

MACHINE LEARNING TECHNIQUES
(AEI3232)

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 one of the below statements regarding the regression line is correct?
 - (a) The prediction equation is another name for a regression line
 - (b) A regression line is also referred to as the line of the average relationship
 - (c) The estimating equation is another name for a regression line
 - (d) All of these
- (ii) Which of the following is a graphical way of showing the frequency distribution in which the height of a bar corresponds to the frequency of a category?
 - (a) Perceptual map (b) Histogram (c) Frequency chart (d) Bar chart
- (iii) Which of the following is NOT a descriptive statistic?
 - (a) t-test (b) Mean (c) Standard deviation (d) Range
- (iv) If the null hypothesis is false then which of the following is accepted?
 - (a) Null Hypothesis (b) Positive Hypothesis
 - (c) Negative Hypothesis (d) Alternative Hypothesis
- (v) The correlation coefficient describes
 - (a) Only magnitude (b) Both magnitude and direction
 - (c) Only direction (d) None of the preceding
- (vi) Which of the following is true about “Ridge” or “Lasso” regression methods in case of feature selection?
 - (a) Ridge regression uses subset selection of features
 - (b) Lasso regression uses subset selection of features
 - (c) Both use subset selection of features
 - (d) None of above
- (vii) What is the name of the procedure, where machine learning model parameters are changed to enhance its performance on training data set?
 - (a) Feature engineering (b) Model evaluation
 - (c) Model training (d) Model optimization

- (viii) How does SVM handle non-linearly separable data?
 (a) By ignoring the data (b) By using a linear kernel
 (c) By applying the kernel trick (d) By reducing the data size
- (ix) A perceptron is
 (a) A neural network that contains feedback
 (b) A single layer feed-forward neural network with pre-processing
 (c) An auto-associative neural network
 (d) A double layer auto-associative neural network
- (x) What is back propagation?
 (a) It is another name given to the curvy function in the perceptron
 (b) It is the transmission of error back through the network to adjust the inputs
 (c) It is the transmission of error back through the network to allow weights to be adjusted so that the network can learn
 (d) None of the mentioned

Fill in the blanks with the correct word

- (xi) The output of training process in machine learning is_____.
- (xii) The problem of finding hidden structure in unlabelled data is called_____ learning.
- (xiii) A statement made about a population for testing purpose is called _____.
- (xiv) Machine learning is a subset of _____.
- (xv) Naive Bayes Algorithm is a _____ machine learning algorithm.

Group - B

2. (a) State the main differences between supervised learning and reinforcement learning. Give four examples of machine learning. [[C01](Remember/LOCQ)]
- (b) Explain the concept of different skewness in data distribution with diagram. Given a left-skewed distribution that has a median of 60, what conclusions can you draw about the mean and the mode of the data? [[C01](Understand/LOCQ)]
- (c) A company wants to improve its sales. The previous sales data indicated that the average sale of 25 salesmen was Rs. 50 per transaction. After training, the recent data showed an average sale of Rs. 80 per transaction. If the standard deviation is Rs.15, find the t-score. [[C01](Analyze/IOCQ)]
- (3 + 2) + (2 + 2) + 3 = 12**
3. (a) State the advantages and disadvantages of supervised learning. [[C01](Remember/LOCQ)]
- (b) A professor wants to know if her introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score above 70 on the test. The six students get the following scores: 62, 92, 75, 68, 83 and 95. Examine whether the professor have 90% confidence that the mean score for the class on the test would be above 70. Given that for 5 degrees of freedom and 90% confidence level, the critical value of t is 2.1328. [[C01](Analyze/IOCQ)]

- (c) A machine learning model is trained to predict tumour in patients. The test data set consists of 100 patients. The model correctly predicts positive class having tumour of 10 patients and negative class not having tumour of 60 patients. It also predicts 8 patients as negative but actually having tumour and 22 patients as positive but don't have tumour. Construct the confusion matrix based on the classifier performance and calculate precision, recall and F1 score of the model. [[C01](Analyse/IOCQ)]

$$2 + 5 + 5 = 12$$

Group - C

4. (a) State and formulate multivariate linear regression problem. Derive an expression for its solution. [[C03](Analyse/IOCQ)]
- (b) Consider a linear regression model to predict target values which suffers from over fitting problem. Suggest potential strategies to overcome this problem and improve model's predictive performance. What are the difficulties it may face? [[C03](Analyse/IOCQ)]
- (c) Explain the merits and demerits of Ridge and LASSO formulation of regression problems. [[C03](Understand/LOCQ)]

$$6 + 3 + 3 = 12$$

5. (a) Define three common performance measure of a regression model. [[C03](Remember/LOCQ)]
- (b) What are the drawbacks of linear regression model? [[C03](Remember/LOCQ)]
- (c) Distinguish between subset selection and shrinkage method. [[C03](Understand/LOCQ)]
- (d) What is principal component regression (PCR)? List the step of PCR. What are the drawbacks of PCR? [[C03](Remember/LOCQ)]

$$3 + 2 + 2 + (1 + 2 + 2) = 12$$

Group - D

6. (a) Explain the process of Linear Discriminant Analysis (LDA). How does LDA differ from PCA in terms of its approach to dimensionality reduction and data classification? [[C04](Analyse/IOCQ)]
- (b) State the limitations of LDA. [[C04](Remember/LOCQ)]
- (c) Use K Means clustering to cluster the following data into two groups: { 2, 4, 10, 12, 3, 20, 30, 11, 25 }. Assume cluster centroid are $m_1=2$ and $m_2=4$. The distance function used is Euclidean distance. [[C02](Apply/HOCQ)]

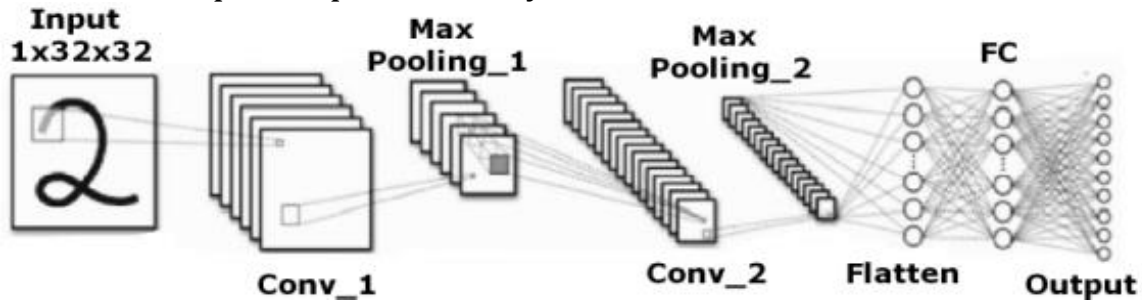
$$5 + 2 + 5 = 12$$

7. (a) State the principle of support vector machine (SVM) classifier. Hence, formulate the hard and soft margin SVM classifier problems. [[C04](Remember, Understand/LOCQ)]
- (b) What are advantages and disadvantages of decision tree? [[C04](Remember/LOCQ)]
- (c) Describe the K-means algorithm used in unsupervised learning. Explain the steps involved in the algorithm, how the number of clusters is determined, and discuss the limitations and advantages of K-means in practical applications. [[C02](Remember/LOCQ)]

$$(2 + 4) + 2 + 4 = 12$$

Group - E

8. (a) Manish wanted to design a convolutional neural network (CNN) model the classification of digits and he considered the following network architecture as shown in figure below. He decided to use 10 and 16 filters in the first and second convolutional layers, respectively, with 5×5 filters and all operations are done with stride 1 and no padding. All the pooling operations performed with 2×2 kernel with stride 2; and used 64 neurons in the fully connected layer (FC). Work out the output shape of each layer.



[[C06](Design/HOCQ)]

- (b) What will be the size of the weight matrix in between the 'flatten' and fully connected (FC) layers in the above problem?
[[C06](Analyse/IOCQ)]
- (c) Instead of taking 32×32 images, Manish now wants to train the same network to classify images of size 68×68 . List two possible architecture changes to support this.
[[C06](Design/HOCQ)]
- (d) Instead of digits, Manish now wants to be able to classify handwritten alphabetic characters (26 characters). What is the minimal change in the network architecture needed in order to support this?
[[C06](Analyse/HOCQ)]

$$6 + 2 + 2 + 2 = 12$$

9. (a) Suppose you are trying to train a neural network with two input features with a cost function $f(x_1, x_2) = x_1^2 + x_1x_2 + x_2^2$, and are trying to find a local minimum using gradient descent. You randomly start with $x_1 = 1.3$ and $x_2 = 5.4$. Perform the first step of gradient descent with learning rate $\alpha = 0.01$. Show the resulting values for x_1 and x_2 as well as all of your calculations.
[[C05](Analyse/IOCQ)]
- (b) Describe the saddle point problem in neural networks, its impact on training, and strategies to mitigate it.
[[C05](Understand/LOCQ)]
- (c) Is overfitting more or less likely when the training set is small or large? Is overfitting more or less likely when the number of parameters to learn (such as the number of weights in a neural network) is small or large?
[[C05](Apply/IOCQ)]
- (d) If we train a neural network for 1,000 epochs (one training example at a time), does it make a difference whether we present all training examples in turn for 1000 times or whether we first present the first training example 1000 times, then the second training example for 1000 times, and so on? Why?

[[C05](Analyse/HOCQ)]

$$4 + 3 + 2 + 3 = 12$$

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	44.79	36.45	18.76