**Roll No:** | | | | | | | | | | | |

# PRESIDENCY UNIVERSITY

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

Presidency University Act, 2013 of the Karnataka Act No. 41 of 2013 | Established under Section 2(f) of UGC Act, 1956
Approved by AICTE, New Delhi | Approved By BCI
Bengaluru

**Mid Term Examinations - October 2026**

**Date:** 05/10/2026 **Time:** 11:45 AM - 01:15 PM

**Course Code:** IST3405 **Course Name:** Applied ML Techniques

**Semester:** Seventh Semester **Max. Marks:** 50 **Weightage:** 50%

| CO - Levels | CO1 | CO2 |
|---|---|---|
| Marks | 46 | 44 |

**PART A: Answer Following Questions. 10 M**

| Q.N. | Questions | M | CO | BT |
|---|---|---|---|---|
| 1 | Four common feature types are numerical/continuous, categorical, ordinal, and binary features. | 2 | CO1 | BT1 |
| 2 | A loss function measures the difference between predicted and actual target values. During training, model parameters are adjusted to minimize the loss. | 2 | CO1 | BT1 |
| 3 | Cross-entropy cost measures the difference between the true class distribution and the probabilities predicted by softmax regression. Minimizing it encourages high probability for the correct class and low probability for incorrect classes. | 2 | CO1 | BT2 |
| 4 | A random subspace is a randomly selected subset of features used to train an individual ensemble learner. One advantage is increased diversity among learners, which can reduce correlation and improve ensemble performance. | 2 | CO2 | BT1 |
| 5 | Stacking combines predictions from multiple base learners and feeds them to a meta-learner, which learns how to combine those predictions to produce the final output. Its main components are the base learners and the meta-learner. | 2 | CO2 | BT2 |

**PART B : Answer Any 1 Following Questions. 10 M**

| Q.N. | Questions | M | CO | BT |
|---|---|---|---|---|
| 6 | The actual values are [10, 14, 18] and the predicted values are [12, 13, 20]. Mean Absolute Error (MAE) is calculated as: MAE = (1/n) $\Sigma\|y_i - \hat{y}_i\|$ Absolute errors are: $\|10 - 12\| = 2$ $\|14 - 13\| = 1$ $\|18 - 20\| = 2$ Therefore, MAE = (2 + 1 + 2)/3 = 5/3 ≈ 1.67. Hence, the MAE is approximately 1.67. This means that, on average, the model's predictions differ from the actual target values by about 1.67 units. A smaller MAE indicates predictions that are, on average, closer to the actual values. | 10 | CO1 | BT3 |

ANSWER KEY

| | | | | |
|---|---|---|---|---|
| 7 | Mean imputation replaces a missing numerical value with the arithmetic mean of the observed values. It is simple but can be strongly affected by extreme values and skewed distributions. Median imputation replaces the missing value with the middle value of the observed data after sorting. The median is more robust to outliers and is therefore often preferred for skewed numerical features. Mode imputation replaces a missing value with the most frequently occurring value. It is particularly useful for categorical features, although it can also be used for discrete variables. For a numerical feature that is skewed and contains outliers, median imputation is recommended because the median is relatively unaffected by extreme observations. Mean imputation could shift the imputed value toward the outliers, while mode imputation may not represent the central tendency of a continuous numerical feature. Therefore, median imputation provides a more robust and representative replacement in this situation. | 10 | CO1 | BT3 |

**PART C: Answer Any 1 Following Questions. 10 M**

| Q.N. | Questions | M | CO | BT |
|---|---|---|---|---|
| 8 | The sigmoid function converts a real-valued linear score into a value between 0 and 1. It is defined as: $\sigma(z) = 1/(1 + e^{-z})$ For a binary classification model, the linear score z may be obtained from $z = w^T x + b$. The sigmoid function maps very large positive values of z close to 1 and very large negative values close to 0, so its output can be interpreted as the estimated probability of the positive class. For z = 0: $\sigma(0) = 1/(1 + e^0) = 1/(1 + 1) = 1/2 = 0.5$. Therefore, the sigmoid output is 0.5. This represents a probability of 50% for the positive class under the model. | 10 | CO1 | BT3 |
| 9 | Softmax regression extends logistic regression to multiclass classification. Suppose there are three customer classes with scores $z_1$, $z_2$ and $z_3$. The softmax probability for class k is: $P(y=k\|x) = e^{zk} / (e^{z1} + e^{z2} + e^{z3})$, for k = 1,2,3. Thus, each class score is exponentiated and divided by the sum of the exponentiated scores of all three classes. The resulting probabilities are non-negative and their sum is equal to 1. The model therefore produces a complete probability distribution over the three customer classes. For a new customer, the three probabilities are calculated and compared. The class having the largest probability is selected as the predicted class. For example, if the probabilities were 0.20, 0.55 and 0.25 for Classes 1, 2 and 3 respectively, Class 2 would be predicted because it has the highest probability. | 10 | CO1 | BT2 |

**PART D: Answer Any 1 Following Questions. 10 M**

| Q.N. | Questions | M | CO | BT |
|---|---|---|---|---|
| 10 | Bagging, or Bootstrap Aggregating, creates several training datasets by drawing bootstrap samples from the original training data with replacement. A separate base learner is trained on each bootstrap sample. After all learners are trained, their predictions are combined: majority voting is normally used for classification, while averaging is used for regression. Bagging is especially useful for high-variance learners such as decision trees. A high-variance learner can change considerably when the training data changes slightly. In bagging, different bootstrap samples produce different learners. These learners make different errors, and aggregation averages or votes across their predictions. As a result, the effect of any one unstable learner is reduced. The ensemble therefore has lower variance and more stable predictions than a single high-variance learner. Bagging does not primarily work by making each individual tree less complex; instead, it reduces the variability of the overall prediction through averaging across multiple models. | 10 | CO2 | BT3 |

ANSWER KEY

| | | | | |
|---|---|---|---|---|
| 11 | Random subspaces are an ensemble technique in which different learners are trained using different randomly selected subsets of the available features. Suppose a dataset has many features. Instead of allowing every learner to use all features, a different subset is selected for each learner. Consequently, individual learners are exposed to different information and may learn different decision boundaries or patterns. This reduces the correlation between learners. When their predictions are combined, the errors made by one learner may be compensated for by the predictions of other learners. Therefore, the ensemble becomes more diverse. Higher diversity with sufficiently accurate component learners is useful because aggregation benefits from less-correlated errors. Random subspaces are particularly useful when the dataset contains many features or when some features dominate the learning process. By forcing learners to consider different feature combinations, the technique can improve robustness and generalization of the ensemble. | 10 | CO2 | BT3 |

| **PART E: Answer Any 1 Following Questions. 10 M** | | | | |
|---|---|---|---|---|
| **Q.N.** | **Questions** | **M** | **CO** | **BT** |
| 12 | AdaBoost builds an ensemble sequentially. Initially, all training observations receive equal weights. A weak learner is trained using these weighted observations. After the learner makes predictions, the observations that are misclassified receive greater importance, while correctly classified observations receive relatively less importance. The next weak learner therefore concentrates more on observations that were difficult for previous learners. This process is repeated for several rounds, and the final classifier combines the weak learners using weights related to their performance. AdaBoost can be sensitive to noise because noisy observations, outliers, or incorrectly labelled examples may be repeatedly misclassified. Since AdaBoost increases the weights of misclassified observations, such problematic observations can receive disproportionately high weights. Later learners may then focus strongly on fitting these noisy cases instead of learning the general underlying pattern. This can reduce generalization performance. Hence, AdaBoost is powerful for emphasizing difficult examples, but excessive emphasis on noisy or mislabeled data can make the method sensitive to noise. | 10 | CO2 | BT3 |
| 13 | Stacking is an ensemble technique that combines several different base learners through a meta-learner. For example, three base learners may be a decision tree, logistic regression model, and support vector machine. Each base learner is trained to make predictions from the original input features. Their predictions are then used as inputs to a second-level model called the meta-learner. The meta-learner learns how to combine the outputs of the base learners to produce the final prediction. In a properly designed stacking system, predictions used to train the meta-learner are preferably generated using validation or out-of-fold predictions so that the meta-learner does not simply learn from overly optimistic predictions on the same data used to train the base models. Stacking can improve prediction when the base learners have complementary strengths and make different types of errors. The meta-learner can learn which learner or combination of learners is more useful for different patterns in the data. Therefore, instead of relying on a single modelling assumption, stacking can exploit the strengths of several diverse learning algorithms and potentially improve generalization. | 10 | CO2 | BT3 |

ANSWER KEY