CS 771A: Intro to Machine Learning, IIT Kanpur				**Midsem Exam** (15 Jun 2024)
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. Write **T** or **F** for True/False **in the box**. Also, **give justification**. **(4 x (1+3) = 16 marks)**

1	All stationary points of the function $f(x) \stackrel{\text{def}}{=} x^3 - x^5$ are either local/global minima or local/global maxima. Justify your answer using first and second derivative tests.	F

$f'(x) = 3x^2 - 5x^4 = x^2(3 - 5x^2)$ and $f''(x) = 6x - 20x^3 = 2x(3 - 10x^2)$. The stationary points of f are $x = 0, \pm\sqrt{\frac{3}{5}}$. At the stationary points $x = \pm\sqrt{\frac{3}{5}}$, $f''(x) = \mp 6\sqrt{\frac{3}{5}}$ indicating a local maximum/minimum. However, $f''(x) = 0$ at $x = 0$ and moreover $f''(-\epsilon) < 0$ and $f''(\epsilon) > 0$ for small $\epsilon \to 0^+$ showing that the function transitions from being concave to being convex around 0 that indicates a saddle point.

2	Let $f: \mathbb{R} \to \mathbb{R}$ be a convex differentiable function. Let $g \stackrel{\text{def}}{=} -f$ i.e., $g(x) = -f(x)$ for all $x \in \mathbb{R}$. Then g can never be convex. Give either a proof or a counter example.	F

Linear/affine functions (that includes constant functions) are both convex and concave and are also closed under negation (the negative of an affine function is also affine). Btw, the difference between affine and linear functions is that the former can have a non-zero constant/bias term whereas the latter cannot. Thus, an affine function and its negation are both convex. More formally, suppose $f(x) = ax + b$ is an affine function. Then it is convex as $f'(x) = a$ and $f(y) = f(x) + a(y - x)$ (convexity requires $f(y) \geq f(x) + a(y - x)$). In this case, $g(x) = -ax - b$ which is also convex since $g(y) = g(x) + (-a)(y - x)$.

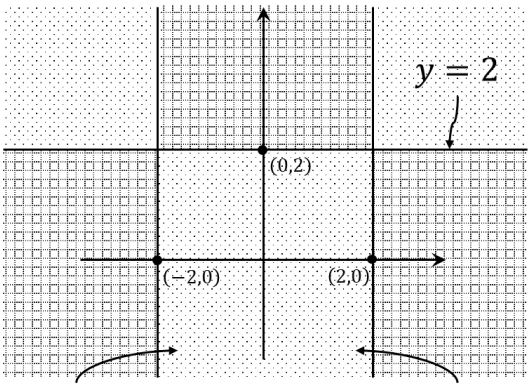
3	The optimum for $\underset{x \in \mathbb{R}}{\text{argmin}} \exp(\lvert x - x_0 \rvert) + (x - x_0)^2$ is always x_0. Justify by deriving the optimum. Note that $x_0 \in \mathbb{R}$ is a constant. (*Hint*: using calculus may be tricky)	T

$f(t) \stackrel{\text{def}}{=} \exp(t)$ and $g(t) \stackrel{\text{def}}{=} t^2$ are both increasing on the positive half of the real line as $f''(t) \geq 0, g''(t) \geq 0$ for $t \geq 0$. Thus, they achieve their minimum value at $t = 0$. Note that the objective is $f(\lvert x - x_0 \rvert) + g(\lvert x - x_0 \rvert)$. Since $\lvert x - x_0 \rvert \geq 0$ always, both the terms of the objective function will simultaneously achieve their minimum at $x = x_0$ which is where the objective will also achieve its minimum value.

4	The dot product of two **Boolean** vectors $\mathbf{u}, \mathbf{v} \in \{0,1\}^3$ cannot be zero unless one of them is the zero vector. If true, give a brief proof, else give a counter example.	F

Let $\mathbf{p} \stackrel{\text{def}}{=} (1,0,1)$ and $\mathbf{q} \stackrel{\text{def}}{=} (0,1,0)$. Neither vector is zero yet $\langle \mathbf{p}, \mathbf{q} \rangle = 0$

Q2. (Chessboard Classifier) Create a feature map $\phi: \mathbb{R}^2 \to \mathbb{R}^D$ for some $D > 0$ so that for any $\mathbf{z} = (x, y) \in \mathbb{R}^2$, $\text{sign}\left(\mathbf{1}^\top \phi(\mathbf{z})\right)$ 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., $(0,0), (3,3), (-3,3)$ are all labelled $+1$ while $(-3,0), (0,3), (3,0)$ are all labelled -1. The lines in the figure are $x = 2$, $x = -2$ and $y = 2$. We don't care what label is given to points lying on the three lines (these are the decision boundaries). $\mathbf{1} = (1,1,\ldots,1) \in \mathbb{R}^D$ is the all-ones vector. No need for derivation – give only the final map below.**(5 marks)**



$\phi(x, y) = [x^2 y, -4y, -2x^2, 8] \in \mathbb{R}^4$ or even $[x^2 y - 4y - 2x^2 + 8] \in \mathbb{R}$. Others solutions are possible too.

Explanation: Let $l_1 \stackrel{\text{def}}{=} \text{sign}(x+2), l_2 \stackrel{\text{def}}{=} \text{sign}(y-2), l_3 \stackrel{\text{def}}{=} \text{sign}(x-2)$ be pseudo labels assigned by the three lines. Then the final label is $\text{XOR}(l_1, l_2, l_3) = \text{sign}(x+2) \cdot \text{sign}(y-2) \cdot \text{sign}(x-2)$ $= \text{sign}\big((x+2) \cdot (y-2) \cdot (x-2)\big) = \text{sign}(x^2 y - 4y - 2x^2 + 8)$

Q3 (Optimal Checkerboard DT) Melbo has received data for the problem in Q2. There are 10 datapoints (given in the table), each with a 2D feature vector (x, y). All 10 points are at the root of a decision tree. Melbo wants to learn a decision stump based on the entropy reduction principle to split the root into two children. Only 3 decision stumps are allowed which ask the questions $(x \le -2?), (x \le 2?)$ and $(y \le 2?)$. **All logs are to base 2, assume $\log_2 3 = 1.58, \log_2 5 = 2.32$ Give your answers correct to at least 2 decimal places.** **(11 x 1 = 11 marks)**

S.	Class	(x,y)	S.	Class	(x,y)	S.	Class	(x,y)	S.	Class	(x,y)	S.	Class	(x,y)
1	−	$(-3,0)$	3	+	$(1,1)$	5	+	$(-1,1)$	7	−	$(1,5)$	9	−	$(-1,5)$
2	+	$(3,3)$	4	+	$(1,-1)$	6	+	$(-1,-1)$	8	−	$(1,3)$	10	−	$(-1,3)$

What is the entropy of the root node?

Class Frequency (CF) is $(5_+, 5_-)$ so $H_{\text{root}} = 1$

What is the entropy of the two child nodes (answers for the two nodes separately) if split is done using the question $x \le -2?$ i.e., $x \le -2$ becomes the left child, $x > -2$ becomes right child)?

CF is $(1_-, 0_+)$ $H_{\text{left}} = 0$

CF is $(4_-, 5_+)$ $H_{\text{right}} = 0.98$

What is the reduction in entropy (i.e., $H_{\text{root}} - H_{\text{children}}$) if the split is done using the question $x \le -2?$ as described above?

$$1 - \left(0 \cdot \frac{1}{10} + 0.98 \cdot \frac{9}{10}\right) = 0.118 \approx 0.12$$

What is the entropy of the two child nodes (answers for the two nodes separately) if split is done using the question $x \leq 2$? i.e., $x \leq 2$ becomes the left child, $x > 2$ becomes right child)?

What is the reduction in entropy (i.e., $H_{\text{root}} - H_{\text{children}}$) if the split is done using the question $x \leq 2$? as described above?

What is the entropy of the two child nodes (answers for the two nodes separately) if split is done using the question $y \leq 2$? i.e., $y \leq 2$ becomes the left child, $y > 2$ becomes right child)?

What is the reduction in entropy (i.e., $H_{\text{root}} - H_{\text{children}}$) if the split is done using the question $y \leq 2$? as described above?

To get the most entropy reduction, which decision stump should we use?

CF is $(0_-, 1_+)$ $H_{\text{left}} = 0$	CF is $(5_-, 4_+)$ $H_{\text{right}} = 0.98$
$1 - \left(0 \cdot \frac{1}{10} + 0.98 \cdot \frac{9}{10}\right)$ $= 0.118 \approx 0.12$	
CF is $(1_-, 4_+)$ $H_{\text{left}} = 0.72$	CF is $(4_-, 1_+)$ $H_{\text{right}} = 0.72$
$1 - \left(0.72 \cdot \frac{1}{2} + 0.72 \cdot \frac{1}{2}\right)$ $= 0.28$	
$y \leq 2?$	

Q4 (Tables are turned). A curious type of regularization is *Morozov regularization* which turns the loss function into a constraint (btw, SVMs & ridge regression use *Tikhonov regularization* instead). Consider the following regression problem where $X \in \mathbb{R}^{N \times d}$ gives us d-dimensional features for N data points and $\mathbf{y} \in \mathbb{R}^N$ gives the labels. Give 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 w_i. Assume $\|\mathbf{y}\|_2^2 \leq 1$ so that the constraint set is not empty (e.g., $\mathbf{w} = \mathbf{0}$ satisfies the constraint). Feel free to define shorthand notation to simplify your answer. **(8 marks)**

$$\min_{\mathbf{w} \in \mathbb{R}^d} \quad \frac{1}{2}\|\mathbf{w}\|_2^2$$
$$\text{s.t.} \quad \|X\mathbf{w} - \mathbf{y}\|_2^2 \leq 1$$

The question has an unintended loophole which lends itself an almost trivial answer. Submissions that exploit the loophole will not be penalized and if correct, will get full marks – same as submissions that do not exploit the loophole.

Loophole Solution: $\mathbf{w} = \mathbf{0}$ is clearly the optimal solution – it satisfies the constraint by assumption and $\|\mathbf{w}\|_2^2 \geq 0$ for all $\mathbf{w}$. If initialized at $\mathbf{w} = \mathbf{0}$, the CM solver will not make any updates.

The loophole solution has limitations – the assumption $\|\mathbf{y}\|_2^2 \leq 1$ is unrealistic for real-life applications. A more realistic way to use Morozov regularization (which I avoided to reduce clutter hence introducing the loophole) is the following ($\mu \in (0,1)$ is a hyperparameter).

$$\min_{\mathbf{w} \in \mathbb{R}^d} \quad \frac{1}{2}\|\mathbf{w}\|_2^2$$
$$\text{s.t.} \quad \|X\mathbf{w} - \mathbf{y}\|_2^2 \leq \mu \cdot \|\mathbf{y}\|_2^2$$

Non-loophole Solution: We define shorthand $C = [c_{ij}] \stackrel{\text{def}}{=} X^\top X \in \mathbb{R}^{d\times d}$ and $\mathbf{z} \stackrel{\text{def}}{=} -X^\top \mathbf{y} \in \mathbb{R}^d$ so that the constraint becomes $\mathbf{w}^\top C\mathbf{w} + 2\mathbf{z}^\top \mathbf{w} + \|\mathbf{y}\|_2^2 \le 1$. Note that C is a symmetric matrix. Also define $\eta_i \stackrel{\text{def}}{=} \|\mathbf{y}\|_2^2 - 1 + 2\sum_{j\neq i} z_j w_j + \sum_{j\neq i}\sum_{k\neq i} w_j w_k c_{jk}$ and $\delta_i \stackrel{\text{def}}{=} 2\left(z_i + \sum_{j\neq i} c_{ij} w_j\right)$. Consider a coordinate $i \in [d]$ and look at the optimization problem restricted to the chosen coordinate.

$$\begin{aligned} \min_{w_i\in\mathbb{R}} \quad & \frac{1}{2}w_i^2 \\ \text{s.t.} \quad & c_{ii}w_i^2 + 2w_i\delta_i + \eta_i \le 0 \end{aligned}$$

Case 1: $c_{ii} = 0$. In this case the constraint becomes $2w_i\delta_i + \eta_i \le 0$.

Case 1.1: $\delta_i = 0$. In this case the constraint is vacuous – assuming we started at a feasible point, we must have $\eta_i \le 0$ so we set $w_i = 0$ to minimize the (now unconstrained) objective.

Case 1.2: $\delta_i > 0$: In this case the constraint becomes $w_i \le \frac{\eta_i}{2\delta_i}$. Applying the QUIN trick, we get $w_i = 0$ if $0 \le \frac{\eta_i}{2\delta_i}$ else $w_i = \frac{\eta_i}{2\delta_i}$.

Case 1.3: $\delta_i < 0$: In this case the constraint becomes $w_i \ge \frac{\eta_i}{2\delta_i}$. Applying the QUIN trick, we get $w_i = 0$ if $0 \ge \frac{\eta_i}{2\delta_i}$ else $w_i = \frac{\eta_i}{2\delta_i}$.

Case 2: $c_{ii} > 0$. In this case we check the discriminant $D \stackrel{\text{def}}{=} 4\delta_i^2 - 4\eta_i c_{ii}$ of the quadratic equation $c_{ii}w_i^2 + 2w_i\delta_i + \eta_i = 0$.

Case 2.1: $D = 0$: In this case only a single value of $w_i = -\frac{\delta_i}{c_{ii}}$ satisfies the constraint which must be the optimal value as well.

Case 2.2: $D > 0$: In this case we get a box constraint $w_i \in \left[\frac{-2\delta_i - \sqrt{D}}{2c_{ii}}, \frac{-2\delta_i + \sqrt{D}}{2c_{ii}}\right]$ to which we can apply the QUIN trick. If $0 < \frac{-2\delta_i - \sqrt{D}}{2c_{ii}}$ then $w_i = \frac{-2\delta_i - \sqrt{D}}{2c_{ii}}$ else if $0 > \frac{-2\delta_i + \sqrt{D}}{2c_{ii}}$ then $w_i = \frac{-2\delta_i + \sqrt{D}}{2c_{ii}}$ else $w_i = 0$.

Case 2.3: $D < 0$: This cannot happen unless we initialized at an infeasible point that violates the constraint. If we are faced with $D < 0$ we may either throw an error and abort the solver or do a projection step of the entire vector $\mathbf{w}$ onto the constraint set $\|X\mathbf{w} - \mathbf{y}\|_2^2 \le 1$ (which will be expensive).

Case 3: $c_{ii} < 0$: This cannot happen as $c_{ii} = \|\mathbf{x}_i\|_2^2 \ge 0$.