CS 771A: Intro to Machine Learning, IIT Kanpur				**Midsem Exam** (18 Jun 2023)
Name	MELBO			**40 marks** Page **1** of **4**
Roll No	230007	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 (Optimal DT) Melbo has a multiclass problem with three classes $+, \times, \square$. There are 16 datapoints in total, each with a 2D feature vector (x, y). x, y can take value 0 or 1. The table below describes each data point. All 16 points are at the root of a decision tree. Melbo wishes to learn a decision stump based on the entropy reduction principle to split this node into two children. Help Melbo finish this task. *Hint: take logs to base 2 so no need for calculator* 😊. **(8 x 0.5 = 4 marks)**

SNo	Class	(x,y)	SNo	Class	(x,y)	SNo	Class	(x,y)	SNo	Class	(x,y)
1	$+$	$(0,1)$	5	$+$	$(0,1)$	9	$\times$	$(1,0)$	13	$\square$	$(1,0)$
2	$+$	$(1,1)$	6	$+$	$(0,1)$	10	$\times$	$(1,0)$	14	$\square$	$(0,0)$
3	$+$	$(0,1)$	7	$+$	$(1,1)$	11	$\times$	$(0,0)$	15	$\square$	$(1,0)$
4	$+$	$(1,1)$	8	$+$	$(1,1)$	12	$\times$	$(0,0)$	16	$\square$	$(0,0)$

What is the entropy of the root node?

Class proportions are $(1/2, 1/4, 1/8)$

$H_{\text{root}} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4}$

$= 1.5$

What is the entropy of the two child nodes (give answers for the two nodes separately) if the split is done using the x feature ($x = 0$ becomes left child, $x = 1$ becomes right child)?

Class proportions remain the same i.e., $H_{\text{left}} = 1.5$

Class proportions remain the same i.e., $H_{\text{right}} = 1.5$

What is the reduction in entropy (i.e., $H_{\text{root}} - H_{\text{children}}$) if the split is done using the x feature as described above?

$H_{root} - \frac{1}{2}(H_{\text{left}} + H_{\text{right}}) = 1.5 - \frac{1}{2}(1.5 + 1.5)$

$= 0$

What is the entropy of the two child nodes (give answers for the two nodes separately) if the split is done using the y feature ($y = 0$ becomes left child, $y = 1$ becomes right child)?

Class proportions are $(0, 1/2, 1/2)$ i.e., $H_{\text{left}} = 1$

Class proportions are $(1,0,0)$ i.e., $H_{\text{right}} = 0$

What is the reduction in entropy (i.e., $H_{\text{root}} - H_{\text{children}}$) if the split is done using the y feature as described above?

$H_{root} - \frac{1}{2}(H_{\text{left}} + H_{\text{right}}) = 1.5 - \frac{1}{2}(1 + 0)$

$= 1$

To get the most entropy reduction, should we split using x feature or y feature?

We should split using the y feature

Q2. Write **T** or **F** for True/False **in the box**. Also, **give justification**. **(4 x (1+3) = 16 marks)**

1	Recall that $\|\mathbf{v}\|_0$ is the number of non-zero coordinates of the vector $\mathbf{v}$. Then for any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$, we always have $\|\mathbf{a}+\mathbf{b}\|_0 \leq \|\mathbf{a}\|_0 + \|\mathbf{b}\|_0$. If true, give a brief proof, else give a counterexample of two 3D vectors that violate this inequality. Show brief calculations in either case.	T

Note that if $a_i = 0 = b_i$ for some $i \in [d]$, then $a_i + b_i = 0$. This means that if $a_i + b_i \neq 0$, then either a_i or b_i must be non-zero. Another way of saying this is to write

$$\mathbb{I}\{a_i + b_i \neq 0\} \leq \mathbb{I}\{a_i \neq 0\} + \mathbb{I}\{b_i \neq 0\}$$

where $\mathbb{I}\{A\} = \begin{cases} 0 & \text{if } A \text{ is false} \\ 1 & \text{if } A \text{ is true} \end{cases}$ is the *indicator function* for some statement A. Taking sums over $i \in [d]$ gives us $\sum_{i\in[d]}\{a_i + b_i \neq 0\} \leq \sum_{i\in[d]} \mathbb{I}\{a_i \neq 0\} + \sum_{i\in[d]} \mathbb{I}\{b_i \neq 0\}$. However, for any vector $\mathbf{v}$, we have $\|\mathbf{v}\|_0 = \sum_{i\in[d]} \mathbb{I}\{v_i \neq 0\}$. This tells us $\|\mathbf{a} + \mathbf{b}\|_0 \leq \|\mathbf{a}\|_0 + \|\mathbf{b}\|_0$.

2	The function $N(p) \stackrel{\text{def}}{=} p \ln p$ is convex over the interval $p \in (0, \infty)$. If true, give a proof via chord definition or tangent definition or second derivative definition of convex functions. If false, give a counter example that violates any one definition.	T

To use the tangent definition, we have to show that for any $p, q \in (0, \infty)$, we always have

$$N(q) \geq N(p) + N'(p)(q - p)$$

Since $N'(p) = (1 + \ln p)$, the above requirement reduces to showing

$$q \ln q - q > q \ln p - p$$

Consider the function $f(x) \stackrel{\text{def}}{=} q \ln x - x$. We have $f'(x) = q/x - 1$ and $f''(x) = -q/x^2$. Thus, $f'(q) = 0$ and $f''(q) < 0$ i.e., q is the global maxima for $f(x)$. This shows that $f(q) \geq f(p)$ or in other words, $q \ln q - q > q \ln p - p$, which proves that the function $N(\cdot)$ is convex.

3	The optimum for $\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w} - \mathbf{w}_0\|_2^2 + \|X\mathbf{w} - \mathbf{y}\|_2^2$ is always achieved at $\mathbf{w}_0$. Justify by deriving the optimum. $X \in \mathbb{R}^{n\times d}, \mathbf{y} \in \mathbb{R}^n, \mathbf{w}_0 \in \mathbb{R}^d$ are all constants.	F

As this is an unconstrained problem with a differentiable objective, first order optimality tells us that the gradient must vanish at the optimum. This means $(\mathbf{w} - \mathbf{w}_0) + X^\top(X\mathbf{w} - \mathbf{y}) = \mathbf{0}$ i.e., $\mathbf{w} = (X^\top X + I)^{-1}(\mathbf{w}_0 + X^\top \mathbf{y})$. This means that the optimum is not always $\mathbf{w}_0$.

4	For any two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$, we always have $\mathbf{u}^\top\mathbf{v} \leq (\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2)/2$. If true, give a brief proof, else give a counter example and calculations with two 3D vectors.	T

By Cauchy-Schwartz inequality, we have $\mathbf{u}^\top\mathbf{v} \leq \|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2$. However, we always have

$$\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2 \leq (\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2)/2$$

since $(\|\mathbf{u}\|_2 - \|\mathbf{v}\|_2)^2 \geq 0$. This completes the proof.

Q3 (Absolute Tilt) Consider the optimization problem $\min_{x\in\mathbb{R}} f(x)$ with objective $f:\mathbb{R}\to\mathbb{R}$ defined as $f(x) \stackrel{\text{def}}{=} |x| + a\cdot x$ where $a\in\mathbb{R}$ is a constant (maybe pos/neg/zero). Find the point x^* at which the optimum is achieved and $f(x^*)$. **Note**: x^* and $f(x^*)$ depend on a. Both $x^*, f(x^*)$ can be ∞ or $-\infty$ for certain cases. You must tell us for each possible case, where is the optimum achieved i.e., x^* and what is $f(x^*)$. E.g., you might say that case 1 is $a<1$, in which case we get $f(x^*)=1$ at $x^*=0.5$, and case 2 is $a\geq 1$, in which case we get $f(x^*)=-1$ at $x^*=\infty$. You may use at most 3 cases to describe your solution. If you don't need those many cases, leave cases blank. Give brief derivations. *Hint: you should not have to derive the dual to solve this problem.* **(8 marks)**

Case No.	Case Condition (write condition such as $a<1$ or $a=1$ etc).	Point x^* where opt. is reached for this case	Optimal objective value $f(x^*)$ for this case
1	$\lvert a\rvert \leq 1$	0	0
2	$\lvert a\rvert > 1$	Either $+\infty$ or $-\infty$	$-\infty$
3			

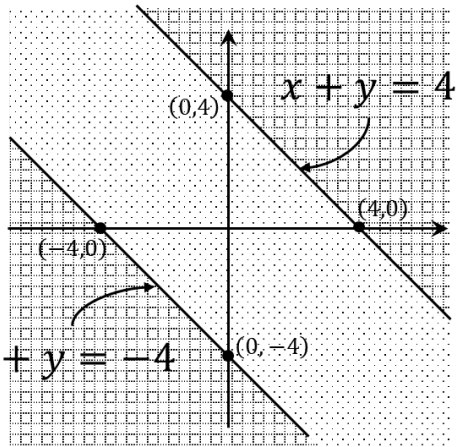
Give brief derivation below.

Note that the first term $|x|$ takes only positive values whereas the second term $a\cdot x$ can take negative values and reduce the objective value.

Notice that if $|a|\leq 1$, the first term i.e., $|x|$ dominates since $a\cdot x \leq |a\cdot x| = |a|\cdot|x| \leq |x|$ i.e., $|x| + a\cdot x \geq |x| - |x| \geq 0$. In this case, the smallest objective value is 0 which is indeed achieved at $x^*=0$.

However, if $|a|>1$, then the term $a\cdot x$ dominates and we can push objective value to $-\infty$. To see this, consider $x=-M\cdot\text{sign}(a)$ for some real $M\geq 0$. We have $f(x)=M-|a|\cdot M<0$ since $|a|>1$. Taking $M\to+\infty$ tells us that $\lim_{M\to\infty} f\big(-M\cdot\text{sign}(a)\big)=-\infty$ which completes the argument.

Q4. (Parallel Classifier) Create a feature map $\phi:\mathbb{R}^2\to\mathbb{R}^D$ for some $D>0$ so that for any $\mathbf{z}=(x,y)\in\mathbb{R}^2$, $\text{sign}\big(\mathbf{1}^\top\phi(\mathbf{z})\big)$ takes value $+1$ if $\mathbf{z}$ is in the dark cross-hatched region and -1 if $\mathbf{z}$ is in the light dotted region (see fig). E.g., $(0,0)$ is labelled -1 while the points $(2,5)$ and $(-6,1)$ are both labelled $+1$. The lines in the figure are $x+y=4$ and $x+y=-4$. We don't care what values are taken on points lying on these two lines (as these are the decision boundaries). $\mathbf{1}=(1,1,\ldots,1)\in\mathbb{R}^D$ is the all-ones vector. No need for derivation – just give the final map below. **(5 marks)**



$\phi(x, y) = (x^2, 2xy, y^2, -16)$

Intuition: The cross-hatched area is where $x + y \geq 4$ or $x + y \leq -4$ i.e., where $(x + y)^2 \geq 16$

Q5 (CM to the rescue) Consider the following problem where $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$ are constants and $\lambda > 0$. Design a coordinate minimization algorithm (choose coordinates cyclically) to solve the primal. Give brief calculations on how you will create a simplified unidimensional problem for a chosen coordinate $i \in [d]$ and then show how to get the optimal value of x_i. **(7 marks)**

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^d} \quad & \frac{1}{2}\|\mathbf{x}\|_2^2 + \mathbf{a}^\top \mathbf{x} \\ \text{s. t.} \quad & \mathbf{b}^\top \mathbf{x} \leq \lambda \\ & \mathbf{c}^\top \mathbf{x} \leq \lambda \end{aligned}$$

Suppose cyclic coordinate choice has presented a coordinate $i \in [d]$. We extract portions of the optimization problem that depend on x_i (and treat $x_j, j \neq i$ as constants) to get the following simplified 1D optimization problem (where $m_i \stackrel{\text{def}}{=} \lambda - \sum_{j \neq i} b_j x_j$ and $n_i \stackrel{\text{def}}{=} \lambda - \sum_{j \neq i} c_j x_j$).

$$\begin{aligned} \min_{x_i \in \mathbb{R}} \quad & \frac{1}{2}x_i^2 + a_i x_i \\ \text{s. t.} \quad & b_i x_i \leq m_i \\ & c_i x_i \leq n_i \end{aligned} \quad \Rightarrow \quad \begin{aligned} \min_{x_i \in \mathbb{R}} \quad & \frac{1}{2}x_i^2 + a_i x_i \\ \text{s. t.} \quad & x_i \in [l_i, u_i] \end{aligned}$$

Now we clean up the constraints to get a single box constraint as shown below

Case	$b_i > 0$ $c_i > 0$	$b_i > 0$ $c_i < 0$	$b_i < 0$ $c_i > 0$	$b_i < 0$ $c_i < 0$
l_i	$-\infty$	$\frac{n_i}{c_i}$	$\frac{m_i}{b_i}$	$\max\left\{\frac{m_i}{b_i}, \frac{n_i}{c_i}\right\}$
u_i	$\min\left\{\frac{m_i}{b_i}, \frac{n_i}{c_i}\right\}$	$\frac{m_i}{b_i}$	$\frac{n_i}{c_i}$	$+\infty$

If either b_i or c_i is 0, that constraint is ignored since it no longer involves x_i. Having converted the pair of constraints into a single box constraint, we can now apply the QUIN trick to solve this problem and obtain the optimal value x_i^* in two simple steps:

1. Find the unconstrained minimum for $\frac{1}{2}x_i^2 + a_i x_i$ which turns out to be $z_i \stackrel{\text{def}}{=} -a_i$
2. If $z_i \in [l_i, u_i]$, then $x_i^* = z_i$ else if $z_i < l_i$, $x_i^* = l_i$ else $x_i^* = u_i$