INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

End-Term Examination (ETE)

Machine Learning (CSN-382)

Time: 180 minutes

Spring Semester 2024-25

Total Marks: 100

Instructions: Each problem has a relatively simple and straightforward solution, and we may deduct points for overly complex answers. Therefore, focus on providing clear and concise solutions that directly address the problem at hand.

Problem 1 (10 marks)

1. (a)

(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**) $\mathbb{P}[y = 1 | \hat{y} = 1]$

Recall (**REC**) $\mathbb{P}[\hat{y} = 1 | y = 1]$

False discovery rate (**FDR**) $\mathbb{P}[y = -1 | \hat{y} = 1]$

False omission rate (**FOR**) $\mathbb{P}[y = 1 | \hat{y} = -1]$

1. (b)

(Kernel Smash) Melbi has created two Mercer kernels $K_1, K_2 \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with the feature map for the kernel K_i being $\phi_i \colon \mathbb{R} \to \mathbb{R}^2$. 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 $\phi_3 \colon \mathbb{R} \to \mathbb{R}^7$ for the kernel K_3.

Note that ϕ_3 must not use more than 7 dimensions.

(5 marks)

$$\phi_3(x) = ?$$

Problem 2 (10 marks = 5+5)

(Positive Linear Regression) We have data features $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^D$ and labels $y_1, \dots, y_N \in \mathbb{R}$ stylized as $X \in \mathbb{R}^{N \times D}, \mathbf{y} \in \mathbb{R}^N$. We wish to fit a linear model with positive coefficients:

$$\underset{\mathbf{w} \in \mathbb{R}^D,}{\arg\min} \frac{1}{2} \|X\mathbf{w} - \mathbf{y}\|_2^2 \text{ s.t. } w_j \geq 0 \text{ for all } j \in [D]$$

1. Write the Lagrangian for this problem by introducing dual variables (no derivation needed).
2. Simplify the dual problem (eliminate $\mathbf{w}$) – show major steps. Assume $X^\top X$ is invertible.

Problem 3 (10 marks)

3. (a) (6 marks)

(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* 😊.

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	$\square$	$(0,0)$

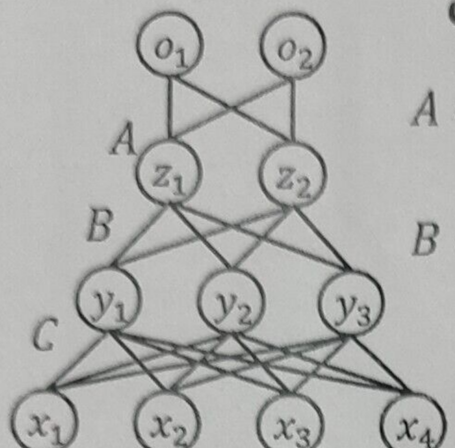
1. What is the entropy of the root node?
2. What is the entropy of the two child nodes (give answers for the two nodes separately) if the split is done using the x feature ($x = 0$ becomes left child, $x = 1$ becomes right child)?
3. What is the reduction in entropy (i.e., $H_{\text{root}} - H_{\text{children}}$) if the split is done using the x feature as described above?
4. 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)?
5. 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?
6. To get the most entropy reduction, should we split using x feature or y feature?

3. (b) (4 marks): What is the role of the learning rate in gradient descent? What can go wrong if it is too high or too low?

Problem 4 (10 marks)

4. (a) (5 marks)

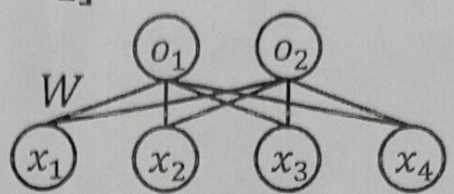
Consider the NN with 2 hidden layers – all nodes use the identity activation function. This NN is clearly equivalent to a network with no hidden layers since all activation functions are linear. Find the weights of this new network



$$\mathbf{o} = A\mathbf{z}, \mathbf{z} = B\mathbf{y}, \mathbf{y} = C\mathbf{x}, \qquad \mathbf{o} = W\mathbf{x}$$

$$A = \begin{bmatrix} 2 & 0 \\ 2 & 2 \end{bmatrix}, C = \begin{bmatrix} 1 & 3 & 0 & 1 \\ 4 & 1 & 4 & 0 \\ 0 & 2 & 0 & 2 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 3 & 1 \end{bmatrix} \qquad W^\top = \begin{bmatrix} \quad ? \quad \end{bmatrix}$$



4. (b) (5 marks): Explain the structure and function of an artificial neural network (ANN). Describe the roles of weights, biases, and activation functions.

Problem 5 (10 marks): Maximum likelihood

Consider the following probability distribution:

$$P_\theta(x) = 2\theta x e^{-\theta x^2}$$

where θ is a parameter and x is a positive real number. Suppose you get m i.i.d. samples x_i drawn from this distribution. Show how one can compute the maximum likelihood estimator for θ based on these samples.

Problem 6 (10 marks)

6. (a) (5 marks)

Let's do principal components analysis (PCA)! Consider this sample of six points $X_i \in \mathbb{R}^2$.

$$\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix} \right\}.$$

6.(b).1. [2 Marks] Compute the mean of the sample points and write the centered design matrix (By subtracting the mean from each sample).

6.(b).2. [3 Marks] Find all the principal components of this sample. Write them as unit vectors.

6. (b) (5 marks)

Support vector machines learn a decision boundary leading to the largest margin from both classes. You are training SVM on a tiny dataset with 4 points shown in the Figure. This dataset consists of two examples with class label +1 (denoted with plus), and two examples with class label -1 (denoted with triangles)

What's the equation corresponding to the decision boundary?

Problem 7 (10 marks)

7. (a) (5 marks): Explain the working of Principal Component Analysis (PCA) and how it achieves dimensionality reduction.

7. (b) (5 marks): How do you determine the optimal number of clusters (K) in K-means?

Problem 8 (10 marks)

8. (a) (5 marks): You applied K-means clustering to a dataset with two features: height (in cm) and weight (in kg). The algorithm formed poor clusters. What might be the issue?

8. (b) (5 marks): In a kernelized SVM using a nonlinear kernel (e.g., RBF), the decision boundary in the input space appears curved. Yet, we say SVM finds a linear separator. Isn't this a contradiction?

Problem 9 (10 marks)

9. (a) (5 marks): Suppose you train a hard-margin SVM on a perfectly linearly separable dataset. Does this guarantee 100% test accuracy? Why or why not?

9. (b) (5 marks): Logistic regression outputs probabilities using a sigmoid function, which is nonlinear. So how can it be called a linear classifier?

Problem 10 (10 marks)

10. (a) (5 marks): Why do we use the mean squared error (MSE) as the cost function in linear regression, and not just absolute error?

10. (b) (5 marks): What problem might arise when applying standard k-fold cross-validation to imbalanced datasets? How can you address it?