

**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
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) Impurity of an attribute is high means
 - (a) this attribute makes the maximum difference to the classification
 - (b) this attribute makes little difference to the classification
 - (c) this attribute can be used for verification of unknown patterns
 - (d) this attribute makes not so much difference to the classification.
- (ii) 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) high bias and low variance.
- (iii) Feature selection property exhibits by
 - (a) ridge regression
 - (b) logistic regression
 - (c) lasso regression
 - (d) all of the above.
- (iv) Gradient descent of a non-convex cost function suffers from
 - (a) two local minima
 - (b) no global minima
 - (c) multiple global minima
 - (d) multiple local minima.
- (v) A type of machine learning that uses agents and environment interaction for creating labelled data is
 - (a) supervised learning
 - (b) unsupervised learning
 - (c) reinforcement learning
 - (d) semi-supervised learning.
- (vi) The Candidate Elimination Algorithm computes the version space by considering
 - (a) only positive instances
 - (b) only negative instances
 - (c) both positive and negative instances
 - (d) none of these.
- (vii) Which of the following is the most desirable option for a learning algorithm to exhibit?
 - (a) Low bias, low variance
 - (b) Low bias, high variance
 - (c) High bias, low variance
 - (d) High bias, high variance.

- (viii) Which of the following is a disadvantage of decision trees?
- Factor analysis
 - Decision trees are robust to outliers
 - Decision trees are prone to be overfit
 - None of the above.
- (ix) Suppose your model is overfitting. Which of the following is NOT a valid way to try to reduce the overfitting?
- Increase the amount of training data
 - Improve the optimization algorithm being used for error minimization
 - Decrease the model complexity
 - Reduce the noise in the training data.
- (x) The full form of SVM in Machine learning is
- State Vector Machine
 - State Variable Machine
 - Support Variable Machine
 - Support Vector Machine.

Fill in the blanks with the correct word

- (xi) Purity of a pure split is obtained by calculating _____.
- (xii) Target variable is a _____ (dependent/independent) variable.
- (xiii) A low bias and low variance is called _____.
- (xiv) The average of absolute residuals is mean absolute error _____ (True/False).
- (xv) A Dictionary in Python is an _ordered _____ collection of data values.

Group - B

2. (a) Mention the different types of Machine learning. [[CO1](Remember/LOCQ)]
- (b) What is supervised learning? How is a classification problem different from a regression problem? [[CO1](Understand/LOCQ)]
- (c) Generate the set of consistent hypotheses using Candidate Elimination algorithm for the data in the following table that describes the symptoms of the people and their Covid-19 test results. [[CO1](Create/HOCQ)]

Instance No.	Fever	Cough	Throat pain	Body pain	Covid 19
1	No	Yes	Yes	Yes	Positive
2	Yes	Yes	Yes	Yes	Positive
3	Yes	No	Yes	Yes	Positive
4	No	Yes	No	No	Negative
5	Yes	Yes	Yes	No	Positive
6	No	No	No	Yes	Negative
7	No	No	no	No	Negative

2 + 4 + 6 = 12

3. (a) What do you understand by multi-label classification? [[CO3](Understand/LOCQ)]
- (b) Construct (induct) a decision tree for the following problem given in Table.2. What is the size and depth of the constructed tree? Hence, calculate the Entropy

and Gini index for short length hair. Discuss the significance of information gain used to induct decision tree.

[[CO4](Evaluate/HOCQ)]

Table. 2

Sl. No	Hair length	Under 170 cm	Gender
1	Short	No	Male
2	Long	Yes	Female
3	Long	Yes	Female
4	Short	No	Female
5	Long	No	Male
6	Short	Yes	Male

- (c) State the min-max normalization process for pre-processing of dataset.

[[CO1](Remember/LOCQ)]

2 + 8 + 2 = 12

Group - C

4. (a) Differentiate between simple regression and multiple regression models.

[[CO4](Understand/LOCQ)]

- (b) State the limitations of linear regression.

[[CO4](Understand/LOCQ)]

- (c) Show how a sigmoid function may be used as a binary classifier in logistic regression.

[[CO4](Apply/IOCQ)]

2 + 2 + 8 = 12

5. (a) Data of weekly sales of a company is provided in the following table. Predict the 9th week sale using linear regression.

[[CO4](Evaluate/HOCQ)]

Table. 3

Sl. No	Week	Sales(in thousands)
1	1	1.2
2	2	1.8
3	3	2.6
4	4	3.2
5	5	3.8

- (b) Discuss how Random Forest (RF) is different from custom bragging technique. Also explain why is normalization necessary for RF?

[[CO4](Understand/LOCQ)]

- (c) Explain the framework of Adaboost technique.

[[CO2](Remember/LOCQ)]

4 + 4 + 4 = 12

Group - D

6. (a) Write down the steps of k-means algorithm. Mention how k-value is set while implementing k-means algorithm. Compare between Hierarchical clustering and k-means clustering.

[[CO2](Analyse/IOCQ)]

- (b) You are provided with coordinate value (X,Y) for 6 in the following table. What will be the updated cluster membership value after 1st iteration?

[[CO2](Evaluate/HOCQ)]

Table. 4

X	1	2	2	4	3	5
Y	1	1	3	3	2	5

7 + 5 = 12

7. (a) What is Clustering? Justify the use of proximity measures in the clustering algorithms. [[CO3](Apply/IOCQ)]
 (b) Calculate the Manhattan distance and the Chebyshev distance of two objects with coordinates (0, 3) and (5, 8). [[CO3](Evaluate/HOCQ)]
 (c) State why the validation of clustering algorithms is difficult as compared to the validation of classification algorithms. How is 'cohesion' different from 'separation' in context to cluster evaluation of clustering algorithms? [[CO3](Apply/IOCQ)]
- 4 + 4 + (2 + 2) = 12**

Group - E

8. (a) What are the basic data types available in Python. [[CO6](Remember/LOCQ)]
 (b) Explain Statements and Expressions in Python programming language with suitable examples. [[CO6](Understand/LOCQ)]
 (c) Write a short note on Scikit learn, NumPy, and Scipy. [[CO6](Remember/LOCQ)]
- 3 + 3 + 6 = 12**
9. (a) More the number of features of an instance, better is the performance of the learning system. Justify the truth/falsity of the statement. [[CO5](Analyse/HOCQ)]
 (b) What is 'Curse of Dimensionality'? How can it be overcome? [[CO5](Apply/IOCQ)]
 (c) Mention the basic idea of Principal Component Analysis (PCA) in relation to feature selection? [[CO5](Understand/LOCQ)]
- 4 + 5 + 3 = 12**

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	38.54	29.17	32.29

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. Design and test Machine Learning based case studies using Python.

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