| CS 771A: Intro to Machine Learning, IIT Kanpur | | | | **Endsem Exam** (14 July 2023) |
|---|---|---|---|---|
| Name | | | | **40 marks** |
| Roll No | | Dept. | |  |

**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**) $\mathbb{P}[\hat{y} = y]$ | |
| Precision (**PRE**) $\mathbb{P}[y = 1 \mid \hat{y} = 1]$ | |
| Recall (**REC**) $\mathbb{P}[\hat{y} = 1 \mid y = 1]$ | |
| False discovery rate (**FDR**) $\mathbb{P}[y = -1 \mid \hat{y} = 1]$ | |
| False omission rate (**FOR**) $\mathbb{P}[y = 1 \mid \hat{y} = -1]$ | |

**Q2. (Kernel Smash)** Melbi has created two Mercer kernels $K_1, K_2 : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ with the feature map for the kernel $K_i$ being $\phi_i : \mathbb{R} \rightarrow \mathbb{R}^2$. 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 $\phi_3 : \mathbb{R} \rightarrow \mathbb{R}^7$ for the kernel $K_3$. Write your answer only in the pace given below. No derivation needed. **Note that $\phi_3$ must not use more than 7 dimensions. If your solution does not require 7 dimensions leave the rest of the dimensions blank.** **(5 marks)**

$\phi_3(x) =$

$(\square, \square, \square, \square, \square, \square, \square)$

**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:

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

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 $\pi r^2$.* **(4 + 6 = 10 marks)**

Write down the density function of your likelihood distribution here.

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

**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)**

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

Derive and simplify the dual for this problem. Show major calculations steps.

**Q5 (Kernelized Anomaly Detection?)** Let's kernelize the 1CSVM. Suppose $d$ is large and instead of receiving $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, we receive pairwise dot products of the features as an $n \times n$ matrix $G = [g_{ij}] \in \mathbb{R}^{n \times n}$ where $g_{ij} \stackrel{\text{def}}{=} \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ 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)**

**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 $s_i \stackrel{\text{def}}{=} 1/(1 + \exp(-y_i \cdot \mathbf{w}^\top \mathbf{x}_i))$ and w.r.t. the entire data is $\mathcal{L}(\mathbf{w}) \stackrel{\text{def}}{=} \prod_{i \in [n]} s_i$. 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 $\tilde{\mathcal{L}}_j(\mathbf{w}) \stackrel{\text{def}}{=} 1/\left(1 + \exp\left(y_j \cdot \mathbf{w}^\top \mathbf{x}_j\right)\right) \cdot \left(\prod_{i \neq j} s_i\right)$.

i. Given a **fixed** model $\mathbf{w}$, $j \in [n]$, give an expression for $\Delta_j(\mathbf{w}) \stackrel{\text{def}}{=} \tilde{\mathcal{L}}_j(\mathbf{w}) / \mathcal{L}(\mathbf{w})$, 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)**

$\Delta_j(\mathbf{w}) =$

$j^* =$ ____ $\Delta_{j^*}(\mathbf{w}) =$ ________ $k^* =$ ____ $\Delta_{k^*}(\mathbf{w}) =$ ________

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