(Please write your Enrollment Number) Enrollment No.________________

# MID-TERM EXAMINATION
## (Course Name : B.Tech CSE-AI/ ECE-AI ) (Semester 5)
## (OCTOBER, 2023) OFF LINE mode, <Subjective>

| | |
|---|---|
| Subject Code: BAI-301 | Subject: Machine Learning(ML) |
| Time : 1 ½Hours | Maximum Marks : 30 |
| Note: Q. 1 is compulsory. | |

**Q1** (2.5*4)

(a) Explain difference between a hard margin and a soft margin in Support Vector Machine (SVM) .

(b)Suppose that we have already trained a classification tree, K-Nearest Neighbors, and logistic regression. When the data is relatively large (e.g., larger than one million points), which one of the techniques is slower in making a prediction?. Explain with reason.

(c) Explain the role of data splitting in the context of data preprocessing. Why is it essential to divide data into training and testing sets before model training?

(d) What role does standardization play in data preprocessing, and how does it impact the training of machine learning models?

**Q2** (Attempt any Two Parts ) UNIT-1 (5,5)

(a) What are different causes of data issues In Machine Learning? What are fallouts?

(b) Enumerate and elucidate various techniques employed in ~~data pre~~-processing, highlighting the significance of each in the context of machine learning.

(c) Describe the role of Exploratory Data Analysis (EDA) in the context of Machine Learning and explain why it is considered a crucial step in the machine learning .

**Q3** (Attempt any Two Parts ) UNIT-2 (5,5)

(a) Suppose you are building a logistic regression model to determine whether or not a person has diabetes. Predicted probabilities(P(Diabetes)) of for 10 patients are given below

| Patient | Diabetes | P(Diabetes) |
|---|---|---|
| A | Yes | 0.82 |
| B | No | 0.37 |
| C | Yes | 0.04 |
| D | No | 0.41 |
| E | Yes | 0.55 |
| F | No | 0.62 |

| | | |
|---|---|---|
| G | No | 0.20 |
| H | Yes | 0.91 |
| I | No | 0.74 |
| J | Yes | 0.34 |

Assuming that you chose a cut-off value of 0.4, wherein if a probability is greater than 0.4, you would conclude that the corresponding patient has diabetes. If it is less than or equal to 0.4, then you would conclude that they do not have diabetes.

Calculate Accuracy, Precision, Recall, Specificity and F1 Score based on the model?

(b)A dataset of patients who have tested positive or negative for COVID-19 follows. Their symptoms are cough (C), fever (F), difficulty breathing (B), and tiredness (T).

Table-1

| | Cough (C) | Fever (F) | Difficulty breathing (B) | Tiredness (T) | Diagnosis |
|---|---|---|---|---|---|
| Patient 1 | no | yes | yes | yes | Sick |
| Patient 2 | yes | yes | no | yes | Sick |
| Patient 3 | yes | no | yes | yes | Sick |
| Patient 4 | yes | yes | yes | no | Sick |
| Patient 5 | yes | no | no | yes | Healthy |
| Patient 6 | no | yes | yes | no | Healthy |
| Patient 7 | no | yes | no | no | Healthy |
| Patient 8 | no | no | no | yes | Healthy |

0-2½

Using the Gini impurity index, build a decision tree of height 1 that classifies (Diagnosis) this data. What is the accuracy of this classifier on the dataset?

(c) Consider the simplified data set given below.

| X | Y |
|---|---|
| 1 | 5 |
| 2 | 11 |
| 3 | 14 |
| 4 | 18 |

Using least square method to find the equation of the regression model fitted on this data (X independent and Y is dependent variable). Calculate R square value of this model.

9th oct
S1
320

(Please write your Enrollment Number) Enrollment No. 02601177021

**End-Term Examination**
**(CBCS)(SUBJECTIVE TYPE)(OffLine)**
**Course Name:< B.Tech CSE AI/ ECE-AI >, Semester:<5th >**
**(November-December, 2023)**

| | |
|---|---|
| **Subject Code: BAI 301** | **Subject: Machine Learning** |
| **Time : 3 Hours** | **Maximum Marks : 60** |
| **Note: Q. 1 is compulsory. Attempt one question each from the Units I, II, III & IV.** | |

| | | |
|---|---|---|
| **Q1.** | | **(2.5*8=20 )** |
| | (a) With reference to logistic regression, if you will travel along with ROC curve, how cutoff value will change? Explain? | |
| | (b) A classifier is designed to identify if patients of a clinic need to go through the rest of the diagnostic steps after the first round of testing. What classification metric would be more or less appropriate? Why? | |
| | (c) Is it a good idea to focus on hyperparameter optimization when we can also improve the quality or quantity of the training data? | |
| | (d) A decision tree classifier learned from a fixed training set achieves 100% accuracy on the test set. Which of the algorithms (Logistic regression, An SVM with a polynomial kernel, k-Nearest neighbors, Naïve Bayesclassifier ) trained using the same training set is guaranteed to give a model with 100% accuracy? Explain. | |
| | (e) What is the curse of dimensionality? | |
| | (f) If a Decision Tree is underfitting the training set, is it a good idea to try scaling the input features? Explain? | |
| | (g) A correlation between age and health of a person found to be 0.99 . On the basis of this you would tell the doctors that age causes health? Explain with reason? | |
| | (h) Suppose you are using Ridge Regression and you notice that the training error and the validation error are almost equal and fairly high. Would you say that the model suffers from high bias or high variance? Should you increase the regularization hyperparameter $\alpha$ or reduce it? Explain? | |
| | UNIT-1 | |
| **Q2.** | (a) Discuss the potential consequences of not handling outliers during data preprocessing<br>(b) How can feature engineering contribute to improving the performance of machine learning models, and what are some common techniques used in feature engineering?<br>(c) How can you use quantiles of a distribution to detect its outliers? | **(3+3+4=10)** |
| **Q3.** | (a)What are different steps in data pre-processing required before any model building? Why these steps are required?<br>(b) Explain the difference between the one-hot and label encoding methods?<br>(c) Discuss the trade-offs between removing outliers and transforming them during the data preprocessing stage | **(3+3+4=10)** |
| | UNIT-2 | |
| **Q4.** | A dataset of patients who have tested positive or negative for COVID-19 follows. Their symptoms are cough (C), fever (F), difficulty breathing (B), and tiredness (T). | **(10)** |

| | Cough (C) | Fever (F) | Difficulty breathing (B) | Tiredness (T) | Diagnosis |
|---|---|---|---|---|---|
| Patient 1 | no | yes | yes | yes | Sick |
| Patient 2 | yes | yes | no | yes | Sick |
| Patient 3 | yes | no | yes | yes | Sick |
| Patient 4 | yes | yes | yes | no | Sick |
| Patient 5 | yes | no | no | yes | Healthy |
| Patient 6 | no | yes | yes | no | Healthy |
| Patient 7 | no | yes | no | no | Healthy |
| Patient 8 | no | no | no | yes | Healthy |

Build a naive Bayes model that predicts the diagnosis from the symptoms. Use the naive Bayes algorithm to find the following probabilities:

Note: For the following questions, the symptoms that are NO mentioned are completely unknown to us. For example, if we know that the patient has a cough, but nothing is said about their fever, it does not mean the patient doesn't have a fever.

a. The probability that a patient is sick given that the patient has a cough

b. The probability that a patient is sick given that the patient is not tired

c. The probability that a patient is sick given that the patient has a cough and a fever

d. The probability that a patient is sick given that the patient has a cough and a fever, but no difficulty breathing.

**Q5.** (a)If it takes one hour to train a Decision Tree on a training set containing 1 million instances, roughly how much time will it take to train another Decision Tree on a training set containing 10 million instances? **(4+6 =10)**

(b) Consider a dataset of labeled points in a 2D space:

Dataset:

Point 1: (2, 3), Label: Class A
Point 2: (5, 8), Label: Class B
Point 3: (1, 2), Label: Class A
Point 4: (8, 8), Label: Class B
Point 5: (7, 3), Label: Class B
Point 6: (6, 4), Label: Class B
Point 7: (1, 6), Label: Class A
Point 8: (4, 7), Label: Class B

Now, suppose you have a new unlabeled point: (3, 5).

Apply the k-Nearest Neighbors (KNN) algorithm to classify the new point. Set k=3 for this problem.

Provide the classification result using Euclidean distance for the new point and discuss any considerations or insights you might have regarding the choice of k and the impact of different distance metrics.

| | UNIT-3 | |
|---|---|---|
| Q6. | Given the following set of points:<br>[(2, 3), (5, 8), (1, 2), (8, 8), (7, 3), (6, 4), (1, 6), (4, 7)]<br>Compute full iterations of k-means clustering for K=2 with initial clusters Centroid μ1 = (1, 2) and μ2 = (6, 4).<br>Make sure to write down the necessary distances (Euclidean distance), explain the steps you follow, and to describe the resulting clusters (centroid and points) at the end of each iterations. | (10) |
| Q7. | (a) What are the main motivations for reducing a dataset's dimensionality? What are the main drawbacks?<br>(b) Can PCA be used to reduce the dimensionality of a highly nonlinear dataset? Explain? | (6+4 =10) |
| | UNIT-4 | |
| Q8. | (a)A boosted strong learner L is formed by three weak learners, L1 , L2 , and L3 . Their weights are 1, 0.4, and 1.2, respectively. For a particular point, L1 and L2 predict that its label is positive, and L3 predicts that it's negative. What is the final prediction the learner L makes on this point?<br>(b) What is the difference between hard and soft voting classifiers?<br>(c) If your AdaBoost ensemble underfits the training data, which hyperparameters should you tweak and how? | (4+2+4=10) |
| Q9. | (a)If you have trained five different models on the exact same training data, and they all achieve 95% precision, is there any chance that you can combine these models to get better results? If so, how? If not, why?<br>(b) Describe Random Forest in the context of ensemble learning. How does Random Forest utilize the concept of ensemble learning, and what distinguishes it from a single Decision Tree? Discuss the benefits of employing an ensemble of trees in Random Forest? | (5+5=10) |