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) What is the purpose of a validation set in supervised learning?
(a) To test the model (b) To fine-tune hyperparameters
(c) To train the model (d) To reduce the model's complexity.

(ii) Which evaluation metric is commonly used for regression models?
(a) Accuracy (b) Precision (c) Mean Squared Error (MSE) (d) F1-Score.

(iii) Which of the following helps in preventing overfitting in a machine learning model?
(a) Adding more layers (b) Removing regularization
(c) Using a smaller dataset (d) Applying cross-validation.

(iv) What is the purpose of hyperparameter tuning in supervised learning?
(a) To reduce the dimensionality of data (b) To improve the model's performance
(c) To increase the training data size (d) To reduce the number of classes in the dataset.

(v) Which of the following is an ensemble learning method?
(a) Decision Tree (b) Logistic Regression (c) Bagging (d) K-means.

(vi) Which of the following is the key difference between bagging and boosting?
(a) Bagging reduces variance, boosting reduces bias
(b) Boosting reduces variance, bagging reduces bias
(c) Bagging increases the training dataset size
(d) Boosting uses parallel learning, bagging uses sequential learning.

(vii) Which of the following is a density-based clustering algorithm?
(a) K-means (b) DBSCAN (c) Hierarchical Clustering (d) KNN.

(viii) What is the role of the convolutional layer in a Convolutional Neural Network (CNN)?
(a) To reduce the number of layers (b) To capture spatial features from the input
(c) To classify data (d) To reduce overfitting.

(ix) Which of the following activation functions is commonly used in deep learning models?
(a) Sigmoid (b) Mean Squared Error
(c) ReLU (d) Accuracy.

(x) What is the primary purpose of dropout in deep learning?
(a) To speed up training (b) To prevent overfitting
(c) To reduce the number of features (d) To increase accuracy.

Fill in the blanks with the correct word

(xi) The _____ algorithm is a popular classification algorithm that works by finding the hyperplane that best separates the data points into classes.

(xii) In a decision tree, the process of splitting the nodes is based on a metric called _____, which measures the purity of the split.

(xiii) The _____ is an ensemble learning technique that reduces variance by training multiple models on different subsets of the data and averaging their predictions.

(xiv) In logistic regression, the _____ function maps the predicted values to probabilities.

(xv) In the context of regularization, the _____ method is also known as Lasso and is used to shrink some model coefficients to zero.

Group – B

2. (a) Explain the difference between classification and regression. Provide examples for each. *[CO3][Understand/LOCQ]*

(b) Apply logistic regression to classify whether customers will make a purchase based on features such as age, income, and previous purchases. Evaluate the model using accuracy and precision. *[CO1][Analyse/IOCQ]*

Customer_ID	Age	Income (k$)	Previous_Purchases	Will_Purchase (0/1)
1	25	40	2	0
2	35	55	5	1
3	45	70	3	1
4	22	30	1	0
5	50	90	6	1

6 + 6 = 12

3. (a) Discuss the steps involved in pre-processing a dataset for supervised learning. What are some common techniques used in data pre-processing? *[CO1][Remember/LOCQ]*

(b) Implement data normalization and feature scaling on a house price dataset and train a support vector machine (SVM) to predict house prices. Compare the results with and without scaling. *[CO1][Analyse/IOCQ]*

House_ID	Size (sqft)	Bedrooms	Age (years)	Price (lakhs)
1	1500	3	10	75
2	1200	2	15	60
3	1800	4	5	90
4	1100	2	20	55
5	1400	3	8	70

6 + 6 = 12

Group – C

4. (a) Describe logistic regression and explain how it differs from linear regression. Provide examples of its applications. *[CO2][Understand/LOCQ]*

(b) Apply logistic regression to classify whether a customer will make a purchase based on the given customer data (age, income, previous purchases). Evaluate the model using accuracy and precision. *[CO2][Analyse/IOCQ]*

Customer_ID	Age	Income (k$)	Previous_Purchases	Will_Purchase (0/1)
1	25	40	2	0
2	35	55	5	1
3	45	70	3	1
4	22	30	1	0
5	50	90	6	1

6 + 6 = 12

5. (a) Discuss the key advantages and limitations of using support vector machines (SVMs) for classification tasks. How do kernel functions enhance the performance of SVMs? *[CO4][Analyse/IOCQ]*

(b) Train a support vector machine (SVM) with a linear and RBF kernel on the following health dataset to classify whether a patient has diabetes. Compare the results in terms of accuracy and recall. *[CO4][Analyse/IOCQ]*

Patient_ID	Age	BMI	Blood_Pressure	Cholesterol	Diabetes (0/1)
1	45	28.5	120	180	1
2	50	32.0	140	200	0
3	35	25.0	110	170	0
4	60	30.5	130	220	1
5	40	27.0	115	190	0

6 + 6 = 12

Group – D

6. (a) Define clustering and explain the differences between hierarchical and partitioning clustering methods. *[CO2][Understand/LOCQ]*

(b) Using the given dataset, perform hierarchical clustering (Agglomerative). Create a dendrogram and use it to decide the optimal number of clusters. Report the clusters formed and evaluate their effectiveness. *[CO1][Analyse/IOCQ]*

Customer_ID	Income (k$)	Spending_Score (1-100)
1	40	60
2	80	90
3	20	30
4	50	70
5	60	50

6 + 6 = 12

7. (a) Explain how a self-organizing map (SOM) works for clustering and visualizing high-dimensional data. Discuss its applications. *[CO6][Analyse/LOCQ]*

(b) Using the provided customer data, create a self-organizing map (SOM) to visualize the clusters of customers. Compare the clustering results with those obtained from K-means and discuss the differences. *[CO6][Create/HOCQ]*

Customer_ID	Age	Income (k$)	Spending_Score (1-100)	Previous_Purchases	Region
1	25	40	60	2	A
2	35	55	90	5	B
3	45	70	30	3	A
4	22	30	70	1	B
5	50	90	50	6	A

6 + 6 = 12

Group – E

8. (a) Describe autoencoders in the context of deep learning and explain how they can be used for dimensionality reduction. What are some of the key differences between PCA and autoencoders? *[CO3][Understand/LOCQ]*

(b) Train an autoencoder on the following dataset to reduce the dimensionality of the features. Compare the reconstruction error of the autoencoder with the error from PCA for the same dataset. *[CO3][Analyse/IOCQ]*

Patient_ID	Age	BMI	Blood_Pressure	Cholesterol	Glucose
1	45	28.5	120	180	95
2	50	32.0	140	200	100
3	35	25.0	110	170	90
4	60	30.5	130	220	105
5	40	27.0	115	190	92

6 + 6 = 12

9. (a) Discuss the architecture of a Recurrent Neural Network (RNN) and explain its advantages for handling sequential data. How does a Long Short-Term Memory (LSTM) network address the limitations of a standard RNN? *[CO6][Analyse/IOCQ]*

(b) Train an LSTM network on the following stock price dataset to predict the future stock prices. Compare the performance of the LSTM model with that of a basic RNN using Mean Absolute Error (MAE) as the evaluation metric. *[CO6][Create/HOCQ]*

Date	Open	High	Low	Close	Volume
2024-01-01	100.5	102.5	99.5	101.0	10000
2024-01-02	101.0	103.0	100.0	102.0	12000
2024-01-03	102.0	104.0	101.0	103.0	14000
2024-01-04	103.0	105.0	102.0	104.0	16000
2024-01-05	104.0	106.0	103.0	105.0	18000

6 + 6 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	37.5	50	12.5