\partial \alpha_2} \leq 0 \ ]$$

$$\begin{cases} \min_{\omega,b} \frac{1}{2}\|\omega\|_2^2 \\ \text{s.t. } y^i(\omega^T x^i + b) \geq 1, \quad \forall i = 1,\dots,N \end{cases} \Rightarrow \begin{cases} \min_{\omega,b} \frac{1}{2}\|\omega\|_2^2 \\ \text{s.t. } 1 - y^i(\omega^T x^i + b) \leq 0 \quad \forall i = 1,\dots,N \end{cases}$$

$$z = \binom{\omega}{b}$$

$\alpha_i \geq 0$ Lagrange multiplier

$$L(\omega,b,\alpha_1,\dots,\alpha_N) = \underbrace{\frac{1}{2}\|\omega\|_2^2}_{\frac{1}{2}\omega^T\omega} + \sum_{i=1}^{N} \alpha_i \left(1 - y^i(\omega^T x^i + b)\right)$$

$\alpha_i \in \mathbb{R}$, $y^i = \pm 1$

$$\boxed{-\alpha_i y^i x^{i^T} \omega}$$

$$\frac{\partial a^T \omega}{\partial \omega} = a$$

$$\frac{\partial L}{\partial \omega} = 0 \Rightarrow \omega + \sum_{i=1}^{N} -\alpha_i y^i x^i = 0$$

$$\Rightarrow \boxed{\omega^* = \sum_{i=1}^{N} \alpha_i^* y^i x^i}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} -\alpha_i y^i = 0 \longrightarrow \boxed{\sum_{i=1}^{N} \alpha_i^* y^i = 0}$$

[Skipping mathematical derivations of Lagrangian wrt α_i's]

plug back $w^* \longrightarrow L(w^*, b, \alpha_1, \dots, \alpha_N) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^i y^j (x^{i^T} x^j) \alpha_i \alpha_j + \sum_{i=1}^{N} \alpha_i$

unknowns

Dual SVM

$$\begin{cases} \max\limits_{\alpha_1, \dots, \alpha_N} -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^i y^j (x^{i^T} x^j) \alpha_i \alpha_j + \sum_{i=1}^{N} \alpha_i \\ \text{s.t.} \quad \sum_{i=1}^{N} \alpha_i y^i = 0 \ , \quad \alpha_i \geq 0 \quad \forall\, i = 1, \dots, N \end{cases}$$

$$\begin{cases} \min\limits_{w, b} \frac{1}{2} \|w\|_2^2 \longrightarrow d+1 \\ \text{s.t.} \quad y^i (w^T x^i + b) \geq 1 \end{cases}$$

only N unknowns

Convex optimization!

$d \gg N$

e.g. MRI scans (d : # pixels)
(N : # patients)

$O(N^3)$ (Naively) $\longrightarrow$ $O(N^2)$ intelligent implementations!

$\alpha_1^*, \dots, \alpha_N^*$

To classify, we need w^*, b^* to compute $\text{Sgn}(w^{*^T} x^{new} + b^*)$!

But from dual SVM we only have $\alpha_1^*, \dots, \alpha_N^*$.

$$\boxed{w^* = \sum_{i=1}^{N} \alpha_i^* y^i x^i}$$

$$\boxed{b^* = \frac{1}{2} \left[\min_{i: y^i = +1} w^{*^T} x^i - \max_{i: y^i = -1} w^{*^T} x^i \right]}$$

Kernel SVM :

Basis function expansion :

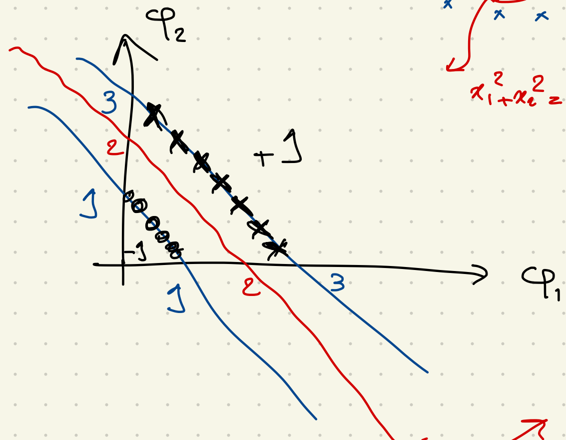
$$\varphi(x) = \begin{pmatrix} \varphi_1(x) \\ \varphi_2(x) \end{pmatrix} = \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix} \qquad \begin{pmatrix} x_1^2 \\ x_1 x_2 \\ x_2^2 \end{pmatrix}$$

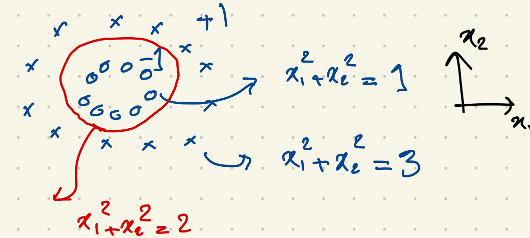
$+1$ x : $1x_1^2 + 1x_2^2 = 3$

$1\varphi_1 + 1\varphi_2 = 3$

-1 o : $1x_1^2 + 1x_2^2 = 1$

$1\varphi_1 + 1\varphi_2 = 1$





$$w^T\varphi(x) + b = 0$$

$$1\varphi_1 + 1\varphi_2 = 2$$

$$1x_1^2 + 1x_2^2 = 2$$

Vanilla SVM

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad y^i(w^T\varphi(x^i) + b) \geq 0 \quad \forall i = 1, \dots, N \end{cases}$$

Dual SVM

$$\begin{cases} \max\limits_{\alpha_1, \dots, \alpha_N} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y^i y^j \left(\varphi(x^i)^T\varphi(x^j)\right)\alpha_i\alpha_j + \sum_i \alpha_i \\ \text{s.t.} \quad \sum_{i=1}^{N}\alpha_i y^i = 0, \quad \alpha_i \geq 0 \ \forall i = 1, \dots, N \end{cases}$$

Challenge: $\varphi : x \longmapsto \varphi(x)$, $\quad \mathbb{R}^d \longrightarrow \mathbb{R}^{d'}$ $\qquad d' \gg d$

Vanilla SVM $\quad O(d'^3)$ ☹

Dual SVM ; have to compute $\varphi(x^i)^T \varphi(x^j)$

N^2 computations of ↗

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}_{d\times 1}, \quad \varphi(x) = \begin{pmatrix} x_1^2 \\ x_1x_2 \\ x_1x_3 \\ \vdots \\ x_1x_d \\ x_2^2 \\ x_2x_3 \\ \vdots \\ x_2x_d \\ \vdots \\ x_d^2 \end{pmatrix} = \begin{pmatrix} \vdots \\ x_\ell x_{\ell'} \\ \vdots \end{pmatrix}$$

$d \times d$

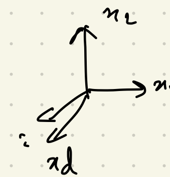
$\simeq d^2 \times 1$

$\ell, \ell' = 1, \dots, d$

$x_1 \times x_1 = x_1^2$

$x_1 \times x_2 = x_1 x_2$

$\vdots$



$\varphi(x^i)^T \varphi(x^j)$

$$\big(\ \big)^T_{d^2\times 1} \big(\ \big)_{d^2\times 1} = \underbrace{\big(\ \big)_{1\times d^2} \big(\ \big)_{d^2\times 1}}$$

$O(1 \times d^2 \times 1) = O(d^2)$

Computation of $\{\varphi(x^i)^T \varphi(x^j)\}_{i,j=1,\dots,N}$: $O(N^2 d^2)$ cost ☹

$$\varphi(x) = \begin{pmatrix} x_1^3 \\ x_1^2 x_2 \\ x_1^2 x_3 \\ \vdots \\ x_1^2 x_d \\ x_1 x_2 x_3 \\ x_1 x_2 x_4 \\ \vdots \\ x_d^3 \end{pmatrix} = \begin{pmatrix} \vdots \\ x_{\underset{d}{\ell}}\, x_{\underset{d}{\ell'}}\, x_{\underset{d}{\ell''}} \\ \vdots \end{pmatrix} \qquad \ell, \ell', \ell'' = 1, \dots, d$$

$\simeq d^3 \times 1$

$x_1 \times x_1 \times x_1 = x_1^3 \qquad \ell = \ell' = \ell'' = 1$

$x_1^1 \times x_2^1 \times x_5^1 \rightarrow 1+1+1=3 \qquad \ell = 1, \ell' = 2, \ell'' = 5$

$$\{\varphi(x^i)^T \varphi(x^j)\}_{i,j=1,\dots,N} \longrightarrow O(N^2 d^3)$$

$$\varphi(x) = \begin{pmatrix} x_1^n \\ x_1^{n-1} x_2 \\ \vdots \\ x_1^{n-1} x_d \\ x_1^{n-2} x_2 x_3 \\ \vdots \\ x_d^n \end{pmatrix} \simeq d^n \times 1 \longrightarrow \{\varphi(x^i)^T \varphi(x^j)\}_{i,j=1,\dots,N} \longrightarrow O(N^2 d^n)$$

Key question: Can we compute $\varphi(x^i)^T\varphi(x^j)$ implicitly, without explicitly computing $\varphi(x^i)\ \forall i$ and then taking their inner product? Yes!

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}, \quad \varphi(x) = \begin{pmatrix} x_1^2 \\ x_1x_2 \\ \vdots \\ x_1x_d \\ \vdots \\ x_d^2 \end{pmatrix} = \begin{pmatrix} \vdots \\ x_\ell x_{\ell'} \\ \vdots \end{pmatrix} \qquad \ell, \ell' = 1, 2, \dots, d$$

$$\varphi(\underbrace{x^i}_{\text{call it } x})^T\varphi(\underbrace{x^j}_{\text{call it } z}) \Rightarrow \varphi(x)^T\varphi(z)$$

$$\underline{\varphi(x)^T\varphi(z)} = \underbrace{\begin{pmatrix} \vdots \\ x_\ell x_{\ell'} \\ \vdots \end{pmatrix}^T}_{d^2\times 1} \underbrace{\begin{pmatrix} \vdots \\ z_\ell z_{\ell'} \\ \vdots \end{pmatrix}}_{d^2\times 1} = \sum_{\ell=1}^{d}\sum_{\ell'=1}^{d} x_\ell x_{\ell'} z_\ell z_{\ell'} = \sum_{\ell=1}^{d}\sum_{\ell'=1}^{d} (x_\ell z_\ell)(x_{\ell'} z_{\ell'})$$

$$\boxed{O(d^2)}$$

$$= \underbrace{\sum_{\ell=1}^{d} x_\ell z_\ell}_{x^Tz} \underbrace{\sum_{\ell'=1}^{d} x_{\ell'} z_{\ell'}}_{x^Tz} = x^Tz \times x^Tz = \underbrace{(x^Tz)^2}_{(\ )^T_{d\times 1}(\ )_{d\times 1}} = \boxed{O(d)}$$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \Longrightarrow \varphi(x) = \begin{pmatrix} x_1^n \\ x_1^{n-1}x_2 \\ \vdots \\ x_d^n \end{pmatrix}_{d^n \times 1}$$

$$\boxed{\begin{array}{c} \varphi(x)^T\varphi(z) \\ \| \\ (x^Tz)^n \end{array}} \begin{array}{l} \longrightarrow O(d^n) \\ \\ \longrightarrow O(d) \end{array}$$

Kernel SVM

$$\max_{\alpha_1,\dots,\alpha_N} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y^i y^j \Big(\underbrace{\cancel{\varphi(x^i)^T\varphi(x^j)}}_{K(x^i,x^j)}\Big)\alpha_i\alpha_j + \sum_{i=1}^{N}\alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{N}\alpha_i y^i = 0, \quad \alpha_i \geqslant 0 \quad \forall i=1,\dots,N$$

$$K(x,z) = \varphi(x)^T\varphi(z) \begin{cases} \xrightarrow{(2)} (x^Tz)^2 \\ \vdots \\ \xrightarrow{(n)} (x^Tz)^n \end{cases}$$

Kernel between x & z

$$K(x^i,x^j) = x^{i\,T}x^j$$

linear kernel

Choice of Kernel :

use cross_validation or hold out data to pick the best Kernel !

* $K(x, z) = (x^T z)^n$ n_th degree monomial $\longrightarrow$ $\varphi(x) = \begin{pmatrix} x_1^n \\ x_1^{n-1} x_2 \\ \vdots \\ x_d^n \end{pmatrix}$ associated BFE for this kernel

* $\boxed{K(x, z) = (x^T z + c)^n}$ monomials upto degree n

c: 0.1

$\varphi(x)^T \varphi(z)$

$$\varphi(x) = \begin{pmatrix} x_1^n \\ \vdots \\ x_d^n \\ x_1^{n-1} \\ \vdots \\ x_d^{n-1} \\ \vdots \\ x_1 \\ \vdots \\ x_d \\ 1 \end{pmatrix} \begin{matrix} \} \text{ degree } n \\ \\ \} \text{ degree } n-1 \\ \\ \} \text{ degree } 1 \\ \} \text{ degree } 0 \end{matrix}$$

* $\boxed{K(x, z) = e^{-\|x - z\|_2^2 / 2\sigma^2}}$ σ $\rightarrow$ hyperpar.

Gaussian / RBF Kernel