MACHINE LEARNING AND INTRODUCTION TO PYTHON
(ECEN 3233)

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

Figures out of the right margin indicate full marks.

*Candidates are required to answer Group A and
<u>any 4 (four)</u> from Group B to E, taking <u>one</u> 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 of the following is an example of Supervised Learning?
            (a) Clustering customers based on behavior
            (b) Predicting house prices based on features
            (c) Detecting anomalies in network traffic
            (d) Playing a video game using trial and error

    (ii)   In k-fold cross-validation, the dataset is split into
            (a) k equal-sized overlapping subsets
            (b) k equal-sized non-overlapping subsets
            (c) 2 unequal subsets
            (d) A training and a testing set only

    (iii)   What is the primary criterion used for splitting in a decision tree (for classification)
            (a) Mean Squared Error
            (b) Log Loss
            (c) Gini Index / Entropy
            (d) number of attributes

    (iv)   What is the purpose of Information Gain in decision trees?
            (a) To measure prediction accuracy
            (b) To find the best split by reducing impurity
            (c) To regularize the model
            (d) To calculate loss

    (v)   What does a neuron in an artificial neural network do?
            (a) Stores data
            (b) Applies a linear transformation only
            (c) Takes input, applies weights, bias, and an activation function
            (d) Computes loss

(vi)   What is the main goal of a Support Vector Machine?
(a) To reduce the number of features
(b) To find a decision boundary that maximizes margin
(c) To calculate the mean of the dataset
(d) To minimize the loss of neural networks

(vii) Which technique focuses on training new models to correct the errors made by previous ones?
(a) Bagging   (b) Boosting   (c) Random Forest   (d) Dropout

(viii) The purpose of the "elbow method" in clustering is
(a) To reduce dimensions   (b) To find the optimal number of clusters
(c) To normalize data   (d) To rank features

(ix)   Which of the following libraries is most commonly used for machine learning in Python?
(a) NumPy   (b) Pandas   (c) Scikit-learn   (d) Matplotlib

(x)   "Curse of dimensionality" refers to
(a) Difficulty in visualizing data in higher dimensions
(b) The increase in computation time as the number of features increases
(c) The ability to capture relationships in high-dimensional data
(d) Both (a) and (b)

Fill in the blanks with the correct word

(xi)   The Candidate Elimination Algorithm computes the version space by considering only positive instances _____ (True/False).

(xii) Combining multiple models to improve performance is the basic idea in _____learning.

(xiii) _____is a classic example of a bagging-based ensemble method.

(xiv) Python uses zero-based indexing, i.e., the first element of a list has index 0 (zero) _____ (True/False).

(xv)   A Dictionary in Python is an _____collection of data values.

Group - B

2. (a)   Mention two normalization techniques that are used in data transformations. Why is normalization done?   *[(CO1) (Remember/LOCQ)]*
   (b)   Consider a data set, D={34, 36,42, 48}. Apply a suitable normalization technique to map the data to a new range 0-1.   *[(CO1) (Apply/IOCQ)]*
                                                                                      6 + 6 = 12

3. (a)   A confusion matrix related to a binary classification is shown in the table. Calculate accuracy, sensitivity and precision of the model.   *[(CO2)(Evaluate/HOCQ)]*

	Actual class 1	Actual class 2
Placed in class 1	26	12
Placed in class 2	08	54

(b) Explain the concept of bias-variance trade-off in machine learning

[(CO1)(Understand/LOCQ)]

7 + 5 = 12

Group - C

4. (a) State the limitation of linear regression. How can it be overcome?

[(CO4)(Understand/LOCQ)]

(b) Train a logistic regression model to predict whether a student will pass based on hours studied. Predict the outcome for a student who studies 4.5 hours. Take coefficients a_0= -1.5 and a_1=0.6, where the symbols have usual meanings.

[(CO4)(Create/HOCQ)]

Hours Studied	Exam Result Pass=1, Fail=0
2	0
3	0
4	0
5	1
6	1
7	1
8	1

4 + 8 = 12

5. (a) Define entropy of a sample space S. Show that the entropy of a homogeneous set of examples is zero.

[(CO4)(Apply/IOCQ)]

(b) Evaluate the Information Gain for the attribute 'Weather' from the following data set where the target attribute is the 'Play Tennis'.

[(CO4)(Evaluate/HOCQ)]

Weather	Wind	Play Tennis
Sunny	Weak	No
Sunny	Strong	No
Cloudy	Weak	Yes
Rain	Weak	Yes
Rain	Strong	No
Cloudy	Strong	Yes
Sunny	Weak	Yes

5 + 7 = 12

Group - D

6. (a) What is a perceptron?

[(CO4)(Understand/LOCQ)]

(b) Explain the Delta rule of a perceptron. State its limitation and how can it be overcome?

[(CO4)(Remember/LOCQ)]

(c) Can the AND and XOR Boolean functions be solved using the Delta rule of a Perceptron? Justify your answer.

[(CO4)(Evaluate/HOCQ)]

1 + 7 + 4 =12

7. (a) Show with a block diagram, why and how a kernel is used in non-linear SVMs.
   [CO4)(Apply/IOCQ)]

   (b) Write 3 common kernel functions used in SVM. *[CO4)(Remember/LOCQ)]*

   6 + 6 = 12

Group - E

8. (a) What is clustering in machine learning? Distinguish between classification and clustering. *[CO5)(Analyze/IOCQ)]*

   (b) In K-Means clustering, what does 'K' represent? What is a centroid ?
   [CO5)(Understand/LOCQ)]

   (c) What kind of distance metric is commonly used in clustering? Why is such a metric necessary? Calculate the Manhattan distance between two objects whose coordinates are (0,3) and (5,8). *[(CO5)(Analyze,Evaluate/IOCQ, HOCQ)]*

   4 + 3 + 5 = 12

9. (a) Explain the Identifiers, Keywords, Statements, Expressions, and Variables in Python programming language with examples. *[(CO6) (Understand/LOCQ)]*

   (b) Write a Python program using suitable libraries and functions to form a confusion matrix. This might include the import of the dataset, splitting of the training and the test data, model training, testing the accuracy for the test data and finally obtaining the confusion matrix. *[(CO6)(Create/HOCQ)]*

   5 + 7 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	38.54	21.88	39.58