INTRODUCTION TO MACHINE LEARNING
(ECEN 4122)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Which are applicable for supervised learning?
(a) Linear Regression (b) DB Scan
(c) Decision Tree (d) Both (a) and (c).

(ii) A typical problem in customer segmentation generally falls under
(a) Supervised learning (b) Unsupervised learning
(c) Both of above (d) Conditional supervised learning.

(iii) In case of linear regression, a hypothetic model is defined as $h_\theta(x) = \theta_0 + \theta_1 x$. Here the θ_1 is associated with
(a) Features (b) Gradient
(c) Learning rate (d) Performance metric.

(iv) An over-fitting model corresponds to
(a) Low bias and high variance (b) High bias and high variance
(c) Low bias and low variance (d) Low bias and high variance.

(v) A linear regression model has training accuracy 90% and test accuracy 80%. The model is
(a) Generalized (b) Under fitting
(c) Over fitting (d) Can not determine.

(vi) A suitable gradient descent curve should be a
(a) Concave function (b) Linear function
(c) Non linear function (d) Convex function.

(vii) Confusion matrix is suitable for a
(a) Classification problem (b) Regression problem
(c) Both of (a) & (b) (d) Any one of (a) & (b).

(viii) A spam classification learning model should be checked by
(a) Precision (b) Recall
(c) Accuracy (d) Both (a) and (c).

(ix) For a binary classification problem, TP=240, TN=300, FP=700, FN=100. Precision of the model is
(a) 25.53% (b) 26.2% (c) 24.44% (d) 28.35%.

(x) In a regression model, training data are given as [1, 2, 3, 4], corresponding outputs from the model are [1.2, 2, 3.1, 4.5]. The R^2 metric will be
(a) 0.952 (b) 0.998 (c) 0.912 (d) 0.900.

Fill in the blanks with the correct word

(xi) Probability of taking a green ball after taking a red ball from a box is given as 70%. The probability of taking a red ball after taking a green ball from that box is _____.

(xii) A high bias and high variance is called _____.

(xiii) Confusion matrix is associated with _____.

(xiv) Sigmoid function is associated with _____.

(xv) Purity of a pure split is obtained by calculating _____.

Group - B

2. (a) A hypothetical model used in linear regression is given as $h_\theta(x) = \theta_0 + \theta_1 x$. Based on the convergence algorithm, find the update equation of θ_0 and θ_1.
[(CO3)(Analyse/HOCQ)]

(b) A learning model in linear regression uses $h_\theta(x) = \theta_1 x$. The training data set is given as $(1,1), (2,2), (3,3)$. Considering the learning rate $\alpha = 0.1$, find the last updated value of θ_1 and corresponding cost function after 3 iteration.
[(CO2)(Understand/LOCQ)]

4 + 8 = 12

3. (a) Explain, how can you handle an over-fitting model using Ridge regularizations.
[(CO3)(Analyse/HOCQ)]

(b) Suppose, your prediction model reached the global minima of corresponding gradient descent. What problem can occur in this case? How can you solve this problem using Lasso regularization? *[(CO4)(Remember/LOCQ)][(CO2)(Apply/IOCQ)]*

6 + 6 = 12

Group - C

4. (a) Why a sigmoid function such as $\frac{1}{1+e^{-z}}$ is not suitable for the logistic regression?
[(CO3)(Analyse/HOCQ)]

(b) A confusion matrix is given as shown below. Calculate Accuracy, Precision, Recall and F-0.5 score.

Predicted \ Actual	1	0
1	10000	4000
0	400	5600

[(CO4)(Remember/LOCQ)]

6 + 6 = 12

5. (a) Based on the dataset given below, decide whether you should drive when you have rainy weather, average road, normal traffic with no engine problem.

SNo.	Weather condition	Road condition	Traffic condition	Engine problem	Accident
1	Rain	bad	high	no	yes
2	snow	average	normal	yes	yes
3	clear	bad	light	no	no
4	clear	good	light	yes	yes
5	snow	good	normal	no	no
6	rain	average	light	no	no
7	rain	good	normal	no	no
8	snow	bad	high	no	yes
9	clear	good	high	yes	no
10	clear	bad	high	yes	yes

[(CO3)(Analyse/HOCQ)]

(b) Derive the Baye's equation for two dependent events A and B. *[(CO4)(Remember/LOCQ)]*

10 + 2 = 12

Group – D

6. (a) What is a pure split and how can you find a pure node? *[(CO3)(Analyse/HOCQ)]*

(b) Why the information gain is important in decision tree? *[(CO4)(Remember/LOCQ)]*

(c) A node f_N with probability $P(yes) = 50\%$ is having two categories as C_1 and C_2. C_1 is having 6 YES, 5 NOs and C_2 with 8 YES. Calculate the information gain of f_N. *[(CO3)(Analyse/HOCQ)]*

2 + 2 + 8 = 12

7. (a) If you want to describe a decision tree based on the following dataset, find which feature will be the first to start with, having highest information gain?

Type of family structure	Age group	Income status	Will they buy a car?
Nuclear	Young	Low	Yes
Extended	Old	Low	No
Childless	Middle-aged	Low	No
Childless	Young	Medium	Yes
Single Parent	Middle-aged	Medium	Yes
Childless	Young	Low	No
Nuclear	Old	High	Yes
Nuclear	Middle-aged	Medium	Yes
Extended	Middle-aged	High	Yes
Single Parent	Old	Low	No

[(CO3)(Analyse/HOCQ)]

(b) Based on the above dataset, draw the decision tree for Income status and calculate entropy of medium income status. *[(CO4)(Remember/LOCQ)]*

8 + 4 = 12

Group - E

8. (a) Describing the K means and Hierarchical clustering , point out the basic difference between them. *[(CO3)(Analyse/HOCQ)]*

(b) How can you validate a clustering model using Silhouette score? *[(CO4)(Remember/LOCQ)]*

6 + 6 = 12

9. (a) Explain the SVM algorithm. *[(CO3)(Analyse/HOCQ)]*

(b) With proper diagram, explain the hyper-plane and marginal plane. Write the equation used to maximize the separation between the marginal planes. *[(CO4)(Remember/LOCQ)]*

6 + 6 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	35.41	6.25	58.34

Course Outcome (CO):

After the completion of the course students will be able to

1. Select an appropriate Machine Learning tool for analyzing data in a given feature space.
2. Apply machine learning techniques such as regression, classification, clustering, and feature selection to detect patterns in the data.
3. Distinguish between supervised, and unsupervised learning.
4. Outline solution for classification and regression approaches in real-world applications.
5. Formulate a machine learning problem.
6. Determine cutting edge technologies related to machine learning applications.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*