10

Q $K(x_i, x_j) = (x_i \cdot x_j + 1)^2$

$x = (x_1, x_2) \in R^2$ ~~then~~ where

$\phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1 x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$

For $K(x_i, x_j) = (x_i \cdot x_j + 1)^3$ what is ϕ function

→ ~~[illegible]~~

~~$K(x_i, x_j)$~~

~~$K(x_i, x_j) =$~~ ~~[illegible]~~

$x_i = (x_{i1}, x_{i2}) \quad x_j = (x_{j1}, x_{j2})$

$K(x_i x_j) = (x_{i1} x_{j1} + x_{i2} x_{j2} + 1)^3$

This can be rewritten as

$K(x_i, x_j) = (x_{i1} x_{j1} + x_{i2} x_{j2} + 1)^2 \cdot (x_{i1} x_{j1} + x_{i2} x_{j2} + 1)$

Expand :- $(x_{i1} x_{j1} + x_{i2} x_{j2} + 1)^2$

$(a+b+c)^2 = a^2 + 2ab + b^2 + 2ac + 2bc + c^2$

$a = x_{i1} x_{j1}$

$b = x_{i2} x_{j2}$

$c = 1$

$(x_{i1} x_{j1} + x_{i2} x_{j2} + 1) = x_{i1}^2 x_{j1}^2 + 2x_{i1} x_{j1} x_{i2} x_{j2} + x_{i2}^2 x_{j2}^2 + 2x_{i1} x_{j1} + 2x_{i2} x_{j2} + 1$

Multiply by $(x_{i1}x_{j1} + x_{i2}x_{j2} + 1)$

$(x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{j1}x_{i2}x_{j2} + x_{i2}^2 x_{j2}^2 + 2x_{i1}x_{j1} + 2x_{i2}x_{j2} + 1) \times (x_{i1}x_{j1} + x_{i2}x_{j2} + 1)$

$K(x_i, x_j) = x_{i1}^3 x_{j1}^3 + 3x_{i1}^2 x_{j1}^2 + 3x_{i1}x_{j1} + 3x_{i1}^2 x_{i2} x_{j1}^2 x_{j2} + 6x_{i1}x_{i2}x_{j1}x_{j2} + 3x_{i1}x_{j1}x_{i2}^2 x_{j2}^2 + 3x_{i2}x_{j2} + 3x_{i2}^2 x_{j2}^2 + x_{i2}^3 x_{j2}^3 + 1$

Rearranging,

$\phi(x) = (x_1^3, \sqrt{3}x_1^2, \sqrt{3}x_1, \sqrt{3}x_1^2 x_2, \sqrt{6}x_1 x_2, \sqrt{3}x_1 x_2^2, \cancel{\sqrt{3x_2^2 x_2^3 x_1}}, \sqrt{3}x_{j2}, \sqrt{3x^2}_2 x^3_{j2}, 1)$

2. $K(x, z) = c_1 k_1(x, z) + c_2 k_2(x, z)$ for $c_1, c_2 > 0$

$k_1(x, z) = (\phi_1(x), \phi_1(z))$
$k_2(x, z) = (\phi_2(x), \phi_2(z))$

$K(x, z) = c_1 k_1(x, z) + c_2 k_2(x, z)$:—

$\phi(x) = (\sqrt{c_1}\,\phi_1(x), \sqrt{c_2}\,\phi_2(x))$

~~...~~ $\phi(x)$ is concatenation of two feature $\phi_1(x)$ $\phi_2(x)$ which is scaled by $\sqrt{c_1}$ & $\sqrt{c_2}$

$$(\phi(x)\,\phi(z)) = (\sqrt{c_1}\,\phi_1(x), \sqrt{c_1}\,\phi_1(z)) + (\sqrt{c_2}\,\phi_2(x), \sqrt{c_2}\,\phi_2(z))$$

$$= c_1(\phi_1(x), \phi_1(z))\ c_2(\phi_2(x), \phi_2(z))$$

This is definition of kernel

$$K(x,z) = c_1 K_1(x,z) + c_2 K_2(x,z)$$

which tells us that this is a valid Kernel

Feature map:- $\phi(x) = (\sqrt{c_1}\,\phi(x), \sqrt{c_2}\,\phi_2(x))$

3.

a) in line 4

$$w \leftarrow w + \gamma\,(y_i - \sigma(w^T x_i))\,x_i$$

In each iteration, modified version of x_i is added based on pred error & the learning rate, resulting in weight vector which becomes linear combination of training data, with α_i telling us how much each training example contribute. So w^* can be expressed as weight sum of training ~~examples~~ examples

$$w^* = \sum_{i=1}^{N} \alpha_i x_i$$

b) By replacing the input vector with their feature mapping using a kernel fn we can kernalize stochastic gradient

Input :- $\{(x_i, y_i)\}_{i=1}^{N}$ where $x_i \in \mathbb{R}^d$ & $y_i \in \{0,1\}$

- Learning rate γ

- kernel function $K(x_i, x_j)$

Output :- Weight vector α

Steps :-

1. Initialize $\alpha = 0$
2. while not converged do
3. for $i = 1 \ldots N$ do

$z = \sum_{j=1}^{N} \alpha_j y_j K(x_i, x_j)$;

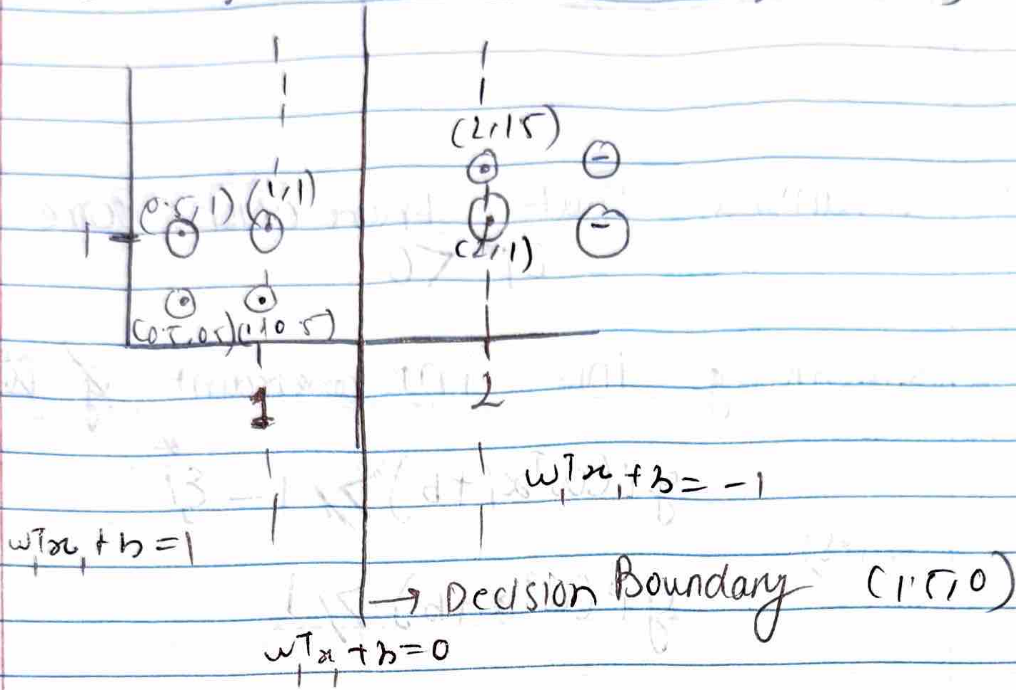
$\alpha_i \leftarrow \alpha_i + \gamma(y_i - \sigma(z))$

end

end

★ α captures influence of each training example on final model

4) (1,1) (1,0.5) (2,1.5) (2,1)



b) Since the ~~ret~~ decision boundary $w_1^T x_1 + b = 0$ is vertical, w_2 must be 0

so,

$$w_1 x_1 + b = 1$$
$$w_1 + b = 1$$
$$w_1 x_1 + b = -1$$
$$2w_1 + b = -1$$

so for

$$w_1 + b = 1$$
$$2w_1 + b = -1$$
$$-w_1 = 2$$
$$w_1 = -2$$

$$2(-2) + b = -1$$
$$b = -1 + 4 \qquad \boxed{b = 3}$$

$\boxed{w_1 = -2, w_2 = 0, b = 3}$

5)

a) Assuming that there exists some $\xi_i^* < 0$

Substituting this into constraint ~~$y_i(w^T x_i + b)$~~

$$y_i(w^T x_i + b) \geq 1 - \xi_i^*$$

gives,

$$y_i(w^T x + b) \geq 1$$

meaning if $\xi_i^* = 0$ we get optimal solution & constraint is satisfied

If we consider $\xi_i^* = 0$ we reduce the objective function value, meaning $\xi_i < 0$ can't be optimal solution.

So solution for optimal, each $\xi_i^* \geq 0$

b)

$$\frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i^2$$

constraint

$$y_i (w^T x_i + b) \geq 1 - \xi_i \quad \text{for all } i$$

Using Lagrange multiplier $\alpha_i \geq 0$ for each constraint, the Lagrangian $L(w, b, \xi)$ becomes:—

$$L(w, b, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i^2 + \sum_{i=1}^{N} \alpha_i (1 - \xi_i - y_i(w^T x_i + b))$$

c) Dual Problem:—

Taking partial Derivatives & setting them to zero

Respect to w:—

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

Respect to b:—

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$$

Respect to ξ_i

$$\frac{\partial L}{\partial \xi_i} \quad -2c\xi_i - \alpha_i = 0 \Rightarrow \xi_i = \frac{\alpha_i}{2c}$$

Substituting back to Lagrangian

$w = \sum_{i=1}^{N}$ & $\xi_i = \frac{\alpha_i}{2C}$

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1} \alpha_i \alpha_j y_i y_j x_i^T x_j + C \sum_{i=1}^{N}\left(\frac{\alpha_i}{2C}\right)^2 T$$

$$\sum_{i=1} \alpha_i - \sum_{i=1}^{N} \frac{\alpha_i^2}{2C} - \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$L(\alpha) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j - \frac{1}{4C} \sum_{i=1}^{M} \alpha_i^2 + \sum_{i=1}^{N} \alpha_i$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \& \quad \alpha_i \geq 0 \text{ for all } i$$

Comparing SVM and hinge loss

→ The objective for the L2 SVM includes $\frac{1}{4C}\sum_{i=1}^{N} \alpha_i^2$ term, which is ~~absent in~~ absent in standard SVM

→ L2 SVM is more sensitive to outliers because squaring the slack variable ξ_i causes penalty. Larger slack value has larger impact on objective of SVM

Hinge loss grows linearly. This quadratic growth can cause models to be more affected by outliers than the standard SVM.