CS 771A: Intro to Machine Learning, IIT Kanpur				**Midsem Exam** (17 Sep 2025)
Name	MELBO			**40 marks** Page **1** of **4**
Roll No	250007	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 (M-SVM Dual) In a previous discussion, we saw that if Mahalanobis distances are used to derive the SVM objective instead of Euclidean distances, then the SVM changes its form as shown below (bias is hidden inside the model vector). Here $y^i \in \{-1,1\}$, $\mathbf{w}, \mathbf{x}^i \in \mathbb{R}^D$, $A \in \mathbb{R}^{D\times D}$ is a symmetric, invertible, positive definite matrix i.e., $\mathbf{x}^\top A\mathbf{x} > 0$ and $\mathbf{x}^\top A^{-1}\mathbf{x} > 0$ for non-zero vectors $\mathbf{x} \in \mathbb{R}^D$. Derive the Lagrangian dual and show all steps as directed below. **(2 + 2 + 2 = 6 marks)**

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^\top A^{-1}\mathbf{w} \quad \text{s.t.} \quad y^i \cdot \left(\mathbf{w}^\top \mathbf{x}^i\right) \geq 1 \quad \forall i \in [n]$$

Write down the Lagrangian by introducing dual variables. No derivation needed.

Let $\boldsymbol{\alpha} \in \mathbb{R}^n_+$ be a vector of non-negative Lagrange variables, $Y = \text{diag}(y_1, \ldots, y_n)$ and feature matrix $X \in \mathbb{R}^{n\times d}$. This gives us $\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^\top A^{-1}\mathbf{w} + \boldsymbol{\alpha}^\top(\mathbf{1} - YX\mathbf{w})$

Give an expression of the model $\mathbf{w}$ in terms of the dual variable. No derivation needed.

Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0}$ gives us $A^{-1}\mathbf{w} - X^\top Y\boldsymbol{\alpha} = \mathbf{0}$ i.e., we get $\mathbf{w} = AX^\top Y\boldsymbol{\alpha}$

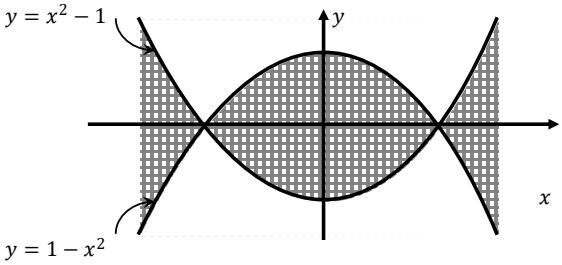
Simplify the dual problem (eliminate $\mathbf{w}$) – show major steps. You may use $\frac{\partial \mathbf{x}^\top B\mathbf{x}}{\partial \mathbf{x}} = (B + B^\top)\mathbf{x}$

The dual problem is $\max_{\boldsymbol{\alpha}} \left\{\min_{\mathbf{w}} \{\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha})\}\right\}$. Putting the above value of w in the Lagrangian yields the following dual problem:

$$\max_{\boldsymbol{\alpha}} \left\{\boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^\top YXAX^\top Y\boldsymbol{\alpha}\right\} = \max_{\boldsymbol{\alpha}} \left\{\sum_{i\in[n]} \alpha^i - \frac{1}{2}\sum_{i,j\in[n]} \alpha^i\alpha^j y^i y^j \left(\mathbf{x}^i\right)^\top A\mathbf{x}^j\right\}$$

$$= \min_{\boldsymbol{\alpha}} \left\{\frac{1}{2}\boldsymbol{\alpha}^\top YXAX^\top Y\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \mathbf{1}\right\} = \min_{\boldsymbol{\alpha}} \left\{\frac{1}{2}\sum_{i,j\in[n]} \alpha^i\alpha^j y^i y^j \left(\mathbf{x}^i\right)^\top A\mathbf{x}^j - \sum_{i\in[n]} \alpha^i\right\}$$

Q2. (Candy Classifier) The figure shows a binary classification task. The bold lines are the curves $y = x^2 - 1$ and $y = 1 - x^2$ where $x \in \mathbb{R}$. Create a feature map $\phi: \mathbb{R}^2 \to \mathbb{R}^D$ for some integer $D > 0$ so that for any 2D vector $\mathbf{z} = (x, y) \in \mathbb{R}^2$, the value of $\text{sign}\left(\mathbf{1}^\top \phi(\mathbf{z})\right)$ is $+1$ if $\mathbf{z}$ is in the cross-hatched region and -1 if $\mathbf{z}$ is in the white region. The D-dimensional all-ones

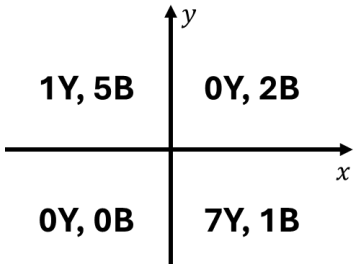


vector is denoted as $\mathbf{1} = (1,1,\ldots,1) \in \mathbb{R}^D$. Write down your feature map in the space below. To create your map, you may use common functions such as polynomials, absolute value, exponential etc i.e., a feature map such as $\phi(\mathbf{z}) = \left(x, y, \exp(x-y), y^2 - x^2, \sqrt{x}\right)$ would be valid (although it might not solve the problem). **No derivation needed.** **(4 marks)**

$\phi(x,y) = (+x^4, +1, -y^2, -2x^2)$

The shaded area is where both curves give opposite signs i.e. $(y - x^2 + 1)(y - 1 + x^2) \leq 0$

Q3. (Axis-aligned DT) Melbo wants to solve a binary classification problem with two labels **Y** and **B** and 2D features using a Decision Tree. We can only ask two kinds of split questions at any non-leaf node. Either we can ask whether $x > 0$ or not. Or else we can ask whether $y > 0$ or not. **Do not use $x < 0$ or $y < 0$ as questions as it will flip your tree.** No data point has either $x = 0$ or $y = 0$. The distribution of data points of the two labels is given on the right e.g. 1 Y and 5 B points have both $x < 0$ and $y > 0$, no point has both $x < 0$ and $y < 0$ etc. Note that the actual coordinates don't matter as we can only ask if a point lies above or below the x-axis, or to the left or right of the y-axis. The tree has 3 levels (one root, its two children and four leaves). Help Melbo construct this DT using entropy minimization. **No derivations needed**. You may use $\log_2 3 = 1.585, \log_2 5 = 2.322, \log_2 7 = 2.807$. **(2+2+2+10+4=20 marks)**



1. What is the entropy at the root note? Write the answer to 2 decimal places. **2 marks**

$H(\text{root}) = 1.00$

There are 8Y and 8B points i.e. proportions are $\left(\frac{1}{2}, \frac{1}{2}\right)$ yielding a unit entropy

2. If we split the root using a horizontal split i.e. ask $y > 0$, what is the entropy of this split? (2 decimal places). **Note**: we need the entropy of the split not entropy of any child. **2 marks**

$\mathrm{H}(y > 0) = 0.54$

This split yields children (1Y,7B), (7Y,1B) i.e. proportions of $\left(\frac{1}{8}, \frac{7}{8}\right), \left(\frac{7}{8}, \frac{1}{8}\right)$ with equal populations in both children giving a split entropy of 0.544 (answers giving correct entropy reduction will also get marks)

3. If we split the root using a vertical split i.e. ask $x > 0$, what is the entropy of this split? (2 decimal places). **Note**: we need the entropy of the split not entropy of any child. **2 marks**

$\mathrm{H}(x > 0) = 0.795 \approx 0.80 \text{ or } 0.79$

This split yields child proportions of $\left(\frac{7}{10}, \frac{3}{10}\right), \left(\frac{1}{6}, \frac{5}{6}\right)$ i.e. entropies $0.882, 0.65$. As child populations are in ratio $10:6$, this gives us a split entropy of 0.795 (answers giving correct entropy reduction will also get marks)

4. In the diagram below, indicate the split question that Melbo should use at the root to get maximum entropy reduction for the root split, as well as the split questions that Melbo should use at the left child and right child of the root. For the root, left child, right child and all 4 leaf nodes, write down, in the space provided in the diagram, how many **Y** and **B** points reached them according to your solution. Pay close attention to which child corresponds to the **Yes** answer to a split question and which child corresponds to the **No** answer to a split question. Mistakes may not get partial marks. **2 x 3 + 1 x 4 = 10 marks**

5. In the space below, write down the entropy after leaves have been created (correct to 2 decimal places). **Note**: we need the entropy of the entire set of 4 leaves and not the entropy of the leaves separately. **Show brief derivation**. **1 + 3 = 4 marks**

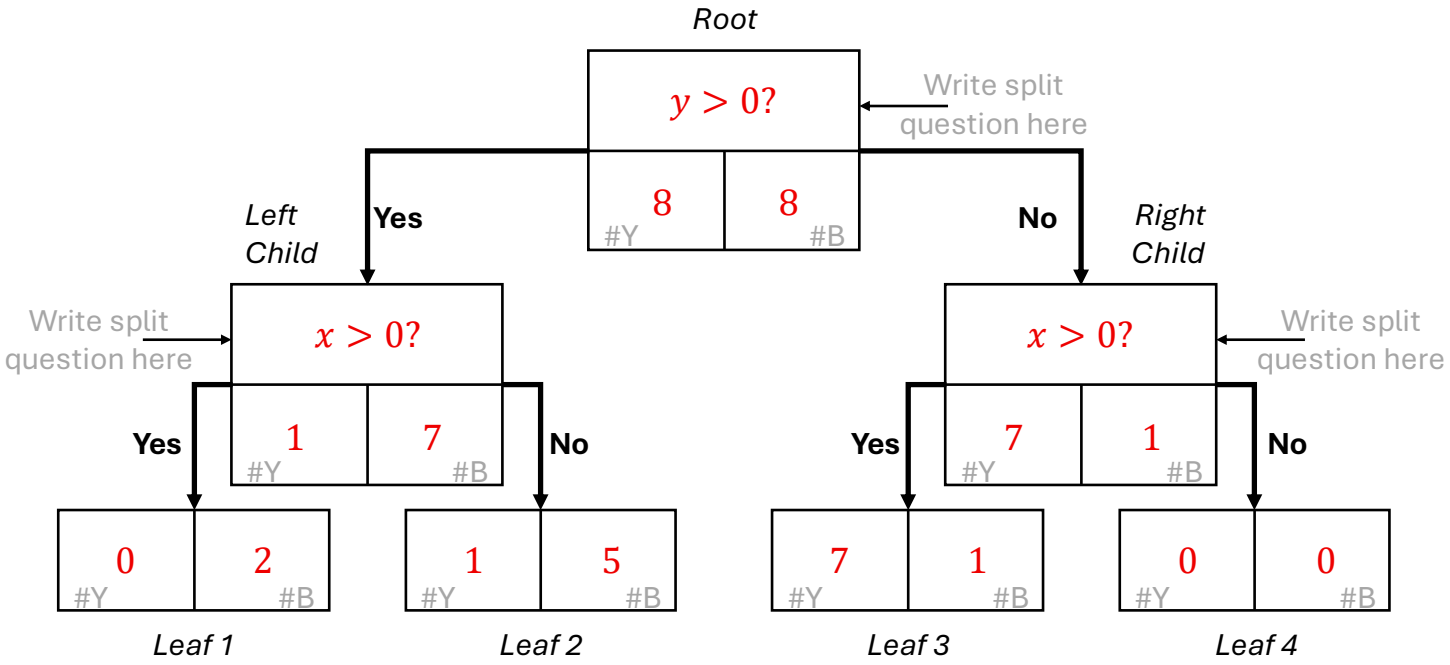
$\mathrm{H}(\text{leaves}) = 0.52 \text{ or } 0.51$

Give brief derivation here

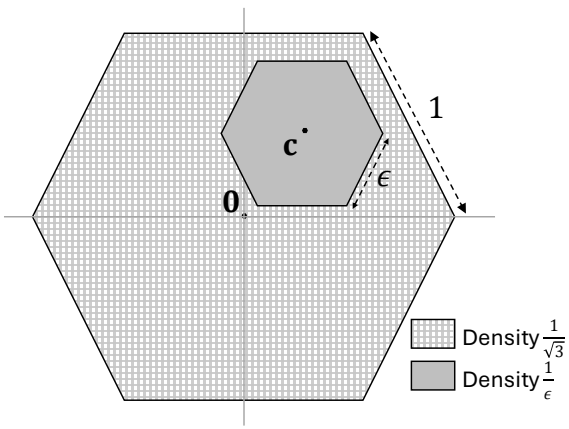
The label distribution in the leaves is $(0,1), \left(\frac{1}{6}, \frac{5}{6}\right), \left(\frac{7}{8}, \frac{1}{8}\right), (0,0)$ i.e. with entropy values $0, 0.65, 0.544, 0$ (since $0 \log 0 \to 0$). As the leaves have populations in ratio $2:6:8:0$, the entropy of the set of leaves is

$$\frac{2}{16} \cdot 0 + \frac{6}{16} \cdot 0.65 + \frac{8}{16} \cdot 0.544 + \frac{0}{16} \cdot 0 = 0.516$$

Note that part 4 and part 5 can be solved even without solving parts 2 and 3 first. It is visibly apparent, without doing entropy calculations, that the horizontal split yields children with higher purity so part 4 can be solved by inspection. Moreover, the entropy of the set of leaves will remain the same even if the suboptimal question $x > 0$ were to be asked at the root and the question $y > 0$ were to be asked at its children.



Q4 (Bestagon Distribution) Melbae has a distribution $\mathcal{D}$ with support over 2D vectors lying inside an equilateral hexagon centered at the origin with all side lengths $= 1$. $\mathcal{D}$ is described by two parameters $\mathbf{c} \in \mathbb{R}^2$ and $\epsilon \in [0,1]$. The density for $\mathcal{D}$ is $\frac{1}{\epsilon}$ in the *dense* equilateral hexagon of side length ϵ centered at $\mathbf{c}$. The density is $\frac{1}{\sqrt{3}}$ in the rest of the support. Assume $\mathbf{c}$ lies within an equilateral hexagon of side length $1 - \epsilon$ i.e., the dense region stays inside the support i.e. inside the bigger hexagon.



a. For which values of ϵ will $\mathcal{D}$ be a proper distribution? **Find them and show calculations**.
b. Find out the mean vector $\boldsymbol{\mu} \in \mathbb{R}^2$ of this distribution. **Show calculations**. **(4 + 6 = 10 marks)**

Hint: The mean of the uniform distribution over an equilateral hexagon is its centre. The area of an equilateral triangle of side s is $\frac{\sqrt{3}}{4}s^2$.

Find value(s) of ϵ for which $\mathcal{D}$ is a proper distribution.

The area of a regular hexagon of side length s is $\frac{3\sqrt{3}}{2}s^2$. As proper distributions are normalized, we need $\frac{1}{\epsilon}\cdot\frac{3\sqrt{3}}{2}\epsilon^2 + \frac{1}{\sqrt{3}}\cdot\left(\frac{3\sqrt{3}}{2} - \frac{3\sqrt{3}}{2}\epsilon^2\right) = 1$ i.e. $3\epsilon^2 - 3\sqrt{3}\epsilon - 1 = 0$. The solutions of this quadratic are $\epsilon = \frac{3\sqrt{3}\pm\sqrt{39}}{6}$. None of these values are in the valid range $[0,1]$ hence there are no valid solutions for ϵ for the given values of density.

Solutions that give the above or similar calculations will be given full marks for this part, irrespective of whether they state that there is no solution, or whether they choose the non-negative solution.

Find out the mean vector of the distribution $\mathcal{D}$.

Since the distribution is not proper for $\epsilon \in [0,1]$, the mean calculation does not make sense. Marks for part 2 would be prorated from marks received in part 1, irrespective of whether part 2 is attempted or not. If part 2 is attempted without attempting part 1, partial marks would be assigned. If this had been a proper distribution, then the mean would have been calculated as follows:

Let $\mathcal{H}$ be large hexagon with density $\mathcal{D}(\mathbf{x}) = P$ and $\mathcal{I}$ be the small hexagon with density $\mathcal{D}(\mathbf{x}) = p$. Note that $p \gg P$. We are not using actual values of $P = \frac{1}{\sqrt{3}}, p = \frac{1}{\epsilon}$ as those values do not yield a distribution anyway. We have $\boldsymbol{\mu} = \int_{\mathcal{H}} \mathbf{x}\cdot\mathcal{D}(\mathbf{x})\,d\mathbf{x} = \underbrace{\int_{\mathcal{I}} \mathbf{x}\cdot\mathcal{D}(\mathbf{x})\,d\mathbf{x}}_{(A)} + \underbrace{\int_{\mathcal{H}\setminus\mathcal{I}} \mathbf{x}\cdot\mathcal{D}(\mathbf{x})\,d\mathbf{x}}_{(B)}$.

$(A) = p\int_{\mathcal{I}} \mathbf{x}\,d\mathbf{x}$. Now $\int_{\mathcal{I}} \mathbf{x}\,d\mathbf{x} = \frac{3\sqrt{3}}{2}\epsilon^2\cdot\int_{\mathcal{I}} \mathbf{x}\cdot\mathcal{U}(\mathbf{x})\,d\mathbf{x}$ where $\mathcal{U}(\mathbf{x}) = \frac{2}{3\sqrt{3}\epsilon^2}$ is the (conditional) uniform distribution inside the dense hexagon. As the mean of a uniform distribution over a hexagon is its centre, we have $\int_{\mathcal{I}} \mathbf{x}\cdot\mathcal{U}(\mathbf{x})\,d\mathbf{x} = \mathbf{c}$ which gives us $(A) = p\cdot\frac{3\sqrt{3}}{2}\epsilon^2\cdot\mathbf{c}$

$(B) = P\int_{\mathcal{H}\setminus\mathcal{I}}\mathbf{x}\,d\mathbf{x} = P\left(\underbrace{\int_{\mathcal{H}}\mathbf{x}\,d\mathbf{x}}_{(C)} - \underbrace{\int_{\mathcal{I}}\mathbf{x}\,d\mathbf{x}}_{(D)}\right)$. Using the same argument as above, we get $(C) = \frac{3\sqrt{3}}{2}1^2\cdot\mathbf{0}$ and $(D) = \frac{3\sqrt{3}}{2}\epsilon^2\cdot\mathbf{c}$ which gives us $(B) = -P\cdot\frac{3\sqrt{3}}{2}\epsilon^2\cdot\mathbf{c}$ giving $\boldsymbol{\mu} = (p-P)\frac{3\sqrt{3}}{2}\epsilon^2\cdot\mathbf{c}$.

Even though a solution does not exist to the mean problem, putting in values of p, P yields

$$\boldsymbol{\mu} = \left(\epsilon\sqrt{3} - \epsilon^2\right)\frac{3}{2}\cdot\mathbf{c} = -\frac{1}{2}\cdot\mathbf{c}$$

since we were supposed to have $3\epsilon^2 - 3\sqrt{3}\epsilon - 1 = 0$ i.e. $\epsilon^2 - \sqrt{3}\epsilon = \frac{1}{3}$. However, this result is absurd since the mean is pointing away from the supposedly "high"-density region.