Register No.:　　.................................　　　Name:　　...

SAINTGITS COLLEGE OF ENGINEERING (AUTONOMOUS)

(AFFILIATED TO APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY, THIRUVANANTHAPURAM)

Ph.D EVEN SEMESTER EXAMINATION (R), MAY 2023

Course Code:　　**21TE204-D**

Course Name:　　**Introduction to Machine Learning**

Max. Marks:　　**60**　　　　　　　　　　　　　　　　**Duration: 3 Hours**

PART A
(Answer all questions. Each question carries 3 marks)

1. Compare unsupervised learning and reinforced learning with examples.
2. What is meant by k-fold cross-validation? Given a data set with 1000 instances, how k-fold cross validation is done with k=1000?
3. Describe the significance of Kernal functions in support vector machines. List any two kernel functions.
4. Distinguish between K-means clustering and hierarchical clustering.
5. List the issues involved in mean-shift clustering.
6. Explain the necessity of dimensionality reduction in machine learning.
7. Differentiate between PCA and ICA techniques used in dimensionality reduction.
8. Distinguish between overfitting and underfitting. How it can affect model generalization?

PART B
(Answer one full question from each module, each question carries 6 marks)

MODULE I

9. Describe the basic approaches in machine learning. Suppose you need to predict the class label or sentiment of a piece of text, like a tweet or a product review. Suggest the most appropriate learning approach for the above scenario. (6)

OR

10. Describe the following terms in the context of machine learning:
    a) Features
    b) Labels
    c) Validation sample　　　　　　　　　　　　　　　　　　　　　　(6)
    d) Loss function
    e) Hypothesis space
    f) Regression

MODULE II

11. Suppose a computer program for recognizing dogs in photographs (6)

identifies eight dogs in a picture containing 12 dogs and some cats. Of the eight dogs identified, five actually are dogs while the rest are cats. Compute the precision and recall of the computer program.

OR

12. State Bayes Theorem. It is estimated 50% of emails are spam mails. Some software has been applied to filter these spam emails before they reach your inbox. A certain brand of software claims that it can detect 99% of spam emails and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a non-spam email?   (6)

MODULE III

13. Describe the practical issues involved in decision tree learning.   (6)

OR

14. Describe the significance of soft margin hyperplane and optimal separating hyperplane and explain how they are computed.   (6)

MODULE IV

15. Identify four measures for finding the distance between numerical data points.   (6)

OR

16. With suitable equations, detail the concept of expectation maximization algorithm for two-component Gaussian mixture.   (6)

MODULE V

17. Explain the procedure for the computation of the principal components of the data.   (6)

OR

18. Describe the forward selection and backward selection algorithm for implementing the subset selection procedure for dimensionality reduction.   (6)

MODULE VI

19. Write in detail the necessity of locally linear embedding (LLE) algorithm and explain the steps in LLE.   (6)

OR

20. Explain independent component analysis in machine learning.   (6)

****