

2054

M.E. (Computer Science and Engineering)

Second Semester

Elective - III

CS-8205: Machine Learning

Time allowed: 3 Hours

Max. Marks: 50

**NOTE:** Attempt five questions in all, including Question No. I which is compulsory and selecting two questions from each Unit.

- \*\_ \*\_ \*-

I. Attempt the following: -

- What is multivariate linear regression and give a real-world example of it?
- List key advantages and disadvantages of logistic regression.
- Discuss the problem of overfitting in machine learning. What effects it can have on training process?
- List the advantages and disadvantages of hierarchical clustering.
- Discuss the challenges associated with collaborative filtering.

(5×2)

**UNIT-I**

II. (a) Which is the most commonly used cost function for estimating the parameters of a linear regression model?

Give its mathematical equation. Explain how the goal of parameter learning in linear regression is to minimize this cost function using optimization techniques.

(b) How Multivariate Linear regression differs from Linear Regression? Explain the parameter optimization process and loss functions in case of multivariate Regression.

(6+4)

III. (a) Define Support Vector Machines (SVMs) and explain the concept of large margin classification. Why the training process in SVM is a slow process as compared to logistic regression?

(b) While training, how can you deduce that there is overfitting of the model? Explain how regularization techniques such as L2 (Ridge) regularization are used to address overfitting in logistic regression and SVMs. Is there any situation where we want to model to overfit?

(5+5)

IV. (a) How we can make our model work on large dataset? Explain the considerations and challenges associated with using large datasets in machine learning.

(b) You are working on a binary classification problem using logistic regression. You have a dataset with 500 samples, out of which 300 belong to class A and 200 belong to class B. After training your logistic regression model, you evaluate its performance using the following confusion matrix:

	Predicted A	Predicted B
Class A	250	50
Class B	30	170

Calculate the accuracy, precision, recall, specificity, and F1 score of the logistic regression model based on this confusion matrix. Summarize the performance of the model on the basis of results.

(4+6)

**P.T.O.**

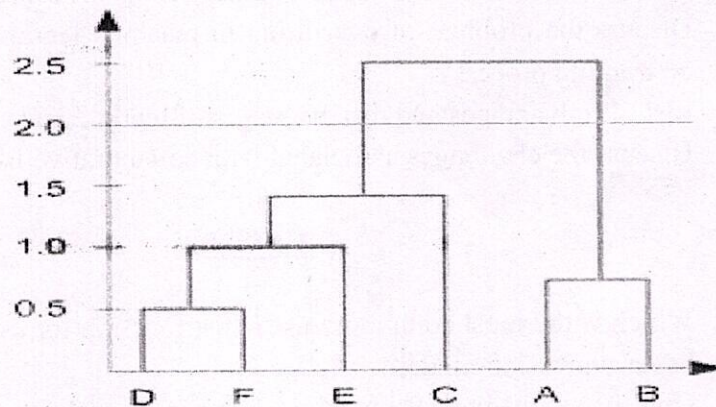
(2)

UNIT-II

- V. (a) Define DBSCAN algorithm. What are the advantages of the DBSCAN density-based Clustering Algorithm? What is the time complexity of the DBSCAN Clustering Algorithm?
- (b) Define Principal Component Analysis. How Principal Component Analysis is used for dimensionality Reduction? How is the first principal component axis selected in PCA?

(5+5)

- VI. (a) In the figure below, if you draw a horizontal line on the y-axis for  $y=2$ ,  $y=1$ . What will be the number of clusters formed? Justify your answer.



- (b) Describe the two main steps of the EM algorithm. Discuss the convergence properties of the EM algorithm. What conditions ensure convergence, and how does the algorithm handle cases where it might get stuck in local optima?

(4+6)

- VII. (a) Provide an example scenario where collaborative filtering would be beneficial. Outline the steps involved in implementing a basic collaborative filtering algorithm to recommend movies to users based on historical rating data.
- (b) Write short note on Multivariate Gaussian Distribution.

(7+3)

- \*\_ \*\_ \*