AI 534: Machine Learning

Assignment II

Name : Nam Nguyen

ID : 934 - 422 - 327

Problem 1:

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^3, \qquad x = [x_1, x_2]$$

- It convenient to use another variables

$$x_i \rightarrow x = [x_1, x_2] \in \mathbb{R}^2$$
$$x_j \rightarrow z = [z_1, z_2] \in \mathbb{R}^2$$

- We need to calculate, $K(x, z) = (x^T z + 1)^3$

$$\Rightarrow K(x, z) = \sum_{k=0}^{3} \binom{3}{k} (x^T z)^k (1)^{3-k}$$

- Now, let's expand this expression:

$$K(x, z) = \binom{3}{0} (x^T z)^0 (1)^3 + \binom{3}{1} (x^T z)^1 (1)^2$$
$$+ \binom{3}{2} (x^T z)^2 (1)^1 + \binom{3}{3} (x^T z)^3 (1)^0$$
$$= 1 + 3(x^T z) + 3(x^T z)^2 + (x^T z)^3$$
$$= 1 + 3(x_1 z_1 + x_2 z_2) + 3(x_1 z_1 + x_2 z_2)^2$$
$$+ (x_1 z_1 + x_2 z_2)^3$$
$$= 1 + 3x_1 z_1 + 3x_2 z_2 + 3x_1^2 z_1^2 + 6x_1 x_2 z_1 z_2$$
$$+ 3x_2^2 z_2^2 + x_1^3 z_1^3 + 3 . x_1^2 z_1^2 x_2 z_2 + 3x_1 z_1 x_2^2 z_2^2$$
$$+ x_2^3 z_2^3$$

$$\Rightarrow K(x,z) = \langle [1 \quad \sqrt{3}x_1 \quad \sqrt{3}x_2 \quad \sqrt{3}x_1^2 \quad \sqrt{6}x_1x_2$$

$$\sqrt{3}x_2^2 \quad x_1^3 \quad \sqrt{3}x_1^2x_2 \quad \sqrt{3}x_1x_2^2 \quad x_2^3]$$

$$\cdot [1 \quad \sqrt{3}z_1 \quad \sqrt{3}z_2 \quad \sqrt{3}z_1^2 \quad \sqrt{6}z_1z_2 \quad \sqrt{3}z_2^2 \quad z_1^3$$

$$\sqrt{3}z_1^2z_2 \quad \sqrt{3}z_1z_2^2 \quad z_2^3] \rangle$$

$$= \langle \phi(x), \phi(z) \rangle$$

- Therefore, the corresponding ϕ function is

$$\phi(x) = (1, \sqrt{3}x_1, \sqrt{3}x_2, \sqrt{3}x_1^2, \sqrt{6}x_1x_2, \sqrt{3}x_2^2,$$
$$x_1^3, \sqrt{3}x_1^2x_2, \sqrt{3}x_1x_2^2, x_2^3)$$

Problem 2:

a) $K'(x, z) = c\,K(x, z)$ for $c > 0$

- This function is indeed a valid kenel. (Closure Property)
- $\phi'(x) = \sqrt{c}\,\phi(x)$
- It's a valid kernel since it can be represented in terms of the feature map $\phi(x)$

b) $K'(x, z) = c\,K(x, z)$ for $c < 0$

- This function is not a valid kenel.
- Let K is kerel matrix corresponding to $K(x, z)$

From Mercer theorem, we have

$$t^T K t \geq 0 \quad \text{for all } t \in \mathbb{R}^n$$

$\Rightarrow$ K' is kenel matrix corresponding to $K'(x, z)$ and

$$t^T K' t = t^T c \cdot K \cdot t = \underbrace{c}_{<0}\,\underbrace{t^T K t}_{\geq 0} \leq 0 \text{ for all } t \in \mathbb{R}^n$$

Hence, this one violates the Mercer theorem.

$K'(x, z)$ is not a valid Kernel

c) $K'(x,z) = c_1 K_1(x,z) + c_2 K_2(x,z)$ for $c_1, c_2 > 0$

- This function is indeed a valid kernel (Closure property)
- $\phi'(x) = \sqrt{c_1}\,\phi_1(x) + \sqrt{c_2}\,\phi_2(x)$
- It's a valid kernel since it can be represented in terms of feature maps.

d) $K'(x,z) = K_1(x,z)\, K_2(x,z)$

- This function is indeed a valid Kernel. (Closure property)
- $\phi'(x) = \phi_1(x) \otimes \phi_2(x)$, where $\otimes$ is the outer product.
- If ϕ_1 has N_1 features and ϕ_2 has N_2 features

Hence, ϕ' will have $(N_1 \times N_2)$ features:

$$\phi_{ij} = \phi_{1i} \cdot \phi_{2j}$$

Problem 3 :

a) • We can pay attention fore these lines :

for $i = 1, \dots, N$ do

$$\underline{w} \leftarrow \underline{w} + \gamma (y_i - \sigma(\underline{w}^T \underline{x}_i)) \underline{x}_i$$

end

• Let $\lambda_i = \gamma (y_i - \sigma(\underline{w}^T \underline{x}_i))$, we have

$$\underline{w} \leftarrow \underline{w} + \lambda_i \underline{x}_i$$

Hence, the solution $\underline{w}^*$ can be expressed as the weighted sum of training examples, $\underline{x}_i$.

b) • Let $\underline{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \underline{x}_i$.

• $$\underline{w}^T \underline{x}_i = \left(\sum_{j=1}^{N} \alpha_j y_j \underline{x}_j \right)^T \underline{x}_i$$

$$= \sum_{j=1}^{N} \alpha_j y_j \underline{x}_j^T \underline{x}_i$$

$$= \sum_{j=1}^{N} \alpha_j y_j K(\underline{x}_j, \underline{x}_i) \quad \text{(Kernelizing step)}$$

• We also have,

$$\underline{w} \leftarrow \underline{w} + \gamma (y_i - \sigma(\underline{w}^T \underline{x}_i)) \underline{x}_i$$

$$\Rightarrow \sum_{i=1}^{N} \alpha_i y_i \underline{x}_i \leftarrow \sum_{i=1}^{N} \alpha_i y_i \underline{x}_i + \gamma \left(y_i - \sigma\left(\sum_{j=1}^{N} \alpha_j y_j K(\underline{x}_j, \underline{x}_i) \right) \right) \underline{x}_i$$

$$\Rightarrow \cancel{\sum} \quad \alpha_i \leftarrow \alpha_i + \gamma \left(y_i - \sigma\left(\sum_{j=1}^{N} \alpha_j y_j K(\underline{x}_j, \underline{x}_i) \right) \right)$$

Hence, we will learn a continuous weights for $\underline{\alpha}$'s insted of $\underline{w}$'s

- Algorithm: Kernelized Stochastic Gradient Descent for Logistic Regression.

Input: $\{ (\underline{x}_i, y_i)_{i=1}^{N} \}$ (training data), γ (learning rate),

Output: learned weight vector $\underline{\alpha}$

Initialize $\underline{\alpha} = 0$;

while not conveged do

  for $i = 1, \dots, N$ do

    $u_i = \sum_{j=1}^{N} \alpha_j y_j K(\underline{x}_i, \underline{x}_j)$ ($K(\cdot,\cdot)$ is Kernel function)

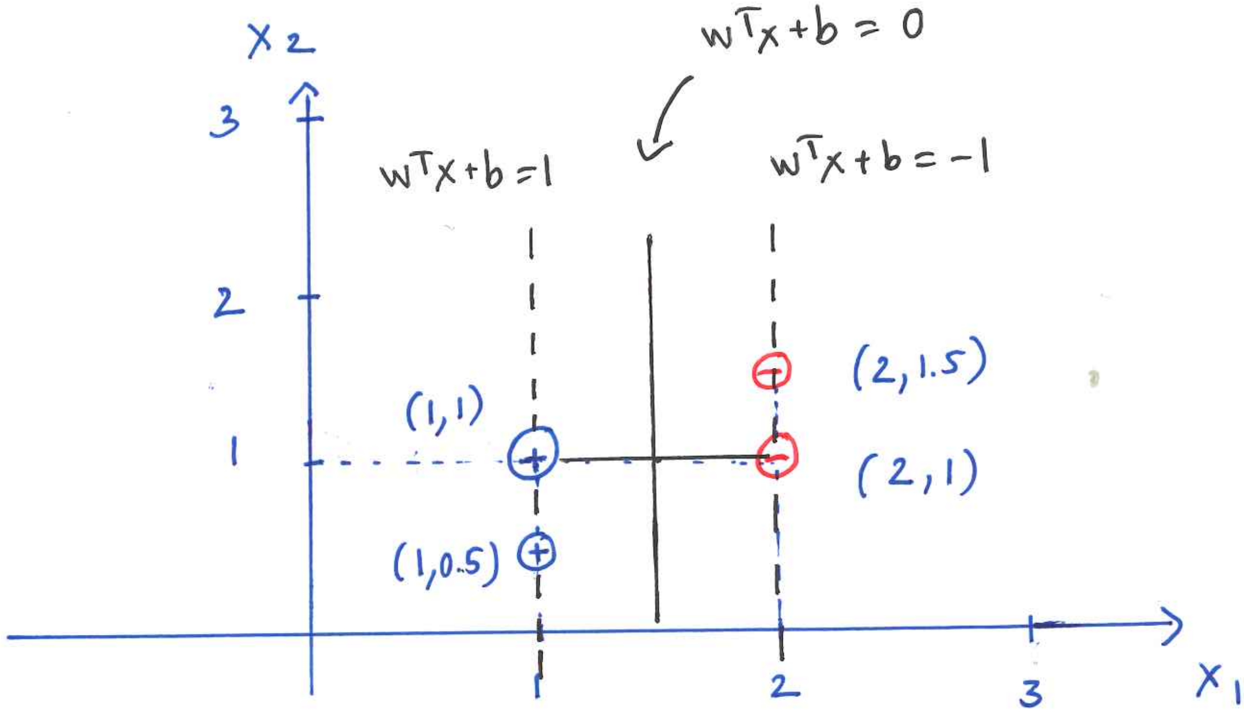
    $\alpha_i = \alpha_i + \gamma \left(y_i - \sigma(u_i) \right)$

  end

end

Problem 4 :

a)



b)

- From the section a), we see that the decision boundary is vertical, hence $w_2 = 0$. where $\underline{w} = [w_1, w_2]$
- Consider the support vector (1, 1)

$$\Rightarrow \quad w_1 + b = 1 \quad (1)$$

- Consider the support vector (2, 1)

$$\Rightarrow \quad 2w_1 + b = -1 \quad (2)$$

- From (1) and (2), we have

$$\begin{cases} w_1 = -2 \\ w_2 = 0 \\ b = 3 \end{cases}$$

Problem 5 :

$$\min_{\underline{w}, b, \xi} \frac{1}{2} \underline{w}^T \underline{w} + C \sum_{i=1}^{N} \xi_i^2$$

$$\text{s.t.} \quad y_i (\underline{w}^T \underline{x}_i + b) \geq 1 - \xi_i , \quad i \in \{1, \dots, N\}$$

$$\xi_i \geq 0, \quad i \in \{1, \dots, N\}$$

a) • Let $(\underline{w}^*, b^*, \underline{\xi}^*)$ be the optimal solution to the problem without the set of constraints, $\xi_i \geq 0$, $i \in \{1, \dots, N\}$

• Assume that $\xi_i^* < 0$ for some i. Hence,

$$y_i (\underline{w}^T \underline{x}_i + b) \geq 1 - \xi_i^* \geq 1 \text{ for some } i$$

This implys that $\left(\xi_i^* = 0\right)$ is the optimal solution of the problem. This one contradics to the assumption that $\left(\xi_i^* < 0\right)$. (Contradiction proof).

Therefore, $\xi_i^* \geq 0$ for all i.

b) . The Lagrangian is

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

where $\alpha_i \geq 0$.

c)

- Tak the gradient of $L(\underline{w}, b, \underline{\xi})$ with respect to $\underline{w}$, b, and $\underline{\xi}$ and set to zero:

$$\frac{\partial L(\underline{w}, b, \underline{\xi})}{\partial \underline{w}} = \underline{w} - \sum_{i=1}^{N} \alpha_i y_i \underline{x}_i = 0$$

$$\Rightarrow \quad \underline{w} = \sum_{i=1}^{N} \alpha_i y_i \underline{x}_i$$

$$\frac{\partial L(\underline{w}, b, \underline{\xi})}{\partial b} = -\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\Rightarrow \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\frac{\partial L(\underline{w}, b, \underline{\xi})}{\partial \xi_i} = 2c\,\xi_i - \alpha_i = 0$$

$$\Rightarrow \quad \xi_i = \frac{\alpha_i}{2c}$$

- Substitute $\underline{w} = \sum_{i=1}^{N} \alpha_i y_i \underline{x}_i$ and $\xi_i = \frac{\alpha_i}{2c}$ into $L(\underline{w}, b, \underline{\xi})$, we have

$$L(\underline{w}, b, \underline{\xi}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \underline{x}_i^T \underline{x}_j + c \sum_{i=1}^{N} \frac{\alpha_i^2}{4c^2}$$

$$\| $$

$$L(\underline{\alpha})$$

$$+ \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \frac{\alpha_i^2}{2c} - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \underline{x}_i^T \underline{x}_j + b \underbrace{\sum_{i=1}^{N} \alpha_i y_i}_{=0}$$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{4c} \sum_{i=1}^{N} \alpha_i^2 - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \underline{x}_i^T \underline{x}_j$$

- The dual problem is therefore

$$\begin{cases} \max\ L(\underline{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{4c} \sum_{i=1}^{N} \alpha_i^2 - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \underline{x}_i^T \underline{x}_j \\ \text{s.t. } \sum_i y_i \alpha_i = 0, \quad \alpha_i \geq 0 \text{ for } i = 1, \dots, N \end{cases}$$

- The differences:
  + The term $\left(- \frac{1}{4c} \sum_{i=1}^{N} \alpha_i^2\right)$
  + In this problem: ($\alpha_i \geq 0$ for $i = 1, \dots, N$)

    In the standard SVM problem: ($0 \leq \alpha_i \leq c$, for $i = 1, \dots, N$)

- This problem is more sensitive to outliers than the standard SVM problem. Because, ξ_i^2 affects much than ξ_i for outliers.