ASSIGNEMENT -5 (MACHINE LEARNING)

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

Ans. To compare models with varying numbers of independent variables, adjusted R-square should be utilized. When choosing significant predictors (independent variables) for the regression model, adjusted R-square should be utilized.

2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other?

Ans. Total sum of squares (TSS) examines the amount of variation in the observed data, whereas residual sum of squares assesses the variance in the error between the observed data and modeled values.

3. What is the need of regularization in machine learning?

Ans. The term "regularization" describes methods for calibrating machine learning models to reduce the adjusted loss function and avoid overfitting or underfitting. We can properly fit our machine learning model on a particular test set using regularization, which lowers the mistakes in the test set.

4. What is Gini–impurity index?

Ans. To identify how the attributes of a dataset should split the nodes to construct the tree, decision trees are built using the Gini Impurity measurement.

5. Are unregularized decision-trees prone to overfitting? If yes, why?

Ans. For instance, you can assess the correctness of the training and validation sets or test sets. The degree of overfitting increases with the size of the gap. When developing decision trees, the following pattern is typically observed

6. What is an ensemble technique in machine learning?

Ans. By mixing numerous models rather of relying just on one, ensemble approaches seek to increase the accuracy of findings in models. The integrated models considerably improve the findings' accuracy. Due of this, ensemble approaches in machine learning have gained prominence.

ENSEMBLE METHODS: (i) BAGGING (II) BOOSTING

7. What is the difference between Bagging and Boosting techniques?

Ans. By using repetitions and combinations to construct several sets of the same data, bagging is a strategy for minimizing prediction variance that produces additional data for training from a dataset. Boosting is an iterative approach for modifying the weight of an observation depending on the classification made before.

8. What is out-of-bag error in random forests?

Ans. The average error for each derived using predictions from the trees that do not contain in their respective bootstrap sample is known as the out-of-bag (OOB) error. This enables fitting and validating the Random Forest Classifier as it is being trained.

9. What is K-fold cross-validation?

Ans. A resampling technique called cross-validation is used to assess machine learning models on a small data sample. The process contains a single parameter, k, that designates how many groups should be created from a given data sample. As a result, the method is frequently referred to as k-fold cross-validation.

10. What is hyper parameter tuning in machine learning and why it is done?

Ans. Finding a set of ideal hyperparameter values for a learning algorithm and using this tuned algorithm on any data set is hyperparameter tuning. By minimizing a predetermined loss function, that set of hyperparameters enhances the model's performance and yields better outcomes with lower error.

11. What issues can occur if we have a large learning rate in Gradient Descent?

Ans. We must choose an adequate value for the learning rate for Gradient Descent to function. This parameter controls how quickly or slowly we will approach the ideal weights. If the learning rate is exceptionally high, the ideal solution will be skipped.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

Ans. When the classes can be distinguished in the feature space by linear bounds, logistic regression is typically utilized as a linear classifier. However, that can be fixed if we have a better understanding of the decision boundary's form.

13. Differentiate between Adaboost and Gradient Boosting?

Ans. The first boosting algorithm with a specific loss function was called AdaBoost. Gradient Boosting, on the other hand, is a general technique that helps in the search for approximations to the additive modeling issue. Gradient Boosting is hence more adaptable than AdaBoost.

14. What is bias-variance trade off in machine learning?

Ans. The bias-variance tradeoff is a characteristic of a model in statistics and machine learning that allows the variance of the parameter estimated across samples to be lowered by increasing the bias in the estimated parameters.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Ans. Linear: The non-linearly separable data may be projected into higher dimensional space using RBF, which is more complex and effective at the same time since it can combine several polynomial kernels numerous times of varying degrees.

Polynomials kernels : The polynomial kernel in machine learning reflects the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing the learning of non-linear models. It is frequently used with support vector machines (SVMs) and other kernelized models.

RBF Kernels : Due of how much the RBF Kernel resembles the K-Nearest Neighborhood Algorithm, it is widely used. Due to the fact that RBF Kernel Support Vector Machines only need to store the support vectors during training and not the complete dataset, it has the benefits of K-NN and solves the space complexity problem.