CS 771A: Intro to Machine Learning, IIT Kanpur				**Endsem Exam** (28 Apr 2024)
Name	MELBO			**40 marks** Page **1** of **4**
Roll No	240007	Dept.	AWSM	

Instructions:

1. This question paper contains 2 pages (4 sides of paper). Please verify.
2. Write your name, roll number, department in **block letters** with **ink** on **each page**.
3. Write your final answers neatly **with a blue/black pen**. Pencil marks may get smudged.
4. Don't overwrite/scratch answers especially in MCQ – ambiguous cases will get 0 marks.

Q1. (**True-False**) Write **T** or **F** for True/False (write **only in the box on the right-hand side**). You must also give a brief justification for your reply in the space provided below. (**3x(1+2) = 9 marks**)

1	Given any N values $a_1, \ldots, a_N > 0$, the optimization problem $\min_{x\in\mathbb{R}} \sum_{i\in[N]} (x-a_i)^3$ always has a solution x^* that is finite i.e. $-\infty < x^* < \infty$.	F
For any value of $x < 0$, we have $\sum_{i\in[N]} (x-a_i)^3 \leq -N\lvert x\rvert^3$ i.e., the objective diverges as we approach $x \to -\infty$.		
2	There exists a unique feature map $\phi_{\text{lin}}: \mathbb{R}^2 \to \mathbb{R}^D$ for the linear kernel such that $\langle \phi_{\text{lin}}(\mathbf{x}), \phi_{\text{lin}}(\mathbf{y}) \rangle = \langle \mathbf{x}, \mathbf{y} \rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$. If T, justify. If F, give two maps (may use same or different values of D for the two maps) that yield the same kernel value.	F
Let $\mathbf{x} \overset{\text{def}}{=} (x_1, x_2) \in \mathbb{R}^2$, then the maps $\phi_1(\mathbf{x}) = (x_1, x_2) \in \mathbb{R}^2$ as well as $\phi_2(\mathbf{x}) = (x_2, x_1) \in \mathbb{R}^2$ both yield the linear kernel value. Other possible maps that also yield the linear kernel include maps of the form $\phi_3(\mathbf{x}) = [a \cdot \mathbf{x}, b \cdot \mathbf{x}] \in \mathbb{R}^4$ where $a^2 + b^2 = 1$, or even maps of the form $\phi_3(\mathbf{x}) = A\mathbf{x} \in \mathbb{R}^2$ where $A \in \mathbb{R}^{2\times 2}$ is any rotation matrix (square symmetric orthonormal) i.e., $A^\top = A$ and $A^\top A = I$.		
3	If $f, g: \mathbb{R} \to \mathbb{R}$ are such that there may be points where either or both functions are not differentiable, then their sum $f + g$ can never be differentiable everywhere.	F
Consider the ReLU function $r(x) \overset{\text{def}}{=} \max\{x, 0\}$ and let $f(x) \overset{\text{def}}{=} r(x)$ and $g(x) \overset{\text{def}}{=} -r(-x)$. Both functions have a kink at $x = 0$ and are not differentiable everywhere. However, their sum is simply $f(x) + g(x) = x$ which is differentiable everywhere.		

Q2. (Kernel Smash) $K_1, K_2, K_3: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are Mercer kernels with 2D feature maps i.e., for any $x, y \in \mathbb{R}$, $K_i(x,y) = \langle \phi_i(x), \phi_i(y) \rangle$ with $\phi_1(x) = (1, x), \phi_2(x) = (x, x^2)$ and $\phi_3(x) = (x^2, x^3)$. Design a feature map $\phi_4: \mathbb{R} \to \mathbb{R}^7$ for kernel K_4 s.t. $K_4(x,y) = \left(K_1(x,y) + K_3(x,y)\right)^2 + K_2(x,y)$ for any $x, y \in \mathbb{R}$. No derivation needed. **Note that ϕ_4 must not use more than 7 dimensions. If your solution does not require 7 dimensions leave the rest of the dimensions blank.** **(5 marks)**

$$\phi_4(x) =$$

$$\left(\boxed{1}, \boxed{\sqrt{3}x}, \boxed{2x^2}, \boxed{2x^3}, \boxed{\sqrt{3}x^4}, \boxed{\sqrt{2}x^5}, \boxed{x^6}\right)$$

Q3. (Total confusion) The *confusion matrix* is a very useful tool for evaluating classification models. For a C-class problem, this is a $C \times C$ matrix that tells us, for any two classes $c, c' \in [C]$, how many instances of class c were classified as c' by the model. In the example below, $C = 2$, there were $P + Q + R + S$ points in the test set where P, Q, R, S are strictly positive integers. The matrix tells us that there were Q points that were in class $+1$ but (incorrectly) classified as -1 by the model, S points were in class -1 and were (correctly) classified as -1 by the model, etc. **Give expressions for the specified quantities in terms of P, Q, R, S.** No derivations needed. Note that y denotes the true class of a test point and $\hat{y}$ is the predicted class for that point. **(6 x 1 = 6 marks)**

		Predicted class $\hat{y}$	
		$+1$	-1
True class y	$+1$	P	Q
	-1	R	S

Confusion Matrix

$\mathbb{P}[\hat{y} = y | \hat{y} = 1]$ $\boxed{\frac{P}{P+R}}$

$\mathbb{P}[\hat{y} \neq y | \hat{y} = 1]$ $\boxed{\frac{R}{P+R}}$

$\mathbb{P}[y = 1 | \hat{y} = 1]$ $\boxed{\frac{P}{P+R}}$

$\mathbb{P}[\hat{y} = 1 | y = 1]$ $\boxed{\frac{P}{P+Q}}$

$\mathbb{P}[\hat{y} = y | y = 1]$ $\boxed{\frac{P}{P+Q}}$

$\mathbb{P}[\hat{y} \neq y | y = 1]$ $\boxed{\frac{Q}{P+Q}}$

Q4 (Opt to Prob) Melbo has come across an anomaly detection algorithm that, given a set of data points in 2D, $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$, solves the following optimization problem:

$$\underset{\mathbf{c} \in \mathbb{R}^2, r \geq 0}{\operatorname{argmin}} \, r \;\; \text{s.t.} \; \|\mathbf{x}_i - \mathbf{c}\|_\infty \leq r \text{ for all } i \in [n]$$

where for any vector $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2$, we have $\|\mathbf{v}\|_\infty \stackrel{\text{def}}{=} \max\{|v_1|, |v_2|\}$. Melbo's friend Melba thinks that this is just an MLE solution but Melbo is doubtful. To convince Melbo, create a likelihood distribution $\mathbb{P}[\mathbf{x} \mid \mathbf{c}, r]$ over the 2D space $\mathbb{R}^2$ with parameters $\mathbf{c} \in \mathbb{R}^2, r \geq 0$ such that $\left[\underset{\mathbf{c} \in \mathbb{R}^2, r \geq 0}{\arg\max} \left\{\prod_{i \in [n]} \mathbb{P}[\mathbf{x}_i \mid \mathbf{c}, r]\right\}\right] = \left[\underset{\mathbf{c} \in \mathbb{R}^2, r \geq 0}{\arg\min} \, r \;\; \text{s.t.} \; \|\mathbf{x}_i - \mathbf{c}\|_\infty \leq r \;\; \text{for all } i \in [n]\right]$. **Your solution must be a proper distribution i.e.,** $\mathbb{P}[\mathbf{x} \mid \mathbf{c}, r] \geq 0$ for any $\mathbf{x} \in \mathbb{R}^2$ and $\int_{\mathbf{x} \in \mathbb{R}^2} \mathbb{P}[\mathbf{x} \mid \mathbf{c}, r] \, d\mathbf{x} = 1$. Give calculations to show why your distribution is correct. **(4 + 6 = 10 marks)**

Write down the density function of your likelihood distribution here.

$$\mathbb{P}[\mathbf{x} \mid \mathbf{c}, r] = \begin{cases} \frac{1}{4r^2} & \text{if } \|\mathbf{x}-\mathbf{c}\|_\infty \leq r \\ 0 & \text{if } \|\mathbf{x}-\mathbf{c}\|_\infty > r \end{cases}$$

The region $\{\mathbf{x}: \|\mathbf{x}-\mathbf{c}\|_\infty \leq r\}$ has an area of $4r^2$ as it is a square of side length $2r$ centered at $\mathbf{c}$ which justifies the normalization constant while defining the uniform distribution over the region.

Give calculations showing why your likelihood distribution does indeed result in the optimization problem as MLE.

Using the above likelihood distribution expression yields the following likelihood value

$$\prod_{i\in[n]} \mathbb{P}[\mathbf{x}_i \mid \mathbf{c}, r] = \begin{cases} \left(\frac{1}{4r^2}\right)^n & \text{if } \|\mathbf{x}_i-\mathbf{c}\|_\infty \leq r \text{ for all } i \in [n] \\ 0 & \text{if } \exists i \text{ such that } \|\mathbf{x}_i-\mathbf{c}\|_\infty > r \end{cases}$$

Thus, likelihood drops to 0 if any data point is outside the square. Since we wish to maximize the likelihood, we are forced to ensure that $\|\mathbf{x}_i-\mathbf{c}\|_\infty > r$ does not happen for any $i \in [n]$. This yields the following optimization problem for the MLE

$$\underset{\mathbf{c}\in\mathbb{R}^2, r\geq 0}{\arg\max} \left(\frac{1}{4r^2}\right)^n \text{ s.t. } \|\mathbf{x}_i-\mathbf{c}\|_\infty \leq r \text{ for all } i \in [n]$$

Since $f(x) \overset{\text{def}}{=} \left(\frac{1}{4x^2}\right)^n$ is a decreasing function of x for all $x \geq 0$ as $n, 4$ are constants, maximizing $f(x)$ is the same as minimizing x. This yields the following problem concluding the argument.

$$\underset{\mathbf{c}\in\mathbb{R}^2, r\geq 0}{\arg\min}\, r \text{ s.t. } \|\mathbf{x}_i-\mathbf{c}\|_\infty \leq r \text{ for all } i \in [n]$$

Q5 (Positive Linear Regression) We have data features $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ and labels $y_1, \ldots, y_N \in \mathbb{R}$ stylized as $X \in \mathbb{R}^{N\times D}, \mathbf{y} \in \mathbb{R}^N$. We wish to fit a linear model with non-negative coefficients:

$$\underset{\mathbf{w}\in\mathbb{R}^D,}{\operatorname{argmin}} \frac{1}{2}\|X\mathbf{w}-\mathbf{y}\|_2^2 \text{ s.t. } w_j \geq 0 \text{ for all } j \in [D]$$

1. Write down the Lagrangian for this optimization problem by introducing dual variables.
2. Simplify the dual problem (eliminate $\mathbf{w}$) – show major steps. Assume $X^\top X$ is invertible.
3. Give a coordinate descent/ascent algorithm to solve the dual. **(2 + 3 + 5 = 10 marks)**

Write down the Lagrangian here (you will need to introduce dual variables and give them names)

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}) = \frac{1}{2}\|X\mathbf{w}-\mathbf{y}\|_2^2 - \boldsymbol{\lambda}^\top \mathbf{w}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^D$ is a dual variable that will be constrained to only have non-negative coordinates.

Derive and simplify the dual. Show major calculations steps.

The dual is simply $\max_{\boldsymbol{\lambda}\geq\mathbf{0}}\left\{\min_{\mathbf{w}}\mathcal{L}(\mathbf{w},\boldsymbol{\lambda})\right\}$. Setting $\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = 0$ gives us $X^\top(X\mathbf{w}-\mathbf{y})-\boldsymbol{\lambda}=\mathbf{0}$ i.e., we get $\mathbf{w}=(X^\top X)^{-1}(X^\top\mathbf{y}+\boldsymbol{\lambda})$. Putting this back into the dual gives us the simplified dual

$$\max_{\boldsymbol{\lambda}\geq\mathbf{0}}\left\{\frac{1}{2}\left\|X(X^\top X)^{-1}(X^\top\mathbf{y}+\boldsymbol{\lambda})-\mathbf{y}\right\|_2^2-\boldsymbol{\lambda}^\top\,(X^\top X)^{-1}(X^\top\mathbf{y}+\boldsymbol{\lambda})\right\}$$
$$\equiv\max_{\boldsymbol{\lambda}\geq\mathbf{0}}\left\{\frac{1}{2}\|\mathbf{r}+B\boldsymbol{\lambda}\|_2^2-\boldsymbol{\lambda}^\top\mathbf{a}\;-\boldsymbol{\lambda}^\top\,C\boldsymbol{\lambda}\right\}\equiv\min_{\boldsymbol{\lambda}\geq\mathbf{0}}\left\{\frac{1}{2}\boldsymbol{\lambda}^\top\,C\boldsymbol{\lambda}+\boldsymbol{\lambda}^\top\mathbf{q}\right\}$$

where $C\overset{\text{def}}{=}(X^\top X)^{-1}\in\mathbb{R}^{D\times D}, \mathbf{a}\overset{\text{def}}{=}CX^\top\mathbf{y}\in\mathbb{R}^D, B\overset{\text{def}}{=}XC\in\mathbb{R}^{N\times D}, \mathbf{r}\overset{\text{def}}{=}(XCX^\top-I)\mathbf{y}\in\mathbb{R}^N$ and $\mathbf{q}\overset{\text{def}}{=}\mathbf{a}-\mathbf{B}^\top\mathbf{r}$. Note that $B^\top B=C$ and we have removed terms that do not depend on $\boldsymbol{\lambda}$.

Give a coordinate descent/ascent algorithm to solve the dual problem.

Choosing a coordinate $j\in[D]$ and focusing only on objective terms and constraints involving λ_j

$$\min_{\lambda_j\geq 0}\left\{\frac{1}{2}C_{jj}\lambda_j^2+\lambda_j\left(q_j+\sum\nolimits_{k\neq j}(C_{jk}+C_{kj})\lambda_k\right)\right\}$$

Note that $C_{jk}=C_{kj}$ since the matrix C is symmetric. Using the QUIN trick gives the optimal value

$$\lambda_j=\max\left\{0,-\frac{1}{C_{jj}}\cdot\left(q_j+2\sum\nolimits_{k\neq j}C_{jk}\lambda_k\right)\right\}$$

Coordinate descent can now cycle through the coordinates in some fashion (say random or else random permutation) and update the chosen coordinate keeping all others fixed.