Name: Nicholas Guerrero
USCID: 4088107452

Problem 1 Kernel function **(32 points)**

Recall from lecture that there are two definitions of a kernel function, $k(\boldsymbol{x}, \boldsymbol{x}')$.

1. First k is called a kernel function if there exists a basis function $\phi : \mathbb{R}^D \to \mathbb{R}^M$ such that $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^T \phi(\boldsymbol{x}')$.
2. Second, we have Mercer's Theorem which states that k is a kernel function if and only if, for any set of $x_1, x_2, \cdots, x_n \in \mathbb{R}^D$, the resulting Gram matrix is PSD.

Throughout this problem, you can use either of these definitions to check or prove that some function is a valid kernel.

1.1 Consider the function $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}' + (\boldsymbol{x}^T \boldsymbol{x}')^2$ over $\boldsymbol{x} \in \mathbb{R}^2$. Is this a valid kernel function? Show why or why not. **(10 points)**

$$x = [x_1,\ x_2]\quad x' = \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}$$

$$\phi : R^2 \to R^M$$

$$k(x, x') = x^T x' + \left(x^T x'\right)^2 = x_1 x_1' + x_2 x_2' + \left(x_1 x_1' + x_2 x_2'\right)^2$$

$$= x_1 x_1' + x_2 x_2' + x_1^2 {x'}_1^2 + 2 x_1 x_2 x_1' x_2' + x_2^2 {x_2'}^2$$

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

$$\phi(x) = \left[x_1,\ x_2,\ x_1^2, \sqrt{2}\, x_1 x_2,\ x_2^2\right]$$

Is indeed a valid kernel function

1.2 Consider the function $k(\boldsymbol{x}, \boldsymbol{x}') = (f(\boldsymbol{x}) + f(\boldsymbol{x}'))^2$ for any function $f : \mathbb{R}^D \to \mathbb{R}$. Is this a valid kernel function? Show why or why not. **(12 points)**

consider $x_1,\ x_2$ *such that* $f(x_1) = a,\ f(x_2) = b$

$$f(x_1,\ x_1) = (f(x_1) + f(x_1))^2 = (a+a)^2 = (2a)^2 = 4a^2$$

$$K = \begin{pmatrix} 4a^2 & (a+b)^2 \\ (b+a)^2 & 4b^2 \end{pmatrix}$$

$$u^TKu = \begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} 4a^2 & (a+b)^2 \\ (b+a)^2 & 4b^2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

$$= \begin{pmatrix} 4a^2x + (a+b)^2y & (b+a)^2x + 4b^2y \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

$$= 4a^2x^2 + (a+b)^2xy + (b+a)^2xy + 4b^2y^2$$

$$= 4a^2x^2 + 4b^2y^2 + 2\,(a+b)^2xy$$

$$= 4a^2x^2 + 4b^2y^2 + 2*\left[a^2 + 2ab + b^2\right]*xy$$

$\therefore$ *not PSD because for a negative x and positive y or vice versa, the express would be negative. so* $\neg\forall u$ *is* $K \geqslant 0$

Alternative solution:

For a matrix to be PSD, its determinant $|K| \geq 0$, *if we set* $a = 0$, $b \neq 0$, *the determinant will equal* $-b^4 < 0$. $\therefore$ *Gram matrix is not PSD and this is not a valid kernel.*

1.3 Now, assume $k_1(\boldsymbol{x}, \boldsymbol{x}')$ and $k_2(\boldsymbol{x}, \boldsymbol{x}')$ are kernel functions. Prove by the Mercer Theorem (from lecture 5) that a linear combination $k(\boldsymbol{x}, \boldsymbol{x}') = \alpha k_1(\boldsymbol{x}, \boldsymbol{x}') + \beta k_2(\boldsymbol{x}, \boldsymbol{x}')$ for some $\alpha, \beta \geq 0$ is also a kernel function. **(10 points)**

$$k_3(x, x') = \alpha k_2(x, x') + \beta k_2(x, x')$$

$$\alpha k_2(x, x') + \beta k_2(x, x') = \alpha\left(u^Tk_1u\right) + \beta\left(u^Tk_2u\right) \geqslant 0$$

$$u^Tk_3u = \alpha\left(u^Tk_1u\right) + \beta\left(u^Tk_2u\right) \geqslant 0$$

Thus k_3 *is a PSD matrix*

Problem 2 Support Vector Machines **(32 points)**

Consider the dataset consisting of points (x, y), where x is a real value, and $y \in \{-1, 1\}$ is the class label. Let's start with three points $(x_1, y_1) = (-1, -1)$, $(x_2, y_2) = (1, -1)$, $(x_3, y_3) = (0, 1)$.

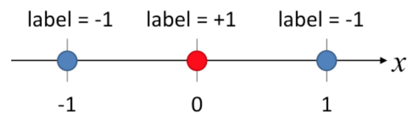


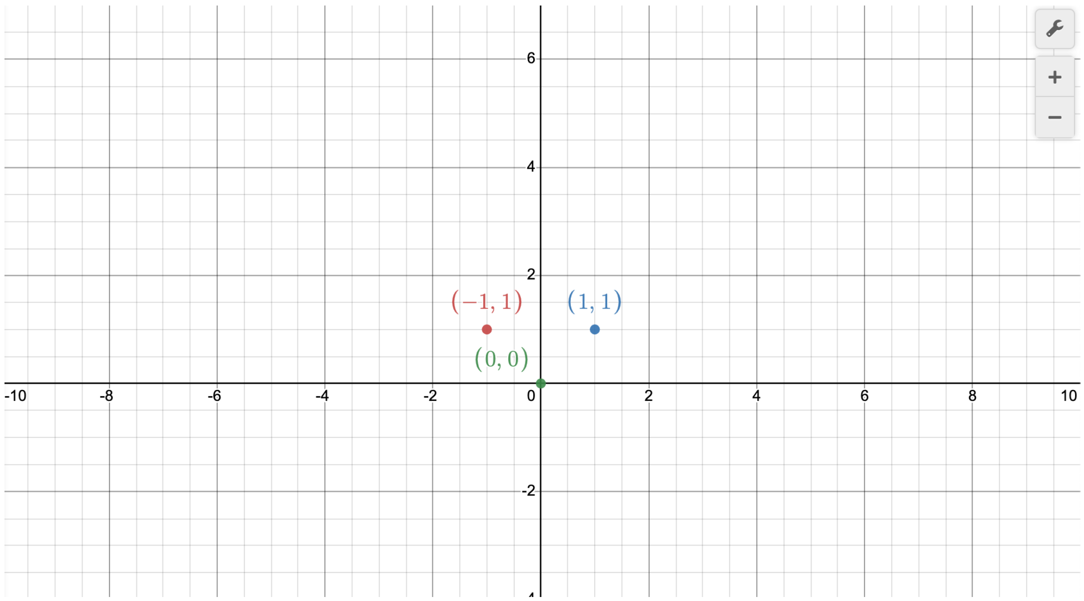
Figure 1: Three data points considered in Problem 2

2.1 Can three points shown in Figure 1, in their current one-dimensional feature space, be perfectly separated with a linear separator? Why or why not? **(4 points)**

No. Since the margin is the smallest distance from all training points to the hyperplane, the best seperating hyperplane would lie right on the point $(x_3,\ y_3)$. This implies that the margin has a value of 0. $If\ y\left[w^T\phi(x)\ +\ b\right]\ \geqslant\ 0$ we can make a correct prediction, but since margin is 0 $w^T\ =0$. When $y=-1$ the points would not be able to be classified currently by the equation above $-b\ \ngeqslant\ 0$. This is the best we can do in a one-dimensional feature space and so the data cannot be seperated with a linear seperator.

2.2 Now we define a simple feature mapping $\phi(x) = [x, x^2]^T$ to transform the three points from one- to two-dimensional feature space. Plot the transformed points in the new two-dimensional feature space. Is there a linear decision boundary that can separate the points in this new feature space? Why or why not? **(4 points)**

$$
\begin{array}{lll}
\phi(x) & \Longrightarrow & \left[x,\ x^2\right]^T \\
(-1,\ -1) & \Longrightarrow & (-1,\ 1)\ \ A\ \ label = \ -1 \\
(1,\ -1) & \Longrightarrow & (1,\ 1)\ \ B\ \ label = \ -1 \\
(0,\ 1) & \Longrightarrow & (0,\ 0)\ \ C\ \ label = \ \ 1
\end{array}
$$



Yes clearly there is a linear decision boundary that can seperate the points since in this picture above the point with a label of 1 is in the positive region of the vertical plane and the points with a label of -1 are on the other side. It is easy to imagine a hyperplane that could seperate them with some non-zero positive margin.

2.3 Given the feature mapping $\phi(x) = [x, x^2]^T$, write down the 3×3 kernel (or Gram) matrix **K** for the three data points. Show that this Gram matrix is positive semi-definite. Write the Kernel function K(x,y)(defined as $K(x, y) = \phi(x)^T\phi(y)$). **(8 points)**

$$\phi(x_1) = [-1, 1]^T \quad , \quad \phi(x_2) = [1, 1]^T, \quad \phi(x_3) = [0 , 0]^T$$

$$k = \begin{pmatrix} \phi(x_1)^T\phi(x_1) & \phi(x_1)^T\phi(x_2) & \phi(x_1)^T\phi(x_3) \\ \phi(x_2)^T\phi(x_1) & \phi(x_2)^T\phi(x_2) & \phi(x_2)^T\phi(x_3) \\ \phi(x_3)^T\phi(x_1) & \phi(x_3)^T\phi(x_2) & \phi(x_3)^T\phi(x_3) \end{pmatrix}$$

$$= \begin{pmatrix} -1*(-1)+1*1 & -1*1+1*1 & -1*0+1*0 \\ 1*(-1)+1*1 & 1*1+1*1 & 1*0+1*0 \\ 0*(-1)+0*1 & 0*1+0*1 & 0*0+0*0 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$u^T\begin{pmatrix}2&0&0\\0&2&0\\0&0&0\end{pmatrix}u \;=\; \begin{pmatrix}a & b & c\end{pmatrix}\begin{pmatrix}2&0&0\\0&2&0\\0&0&0\end{pmatrix}\begin{pmatrix}a\\b\\c\end{pmatrix} \;=\; \begin{pmatrix}a*2 & b*2 & 0\end{pmatrix}\begin{pmatrix}a\\b\\c\end{pmatrix} \;=\; a^2*2 \;+\; b^2*2 \;\geqslant\; 0$$

$$Thus\ \forall u\quad u^T\begin{pmatrix}2&0&0\\0&2&0\\0&0&0\end{pmatrix}u \;\geqslant\; 0\ and\ PSD$$

$$k(x,\ y) \;=\; x\cdot y \;=\; x_1y_1$$

$$\phi(x)^T\phi(y) \;=\; x_1y_1 + x_1^2y_1^2 \;= x_1y_1 \;+\; (x_1y_1)^2 \;=\; x\cdot y \;+\; (x\cdot y)^2 \;=\; k(x,\ y)$$

2.4 Write down the dual formulation of this problem (plugging in the numerical values evaluated using the kernel function). **(8 points)**

$$\max_{\{\alpha_n\}} \;\sum_n \alpha_n \;-\; \frac{1}{2}\sum_{m,\,n} y_m y_n \alpha_m \alpha_n k(x_m, x_n)$$

$$s.t.\;\; 0 \;\leqslant\; \alpha_n,\;\; \forall n$$

$$\sum_n \alpha_n y_n = 0$$

$$plug\ in\ values:\quad k(x_1, x_1) \;=\; 2$$

$$(x_1,\ y_1) \;=\; (-1,\ -1)$$
$$(x_2,\ y_2) \;=\; (1,\ -1)$$
$$(x_3,\ y_3) \;=\; (0,\ 1)$$

$$\max_{\{\alpha_n\}} \;\sum_n \alpha_n \;-\; \frac{1}{2}\sum_{m,\,n} y_m y_n \alpha_m \alpha_n k(x_m, x_n)$$

$$= \;\max_{\alpha_1, \alpha_2,\ \alpha_3 \geqslant 0} \quad \alpha_1 + \alpha_2 \;+\; \alpha_3 \;-\frac{1}{2}[y_1y_1\alpha_1\alpha_1k(x_1,x_1) \;+ y_2y_2\alpha_2\alpha_2k(x_2,x_2) \;+ y_3y_3\alpha_3\alpha_3k(x_3,x_3)]$$

$$= \max_{\alpha_1, \alpha_2,\ \alpha_3 \geqslant 0} \quad \alpha_1 + \alpha_2 \ + \ \alpha_3 \ - \frac{1}{2}[\alpha_1 \alpha_1 2 \ + \alpha_2 \alpha_2 2 \ + \alpha_3 \alpha_3 0]$$

$$= \max_{\alpha_1, \alpha_2,\ \alpha_3 \geqslant 0} \quad \alpha_1 + \alpha_2 \ + \ \alpha_3 \ - \alpha_1^2 \ - \ \alpha_2^2$$

$$s.t. \quad \alpha_1(-1) \ + \ \alpha_2(-1) \ + \ \alpha_3(1) \ = \ -\alpha_1 \ -\alpha_2 + \alpha_3 \ = \ 0$$
$$\implies \alpha_3 \ = \ \alpha_1 \ + \ \alpha_2$$

2.5 Solve the dual form analytically. Then obtain primal solution $\mathbf{w}^*, b^*$ using dual solution. **(8 points)**

$$w^* \ = \ \sum_n \alpha_n^* y_n \phi(x_n) \ = \ \sum_{n\,:\,\alpha_n > 0} \alpha_n^* y_n \phi(x_n)$$

$$b^* \ = \ y_n \ - \ {w^*}^T \phi(x_n) \ = \ y_n \ - \ \sum_m y_m \alpha_m^* k(x_m, x_n)$$

$$using\ \alpha_3 \ = \ \alpha_1 \ + \ \alpha_2\ an\ substituting\ into\ objective\ function\ we\ obtain$$
$$obj : f \ = \ max_{\{\alpha_1,\, \alpha_2\, \geqslant 0\}}\ 2\alpha_1 + 2\alpha_2 \ \ -\alpha_1^2 \ - \ \alpha_2^2$$

$$\frac{\partial f}{\alpha_1} \ = \ 2 \ -2\alpha_1 \ = \ 0 \ \implies \ \alpha_1 = 1$$

$$\frac{\partial f}{\alpha_2} \ = \ 2 \ - \ 2\alpha_2 \ = \ 0 \ \implies \alpha_2 = 1$$

$$so\ \alpha_3 = 1 + 1 \ = \ 2 \ \implies \ \alpha_3 \ = \ 2$$

$$primal\ solutions:$$
$$b^* \ = \ y_n \ - \ {w^*}^T \phi(x_n) \ = \ y_n \ - \ \sum_m y_m \alpha_m^* k(x_m, x_n)$$

$b^* = y_1 - y_1\alpha_1^* 2 = -1 - (-1)*1*2 = -1 + 2 = 1$

$$w^* = \sum_{n=1}^{3} \alpha_n^* y_n \phi(x_n) = [1, -1]^T + [-1, -1]^T = [0, -2]^T$$

$\alpha_1 y_1 \phi(x_1) = 1*(-1)*[-1, 1]^T = [1, -1]$

$\alpha_2 y_2 \phi(x_2) = 1*(-1)*[1, 1]^T = [-1, -1]$

$\alpha_3 y_3 \phi(x_3) = 2*(1)*[0, 0]^T = [0, 0]$

Problem 3 Constrained Optimization (36 points)

Machine learning problems, especially clustering problems, sometimes involve optimization over a **simplex**. In this exercise, you will solve two optimization problems over the simplex. Recall a $K-1$ dimensional simplex Δ is defined as:

$$\Delta = \{\boldsymbol{q} \in \mathbb{R}^K | q_k \geq 0, \forall k \text{ and } \sum_{k=1}^{K} q_k = 1\},$$

which means that a $K-1$ dimensional simplex has K non-negative entries, and the sum of all K entries is 1. This property coincides with the property of the probability distribution of a discrete random variable of K possible outcomes. Thus, the simplex is usually seen in clustering problems.

3.1 Given $a_1, ..., a_K \in \mathbb{R}_{\neq 0}$ (the set of non-zero real numbers), solve the following optimization over the simplex. (find the optimal value $\boldsymbol{q}^*$ of $\boldsymbol{q}$) **(18 points)**

$$\arg\max_{\boldsymbol{q}\in\Delta} \sum_{k=1}^{K} a_k^2 \ln q_k$$

(a) Write down the Lagrangian of this problem. (Hint: use the constraints on q_k given by the simplex Δ) **(4 points)**

$$L(q, \alpha, \lambda_k) = \sum_{k=1}^{K} a_k^2 ln(q_k) + \sum_{k=1}^{K} \lambda_k q_k + \alpha\left(\sum_{k=1}^{K} q_k - 1\right)$$

where the lagrangian multiples are $\lambda_1, \lambda_2, ..., \lambda_k \geqslant 0$ and $\alpha \neq 0$

(b) Apply KKT conditions on the Lagrangian you derived above to find $\boldsymbol{q}^*$. (Hint: the solution can be written in the form of $q_k^* = ...$) **(12 points)**

(1) *apply stationarity*: $\nabla L(q, \alpha, \{\lambda_k\}) = 0$

$$\frac{\partial L}{\partial q} = \sum_{k=1}^{K} \frac{a_k^2}{q_k} + \sum_{k=1}^{K} \lambda_k + \alpha$$

$$\implies \textit{for each } k : \frac{a_k^2}{q_k^*} + \lambda_k + \alpha = 0$$

$$\frac{a_k^2}{q_k^*} = -(\lambda_k + \alpha) \implies q_k^* = -\frac{a_k^2}{\lambda_k + \alpha} \neq 0$$

since $a_1, \ldots, a_k$ *are non* $-$ *zero real numbers,* $\lambda_1, \lambda_2, \ldots, \lambda_k \geqslant 0$ *and* $\alpha \neq 0$

(2) *apply complimentary slackness* : $\lambda_k q_k^* = 0$

$$\implies \textit{implies } \lambda_k = 0 \textit{ since } q_k^* \neq 0$$

(3) *apply Feasibility Conditions* : $\displaystyle\sum_{k=1}^{K} q_k^* = 1$

$$q_k^* = -\frac{a_k^2}{\lambda_k + \alpha} \implies q_k^* = -\frac{a_k^2}{\alpha}$$

$$\implies \sum_{k=1}^{K} \left(-\frac{a_k^2}{\alpha}\right) = 1$$

$$\implies \alpha = \sum_{k=1}^{K} -a_k^2$$

$$q_k^* = -\frac{a_k^2}{\sum_{k=1}^{K} -a_k^2} = \frac{a_k^2}{a_k^2} = 1$$

(c) The solution you acquired will not have a simple form if a_k is allowed to be 0. Explain why. (Hint: point out the relevant variable. One sentence explanation is sufficient) **(2 points)**

the denominator of the expression $q_k^* = \frac{a_k^2}{a_k^2}$ ill be undefined as the denominator will be 0

3.2 Next given $c_1, ..., c_K \in \mathbb{R}$, solve the following optimization problem following the same steps in part 1.1. $\boldsymbol{q}$ is under the same constraints as in part 1.1: **(18 points)**

$$\arg\max_{q \in \Delta} \sum_{k=1}^{K} (q_k c_k - q_k \ln q_k)$$

$$(a.\ )L(q,\ \alpha,\ \{\ \lambda_k\ \}\ ) = \sum_{k=1}^{K} (q_k c_k - q_k ln(q_k)) + \sum_{k=1}^{K} \lambda_k q_k + \alpha \left(\sum_{k=1}^{K} q_k - 1 \right)$$

where the lagrangian multiples are $\lambda_1, \lambda_2, ..., \lambda_k \geqslant 0$ *and* $\alpha \neq 0$

$(b.)$ *apply stationarity*: $\nabla L(q,\ \alpha,\ \{\lambda_k\}) = 0$

$$\frac{\partial L}{\partial q} = \sum_{k=1}^{K} c_k + \sum_{k=1}^{K} (lnq_k + 1) + \sum_{k=1}^{K} \lambda_k + \alpha$$

for each k : $c_k + lnq_k + 1 + \lambda_k + \alpha = 0$

$lnq_k = -(c_k + \lambda_k + \alpha + 1)$

$\implies q_k^* = e^{-(c_k + \lambda_k + \alpha + 1)}$

apply complimentary slackness: $\lambda_k q_k^* = 0$

$\implies$ *implies* $\lambda_k = 0$ *since* $q_k^* \neq 0$

$$q_k^* = e^{-(c_k + \alpha + 1)}$$

apply Feasibility Conditions : $\sum_{k=1}^{K} q_k^* = 1$

$$\sum_{k=1}^{K} e^{-(c_k + \alpha + 1)} = 1 \implies -(c_k + \alpha + 1) = ln(1) \implies - \alpha = 0 + c_k + 1$$

$$\implies \alpha = c_k + 1$$

$$\implies q_k^* = e^{-(2c_k + \lambda_k + 2)}$$