CS 771A: Intro to Machine Learning, IIT Kanpur				**Endsem Exam** (14 July 2023)
Name	MELBO			**40 marks**
Roll No	230007	Dept.	AWSM	Page **1** of **4**

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 may get 0 marks.

Q1. (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. **(5 x 1 = 5 marks)**

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

Confusion Matrix

Accuracy (**ACC**) $\mathbb{P}[\hat{y} = y]$	$\frac{P+S}{P+Q+R+S}$
Precision (**PRE**) $\mathbb{P}[y = 1 \mid \hat{y} = 1]$	$\frac{P}{P+R}$
Recall (**REC**) $\mathbb{P}[\hat{y} = 1 \mid y = 1]$	$\frac{P}{P+Q}$
False discovery rate (**FDR**) $\mathbb{P}[y = -1 \mid \hat{y} = 1]$	$\frac{R}{P+R}$
False omission rate (**FOR**) $\mathbb{P}[y = 1 \mid \hat{y} = -1]$	$\frac{Q}{Q+S}$

Q2. (Kernel Smash) Melbi has created two Mercer kernels $K_1, K_2: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with the feature map for the kernel K_i being $\phi_i: \mathbb{R} \to \mathbb{R}^2$. Thus, for any $x, y \in \mathbb{R}$, we have $K_i(x, y) = \langle \phi_i(x), \phi_i(y) \rangle$ for $i \in \{1,2\}$. Melbi knows that $\phi_1(x) = (x, x^3)$ and $\phi_2(x) = (1, x^2)$. Melbo has created a new kernel K_3 using Melbi's kernels so that for any $x, y \in \mathbb{R}$, $K_3(x, y) = \left(K_1(x, y) + 3 \cdot K_2(x, y)\right)^2$. Design a feature map $\phi_3: \mathbb{R} \to \mathbb{R}^7$ for the kernel K_3. Write your answer only in the pace given below. No derivation needed. **Note that ϕ_3 must not use more than 7 dimensions. If your solution does not require 7 dimensions leave the rest of the dimensions blank.** **(5 marks)**

$$\phi_3(x) = \left(\boxed{3}, \boxed{x\sqrt{6}}, \boxed{x^2\sqrt{19}}, \boxed{x^3\sqrt{12}}, \boxed{x^4\sqrt{11}}, \boxed{x^5\sqrt{6}}, \boxed{x^6}\right)$$

Q3 (Opt to Prob) Melbo enrolled in an advanced ML course and learnt an unsupervised learning technique called support vector data description (SVDD). Given a set of data points, say in 2D for sake of simplicity, $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^2$, SVDD solves the following optimization problem:

$$\min_{\mathbf{c} \in \mathbb{R}^2, r \in \mathbb{R}} r^2 \text{ s.t. } \|\mathbf{x}_i - \mathbf{c}\|_2^2 \leq r^2 \text{ for all } i \in [n]$$

Melbo's friend Melba saw this and exclaimed that this is just an MLE solution. To convince Melbo, create a likelihood distribution $\mathbb{P}[\mathbf{x} \mid \mathbf{c}, r]$ over the 2D space $\mathbb{R}^2$ with parameters $\mathbf{c} \in \mathbb{R}^2, r \geq 0$ s.t. $\left[\underset{\mathbf{c} \in \mathbb{R}^2, r \geq 0}{\arg\max} \left\{\prod_{i \in [n]} \mathbb{P}[\mathbf{x}_i \mid \mathbf{c}, r]\right\}\right] = \left[\underset{\mathbf{c} \in \mathbb{R}^2, r \geq 0}{\arg\min}\, r^2 \text{ s.t. } \|\mathbf{x}_i - \mathbf{c}\|_2^2 \leq r^2 \text{ for all } i \in [n]\right]$. **Your solution must be a proper distribution i.e.,** $\mathbb{P}[\mathbf{x} \mid \mathbf{c}, r] \geq 0$ 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. *Hint: area of a circle of radius r is πr^2.* **(4 + 6 = 10 marks)**

$$\mathbb{P}[\mathbf{x} \mid \mathbf{c}, r] = \begin{cases} \dfrac{1}{\pi r^2} & \|\mathbf{x} - \mathbf{c}\|_2 \leq r \\ 0 & \|\mathbf{x} - \mathbf{c}\|_2 > r \end{cases}$$

Using the above likelihood distribution expression yields the following likelihood value

$$\prod_{i \in [n]} \mathbb{P}[\mathbf{x}_i \mid \mathbf{c}, r] = \begin{cases} \left(\dfrac{1}{\pi r^2}\right)^n & \|\mathbf{x}_i - \mathbf{c}\|_2^2 \leq r^2 \text{ for all } i \in [n] \\ 0 & \exists i \text{ such that } \|\mathbf{x}_i - \mathbf{c}\|_2 > r \end{cases}$$

Thus, likelihood drops to 0 if any data point is outside the circle. Since we wish to maximize the likelihood, we are forced to ensure that $\|\mathbf{x}_i - \mathbf{c}\|_2 > r$ does not happen for any $i \in [n]$. This yields the following optimization problem for the MLE

$$\underset{\mathbf{c} \in \mathbb{R}^2, r \geq 0}{\arg\max} \left(\frac{1}{\pi r^2}\right)^n \text{ s.t. } \|\mathbf{x}_i - \mathbf{c}\|_2^2 \leq r^2 \text{ for all } i \in [n]$$

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

$$\underset{\mathbf{c} \in \mathbb{R}^2, r \geq 0}{\arg\min}\, r^2 \text{ s.t. } \|\mathbf{x}_i - \mathbf{c}\|_2^2 \leq r^2 \text{ for all } i \in [n]$$

Q4. (A one-class SVM?) For anomaly detection tasks, the "one-class" approach is often used. Given a set of data points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, the 1CSVM solves the following problem:

$$\min_{\mathbf{w} \in \mathbb{R}^d, \boldsymbol{\xi} \in \mathbb{R}^n, \rho \in \mathbb{R}} \left\{\frac{1}{2}\|\mathbf{w}\|_2^2 - \rho + \sum_{i \in [n]} \xi_i\right\} \text{ s.t. } \mathbf{w}^\top \mathbf{x}_i \geq \rho - \xi_i \text{ and } \xi_i \geq 0 \text{ for all } i \in [n]$$

1. Write down the Lagrangian for this optimization problem by introducing dual variables.
2. Write down the dual problem as a max-min problem (no need to simplify it at this stage).
3. Now simplify the dual problem (eliminate $\mathbf{w}, \boldsymbol{\xi}, \rho$). Show major steps. **(3 + 2 + 5 = 10 marks)**

Introducing dual variables $\alpha_i, \beta_i, i \in [n]$ for the first and second set of constraints respectively (styled as vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ for notational brevity) and using $\mathbf{1} \in \mathbb{R}^n$ to denote the all-ones vector and $X \in \mathbb{R}^{n \times d}$ to denote the feature matrix allows us to write the Lagrangian in a compact form.

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|_2^2 - \rho + \boldsymbol{\xi}^\top \mathbf{1} + \boldsymbol{\alpha}^\top(\rho \cdot \mathbf{1} - \boldsymbol{\xi} - X\mathbf{w}) - \boldsymbol{\beta}^\top \boldsymbol{\xi}$$

The dual problem is simply $\max_{\boldsymbol{\alpha}, \boldsymbol{\beta} \geq \mathbf{0}} \left\{ \min_{\mathbf{w} \in \mathbb{R}^d, \boldsymbol{\xi} \in \mathbb{R}^n, \rho \in \mathbb{R}} \{\mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta})\} \right\}$

To simplify the dual, we eliminate $\mathbf{w}, \boldsymbol{\xi}, \rho$ by using first-order optimality to get

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \Rightarrow \mathbf{w} = X^\top \boldsymbol{\alpha} \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}} = \mathbf{0} \Rightarrow \boldsymbol{\alpha} + \boldsymbol{\beta} = \mathbf{1} \qquad \frac{\partial \mathcal{L}}{\partial \rho} = \mathbf{0} \Rightarrow \boldsymbol{\alpha}^\top \mathbf{1} = 1$$

Putting these back into the dual gives us the following form of the dual with constraints.

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n} \left\{ -\frac{1}{2} \boldsymbol{\alpha}^\top X X^\top \boldsymbol{\alpha} \right\} \quad \text{s.t.} \quad \boldsymbol{\alpha}, \boldsymbol{\beta} \geq \mathbf{0} \quad \text{and} \quad \boldsymbol{\alpha} + \boldsymbol{\beta} = \mathbf{1} \quad \text{and} \quad \boldsymbol{\alpha}^\top \mathbf{1} = 1$$

We now eliminate $\boldsymbol{\beta}$ by setting $\boldsymbol{\beta} = \mathbf{1} - \boldsymbol{\alpha}$. Note that this introduces a new constraint $\boldsymbol{\alpha} \leq \mathbf{1}$ (i.e., $\alpha_i \leq 1$ for all $i \in [n]$) since $\boldsymbol{\beta} \geq \mathbf{0}$. This simplifies the dual further to

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2} \boldsymbol{\alpha}^\top X X^\top \boldsymbol{\alpha} \quad \text{s.t.} \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1} \quad \text{and} \quad \boldsymbol{\alpha}^\top \mathbf{1} = 1$$

Actually, the constraint $\boldsymbol{\alpha} \leq \mathbf{1}$ is vacuous since $\mathbf{0} \leq \boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^\top \mathbf{1} = 1$ together ensure $\boldsymbol{\alpha} \leq \mathbf{1}$. Thus, and even more simplified version of the dual is

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2} \boldsymbol{\alpha}^\top X X^\top \boldsymbol{\alpha} \quad \text{s.t.} \quad \boldsymbol{\alpha} \geq \mathbf{0} \quad \text{and} \quad \boldsymbol{\alpha}^\top \mathbf{1} = 1$$

Q5 (Kernelized Anomaly Detection?) Let's kernelize the 1CSVM. Suppose d is large and instead of receiving $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, we receive pairwise dot products of the features as an $n \times n$ matrix $G = [g_{ij}] \in \mathbb{R}^{n \times n}$ where $g_{ij} \overset{\text{def}}{=} \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ for all $i, j \in [n]$. Rewrite **the (simplified) dual that you derived in Q4** but using only the dot products g_{ij}. No derivations required – just rewrite the dual using the dot products. **Note**: **your rewritten dual must not use feature vectors $\mathbf{x}_i$ at all.** **(2 marks)**

Note that $XX^\top = G$. This allows us to rewrite the simplified dual in terms of just the dot products.

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2} \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} \quad \text{s.t.} \quad \mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{1} \quad \text{and} \quad \boldsymbol{\alpha}^\top \mathbf{1} = 1$$

or else the further simplified form

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2} \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} \quad \text{s.t.} \quad \boldsymbol{\alpha} \ge \mathbf{0} \quad \text{and} \quad \boldsymbol{\alpha}^\top \mathbf{1} = 1$$

Q6 (Delta Likelihood) Melbo has n data points $\{\mathbf{x}_i, y_i\}$ with $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}$. The likelihood of a model $\mathbf{w} \in \mathbb{R}^d$ w.r.t. data point i is $s_i \overset{\text{def}}{=} 1/(1 + \exp(-y_i \cdot \mathbf{w}^\top \mathbf{x}_i))$ and w.r.t. the entire data is $\mathcal{L}(\mathbf{w}) \overset{\text{def}}{=} \prod_{i \in [n]} s_i$. Notice that if the label of the j-th point is flipped (for any single $j \in [n]$), then the likelihood of the same model $\mathbf{w}$ changes to $\tilde{\mathcal{L}}_j(\mathbf{w}) \overset{\text{def}}{=} 1/\left(1 + \exp\left(y_j \cdot \mathbf{w}^\top \mathbf{x}_j\right)\right) \cdot \left(\prod_{i \neq j} s_i\right)$.

i. Given a **fixed** model $\mathbf{w}$, $j \in [n]$, give an expression for $\Delta_j(\mathbf{w}) \overset{\text{def}}{=} \tilde{\mathcal{L}}_j(\mathbf{w}) / \mathcal{L}(\mathbf{w})$, i.e., the factor by which likelihood of $\mathbf{w}$ changes if j-th label is flipped. Give brief derivation.
ii. If $n = 5$ and $s_1 = 0.1, s_2 = 0.3, s_3 = 0.9, s_4 = 0.6, s_5 = 0.2$, find the point $j^* \in [5]$ for which $\Delta_{j^*}(\mathbf{w})$ is the largest and value of $\Delta_{j^*}(\mathbf{w})$. Give brief justification.
iii. If $n = 5$ and $s_1 = 0.4, s_2 = 0.6, s_3 = 0.2, s_4 = 0.7, s_5 = 0.8$, find $k^* \in [5]$ for which $\Delta_{k^*}(\mathbf{w})$ is the smallest and value of $\Delta_{k^*}(\mathbf{w})$. Give brief justification. **(2 + 3 + 3 = 8 marks)**

$\Delta_j(\mathbf{w}) = \exp\left(-y_j \cdot \mathbf{w}^\top \mathbf{x}_j\right)$ or equivalently, $\frac{(1+\exp(-y_j \cdot \mathbf{w}^\top \mathbf{x}_j))}{(1+\exp(y_j \cdot \mathbf{w}^\top \mathbf{x}_j))}$

$j^* = 1$ $\quad$ $\Delta_{j^*}(\mathbf{w}) = 9$ $\quad$ $k^* = 5$ $\quad$ $\Delta_{k^*}(\mathbf{w}) = 0.25$

Give brief derivation for part i and justification for parts ii and iii below.

We have $\Delta_j(\mathbf{w}) = \left(1 + \exp\left(-y_j \cdot \mathbf{w}^\top \mathbf{x}_j\right)\right) / \left(1 + \exp\left(y_j \cdot \mathbf{w}^\top \mathbf{x}_j\right)\right) = \exp\left(-y_j \cdot \mathbf{w}^\top \mathbf{x}_j\right) = \frac{1}{s_i} - 1$.

Thus, $\Delta_j(\mathbf{w})$ is the largest when s_j is the smallest giving us $j^* = 1$ and $\Delta_{j^*}(\mathbf{w}) = \frac{1}{0.1} - 1 = 9$.

Also, $\Delta_k(\mathbf{w})$ is the smallest when s_j is the largest giving us $k^* = 5$ and $\Delta_{k^*}(\mathbf{w}) = \frac{1}{0.8} - 1 = 0.25$.

Note that this makes sense since in part ii, point 1 is indeed the worst classified point (misclassified with a large margin) and thus, flipping y_1 will increase the likelihood the most.

Similarly in part iii, point 5 is the best classified point (correctly classified with a large margin) and thus, flipping y_5 will decrease the likelihood the most.