ID5001W: Machine learning and its applications
Midsem Exam

Name : **Roll No** :

(a) Answer any 5 out of 6 questions below.

(b) You may use any result proved in class (not tutorials) without proof.

(c) All figures must be neatly drawn using a ruler.

(d) No striking.

(e) Submit the answers in order.

(f) Insert the pages corresponding to the questions from this pdf before the first page of the answer to that question.

(g) You may refer to material covered in class including class notes.

(h) No seeking help from others.

(i) Write, scan, convert to pdf, insert the question pages appropriately and submit.

(j) Deadline is 12 Jan, 2023, 09:00 PM.

(k) Submit on Moodle.

(l) If you cannot, access Moodle then email the pdf to `cs18d006@smail.iitm.ac.in` *and cc* `hariguru@cse.iitm.ac.in`.

(m) As a general hint, plots are a very useful tool, use them whenever you can.

(1) **(Bayes Classifier.)**

i. Consider the following cost matrices for a 3 class classification problem.

$$L_{\text{zo}} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad L_{\text{ordinal}} = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} \quad L_{\text{abstain}} = \begin{bmatrix} 0 & 1 & 1 & \frac{1}{2} \\ 1 & 0 & 1 & \frac{1}{2} \\ 1 & 1 & 0 & \frac{1}{2} \end{bmatrix}$$

where $(i.j)$th entry in the cost matrix corresponds to the loss of predicting j when the truth is i. The abstain loss has 4 possible predictions for this three class problem corresponding to the three classes, and an 'abstain' option. Derive the Bayes classifier for all three cost matrices. In other words, give a mapping which takes as input a 3-dimensional probability vector corresponding to the class conditional probabilities and outputs one of the 3 classes for the zero-one and ordinal losses, and outputs one of the 4 possibilities for the abstain loss. Denote the option of 'abstaining' by the symbol $\perp$.

ii. Consider the following distribution of (X, Y) over $\mathbb{R} \times \{1, 2, 3\}$. with $P(Y = 1) = P(Y = 2) = P(Y = 3) = \frac{1}{3}$.

$$f_X(x|Y=1) = \begin{cases} \frac{3}{14} & \text{if } x \in [0,2] \\ \frac{1}{14} & \text{if } x \in [2,10] \\ 0 & \text{otherwise} \end{cases}$$

$$f_X(x|Y=2) = \begin{cases} \frac{3}{14} & \text{if } x \in [4,6] \\ \frac{1}{14} & \text{if } x \in [0,4] \cup [6,10] \\ 0 & \text{otherwise} \end{cases}$$

$$f_X(x|Y=3) = \begin{cases} \frac{3}{14} & \text{if } x \in [8,10] \\ \frac{1}{14} & \text{if } x \in [0,8] \\ 0 & \text{otherwise} \end{cases}$$

where $f_X(x|Y = a)$ is the conditional density of X given that $Y = a$. Give the Bayes classifier for all three cost matrices for this distribution.

iii. Repeat the previous sub-problem for the below distribution with $P(Y = 1) = P(Y = 2) = P(Y = 3) = \frac{1}{3}$.

$$f_X(x|Y=1) = \begin{cases} \frac{x-1}{9} & \text{if } x \in [1,4] \\ \frac{7-x}{9} & \text{if } x \in [4,7] \\ 0 & \text{otherwise} \end{cases}$$

$$f_X(x|Y=2) = \begin{cases} \frac{x-2}{9} & \text{if } x \in [2,5] \\ \frac{8-x}{9} & \text{if } x \in [5,8] \\ 0 & \text{otherwise} \end{cases}$$

$$f_X(x|Y=3) = \begin{cases} \frac{x-3}{9} & \text{if } x \in [3,6] \\ \frac{9-x}{9} & \text{if } x \in [6,9] \\ 0 & \text{otherwise} \end{cases}$$

(2+2+2 points)

(2) **(Multiclass Logistic Regression.)**

i. Let $X|Y = i$ be distributed as the multivariate normal given by $\mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 I)$ for all $i \in [K]$. Let π_i be equal to $P(Y = i)$. What is the posterior probability $P(Y = i|X = \mathbf{x})$?

ii. Consider the three class, 1-dimensional dataset, with 6 data points. With feature given by x and class label given by y.

x	3	2	5	5	7	8
y	1	1	2	2	3	3

The multinomial logistic loss is given as :

$$L = \sum_{i=1}^{6} -\log\left(\left[\mathrm{SM}\left(w_1 x_i + b_1, w_2 x_i + b_2, w_3 x_i + b_3\right)\right]_{y_i}\right)$$

where SM is the softmax function from $\mathbb{R}^3 \rightarrow \mathbb{R}^3_+$ and the parameters are w_j, b_j for $j \in \{1, 2, 3\}$. Give a setting for the parameters so that $L < 0.1$. Argue that the loss can be made arbitrarily close to zero for some setting of the parameters.

iii. Consider the same dataset as above. The loss minimised in one-vs-all logistic regression is:

$$L = \sum_{i=1}^{6}\sum_{j=1}^{3} -\log\left(\sigma(y_{ij}(w_j x_i + b_j))\right)$$

where σ is the sigmoid function. $y_{ij} = +1$ if $y_i = j$ and -1 otherwise. Show that for any setting of $w_1, w_2, w_3, b_1, b_2, b_3$ the loss L is greater than $2\log(2)$.

iv. Repeat the two sub-problems above, with the 2-dimensional 4-class dataset with 8 points given below as well. Note that the parameters are $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4$ and b_1, b_2, b_3 and b_4, with $\mathbf{w}_j \in \mathbb{R}^2$ and $b_j \in \mathbb{R}$. The multinomial logistic and one-vs-all loss expressions also change appropriately.

x_1	1	2	3	4	3	4	7	7
x_2	1	0	4	3	6	6	2	3
y	1	1	2	2	3	3	4	4

(1+1+2+2 points)

(3) **(Kernel Regression)** Consider the following kernel regression problem. The data matrix containing 3 points with one dimension is given by $X^\top = [-1, 0, 2]$. The regression targets are given by $\mathbf{y}^\top = [1, 2, 0]$. Consider the feature vector regression problem given by the objective:

$$R(\mathbf{w}) = \sum_{i=1}^{3} (\mathbf{w}^\top \phi(x_i) - y_i)^2$$

where $\phi : \mathbb{R} \rightarrow \mathbb{R}^d$ is a feature vector corresponding to the kernel $k(u, v) = \sin(u)\sin(v) + \cos(u)\cos(v) + 1$.

i. Solve the kernel regression problem and give the solution $\alpha_1^*, \alpha_2^*, \alpha_3^*$. Does the problem have unique or multiple solutions?

ii. Use the above $\boldsymbol{\alpha}^*$ to make predictions at the 11 points ranging from $x = -5$ to $x = 5$ in steps of 1. Plot this as a curve.

iii. Give any feature mapping $\phi : \mathbb{R} \rightarrow \mathbb{R}^d$ such that $\phi(u)^\top \phi(v) = k(u, v)$

iv. Give the solution $\mathbf{w}^*$ to the feature vector regression problem assuming the feature function ϕ got above. **(2+2+1+1 points)**

(4) **(Maximum Likelihhood.)** Consider the following parameter estimation problem. Let $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ be known constants. The d-dimensional instance vectors $X_1, X_2, \ldots, X_n$ are drawn from some distributions in an i.i.d. fashion. The real valued targets $Y_1, \ldots, Y_n$ are such that Y_i is drawn from a Normal distribution with mean $\mathbf{x}_i^\top \mathbf{w}^*$ and variance σ_i^2, for some fixed but unknown parameter $\mathbf{w}^* \in \mathbb{R}^d$. Derive the maximum likelihood estimate of $\mathbf{w}^*$.

Assume you have access to an equation solver sub-routine that takes in $A \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ and returns a solution to $A\mathbf{x} = \mathbf{b}$ (if a solution exists). How will you use this solver for this parameter estimation problem, for a given dataset with instances $\mathbf{x}_1, \ldots, \mathbf{x}_n$, targets $y_1, \ldots, y_n$ and noise variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. **(4+2 points)**

(5) **(AdaBoost.)** Consider the following binary classification dataset. Run AdaBoost for 3 iterations on the dataset, with the weak learner returning a best decision stump (equivalently a decision tree with one node) (equivalently a horizontal or vertical separator). Ties can be broken arbitrarily. Give the objects asked for below. Highlight your answer by boxing it.

x_1	x_2	y
1	1	$+1$
1	2	-1
1	3	$+1$
2	1	-1
2	2	-1
2	3	-1
3	1	$+1$
3	2	$+1$
3	3	$+1$

i. Give the weak learners h_t for $t = 1, 2, 3$.

ii. Give the "edge over random" γ_t, and the multiplicative factor β_t for $t = 1, 2, 3$.

iii. Give the predictions of the final weighted classifier h on the training points.

(2+2+2 points)

(6) (**Naive Bayes Methods**) Consider a distribution over (X, Y) given by the following assumptions:
$Y \in \{-1, +1\}, X \in \{0, 1\}^3$.
$P(Y = +1) = a, \mathbf{P}(Y = -1) = 1 - a$,
$X|Y = -1 \sim \text{Bern}(\theta_1) \times \text{Bern}(\theta_2) \times \text{Bern}(\theta_3)$,
$X|Y = +1 \sim \text{Bern}(\tau_1) \times \text{Bern}(\tau_2) \times \text{Bern}(\tau_3)$.

We have 10 training points from the above distribution, given by the table below.

X_1	X_2	X_3	Y
1	0	0	$+1$
0	1	1	-1
0	1	0	$+1$
1	1	0	$+1$
1	1	1	-1
1	0	0	$+1$
1	0	1	$+1$
0	0	1	-1
0	1	1	$+1$
0	0	0	-1

i. Give the ML estimates for $a, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3$.

ii. For all the 8 points X in the instance space $\{0, 1\}^3$, give the estimate of the posterior probability $\mathbf{P}(Y = +1|X)$, and give the prediction that minimises the misclassification rate (or the Bayes classifier for the zero-one loss), in the form of a table with 8 rows. **(3+3 Points)**