Department of Computer Science
University of Delhi
Master of Computer Application
MCAC 204: Machine Learning
Unique Paper Code: 223421206

Semester II
Year of admission: 2023

Time: Three Hours **Max. Marks: 70**

Instructions:

1. All questions carry equal marks.
2. Use proper notation and show complete working for full credit.

1	a. How does the choice of learning rate (α) affect the convergence and stability of the gradient descent algorithm in linear regression? Illustrate with the help of a suitable example.	[4]
	b. Consider the following dataset: Use the candidate elimination learning algorithm to determine the most general and the most specific hypothesis for the given training data.	[4]
	c. Consider the training examples $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_m, y_m)\}$ and the following cost function for a linear regression problem: $$J(w,b) = \frac{1}{2m}\sum_{i=1}^{m}\left(wx^{(i)} + b - y^{((i))}\right)^2$$ Prove that the parameters w and b that minimizes $J(w,b)$ can be derived as $\frac{cov(x,y)}{var(x)}$ and $\bar{y} - w\bar{x}$, respectively.	[6]
2	a. What is the purpose of feature scaling in multiple linear regression? Illustrate the same with the help of a suitable example.	[4]
	b. How does the sigmoid function define the decision boundary in logistic regression? Also, give the intuition behind the cost function used in logistic regression with regularization.	[5]

Size	Color	Shape	Class/ Label
Big	Red	Circle	No
Small	Red	Triangle	No
Small	Red	Circle	Yes
Big	Blue	Circle	No
Small	Blue	Circle	Yes

c. Consider a dataset with binary class labels (+1 or -1). You are using the AdaBoost algorithm with decision stumps as the base classifiers. Initially, all samples are assigned equal weights. The first decision stump is trained and achieves an error rate of 0.2. [5]

i. Calculate the Amount of Say (alpha) assigned to this weak learner in the final ensemble.

ii. After updating the sample weights based on the performance of the first weak learner, determine the weight assigned to a sample that was misclassified by this weak learner.

2 a. Each binary classifier in an ensemble makes predictions on an input $\mathbf{x}$, as shown in the table below. Using this table, find the ensemble model's aggregated prediction for $\mathbf{x}$: [2]

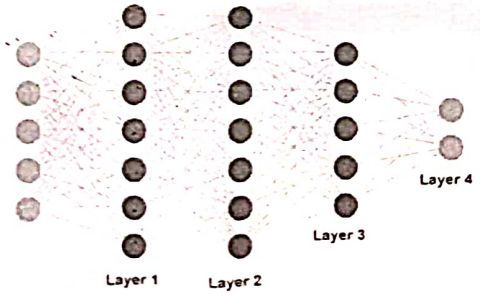
	Classifier's Confidence	Prediction
Classifier 1	0.61	+1
Classifier 2	0.53	-1
Classifier 3	0.88	-1
Classifier 4	0.34	+1

b. With the help of a suitable diagram, prove that the width of the margin in the linear support vector classifier (LSVM) is $\frac{2}{\|\vec{w}\|}$, where $\vec{w}$ is a vector perpendicular to the decision boundary. [6]

c. Find a principal component of the dataset given below (Show all the intermediate steps): [6]

Age	Weight
20	65
25	60
5	15
10	20

3 a. Consider the following neural network, comprising four layers: layer 1, layer 2, layer 3, and layer 4. Layers 1, 2, and 3 are hidden layers, and layer 4 is the output layer. [6]

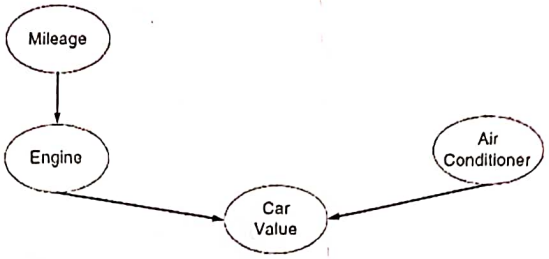


Determine the following:

i. dimensions of $W^{[1]}, b^{[1]}, W^{[2]}, b^{[2]}, W^{[3]}, b^{[3]}, W^{[4]}, b^{[4]}$

ii. number of trainable parameters for Layer 1.

b. Given the Bayesian belief network (BBN) and the corresponding dataset shown below: [8]



Mileage	Engine	Air Conditioner	Number of Records with Car Value = High	Number of Records with Car Value = Low
High	Good	Working	3	4
High	Good	Broken	1	2
High	Bad	Working	1	5
High	Bad	Broken	0	4
Low	Good	Working	10	0
Low	Good	Broken	3	1
Low	Bad	Working	2	2
Low	Bad	Broken	0	2

i. Draw the probability table for each node in the BBN.

ii. Use the BBN to compute $P(Engine = Bad, Air\ Conditioner = Broken)$.

4 a. For a classifier that distinguishes between cat vs non-cat, which of the four activations should be employed at the output layer: Sigmoid, Leaky ReLU, and Tanh? Justify your answer. [2]

b. For a neural network that outputs the likelihood of occurrence of a disease out of six, which activation functions should be employed: Sigmoid, ReLU, SoftMax, and Tanh? Justify your answer. [2]

c. Consider a dataset comprising 1 billion instances. Should a 70:30 train-test split be a good choice? Justify your choice. [2]

d. For the following classification tasks, identify which metric, precision, or recall would be most suitable. Justify your answer. [3]

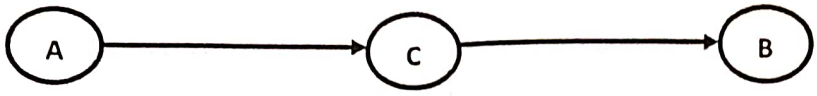
i) COVID-19 identification using chest X-rays.

ii) Spam e-mail classification.

e. Write kernel K-means algorithm. [5]

5

a. Consider the following diagram. [3]



Show that when conditioned on **C**, **A** and **B** are orthogonal to each other.

b. Assume we have two users and three movies. The 2×3 matrix Y is given below: [4]

$$Y = \begin{bmatrix} 1 & 2 & ? \\ ? & 10 & 15 \end{bmatrix}$$

Our goal is to find the matrices U and V such that $X = UV^T$ closely approximates the observed ratings in Y. Assume we start by fixing V to initial values of $[2 \quad 4 \quad 6]^T$. Find the optimal 2×1 vector U in this case. (Express your answer in terms of λ, where λ is a regularization parameter).

c. Consider a problem with three binary random variables relating to the fuel system on a car. The variables are called **B**, representing the state of a battery that is either **charged (B = 1)** or **flat (B = 0)**, **F** representing the state of the fuel tank that is either **full of fuel (F = 1)** or **empty (F = 0)**, and **G**, which is the state of an electric fuel gauge and which indicates either **full (G = 1)** or **empty (G = 0)**. The battery is either charged or flat, and independently, the fuel tank is either full or empty, with prior probabilities given below: [2+2 +3]

```
p (B = 1) = 0.9
p (F = 1) = 0.9
```

Also, consider the probability table given below:

```
p (G = 1|B = 1, F = 1) = 0.8
p (G = 1|B = 1, F = 0) = 0.2
p (G = 1|B = 0, F = 1) = 0.2
p (G = 1|B = 0, F = 0) = 0.1
```

Compute the following:

```
i.   p (G = 0)
ii.  p (F = 0 | G = 0)
iii. p (F = 0 | G = 0, B = 0)
```