Machine Learning Assignment

1. R-squared vs. RSS -

- R-squared is a better measure of goodness of fit because it indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. Higher R-squared values represent better fit.

- RSS measures the total squared deviation of the predicted values from the actual values. Lower RSS indicates a better fit, but it does not provide a normalized measure of fit like R-squared.

2. TSS, ESS, and RSS -

- TSS (Total Sum of Squares) : Measures the total variance in the observed data.

- ESS (Explained Sum of Squares) : Measures the variance explained by the regression model.

- RSS (Residual Sum of Squares): Measures the variance not explained by the model.

- Equation : (TSS = ESS + RSS).

3. Need for Regularization -

- Regularization is needed to prevent overfitting by penalizing large coefficients in the model. It helps improve the generalization of the model to unseen data.

4. Gini-impurity Index -

- A measure of how often a randomly chosen element would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the dataset. It is used in decision trees to decide the best split.

5. Unregularized Decision Trees and Overfitting -

- Yes, unregularized decision trees are prone to overfitting because they can create very complex trees that perfectly fit the training data, including noise, leading to poor generalization.

6. Ensemble Technique -

- A method in machine learning where multiple models (usually of the same type) are trained and combined to solve a problem, often leading to better performance than individual models.

7. Bagging vs. Boosting -

- Bagging (Bootstrap Aggregating) : Involves training multiple models independently on different subsets of the data and averaging their predictions.

- Boosting : Involves training models sequentially, where each model tries to correct the errors of the previous one.

8. Out-of-Bag Error in Random Forests -

- An estimate of the prediction error for a random forest model. It is calculated using the data that was not used during the bootstrapping process for training each tree.

9. K-fold Cross-validation -

- A technique for evaluating a model's performance by dividing the dataset into K equally sized folds. The model is trained on K-1 folds and tested on the remaining fold. This process is repeated K times, and the results are averaged.

10. Hyperparameter Tuning -

- The process of selecting the best hyperparameters for a machine learning model to optimize its performance. It is done to enhance model accuracy and prevent overfitting or underfitting.

11. Issues with Large Learning Rate in Gradient Descent -

- A large learning rate can cause the algorithm to overshoot the minimum of the cost function, leading to divergence or oscillation around the minimum, resulting in poor convergence.

12. Logistic Regression for Non-Linear Data -

- Logistic Regression inherently models linear decision boundaries. It cannot handle non-linear data unless transformed features or kernel methods are used to introduce non-linearity.

13. Adaboost vs. Gradient Boosting -

- Adaboost : Focuses on adjusting the weights of incorrectly classified instances and uses the same weak learner iteratively.

- Gradient Boosting : Optimizes the model by sequentially fitting new models to minimize the residual errors of previous models using gradient descent.

14. Bias-Variance Trade-off -

- The trade-off between the error due to bias (error from incorrect assumptions in the learning algorithm) and variance (error from sensitivity to small fluctuations in the training set). Finding the right balance minimizes the total error.

15. Kernels in SVM -

- Linear Kernel : Computes the linear separation between data points.

- RBF (Radial Basis Function) Kernel : Measures similarity based on the distance from a central point, effective for non-linear data.

- Polynomial Kernel : Computes non-linear boundaries by taking polynomial combinations of the input features, useful for capturing complex relationships.