Machine Learning Assignment - 5

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?

R-squared is generally considered a better measure of goodness of fit because it provides a proportion of the variance in the dependent variable that is predictable from the independent variables. RSS, on the other hand, only measures the total deviation of the predicted values from the actual values.

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.

- TSS (Total Sum of Squares): Total variation in the dependent variable.
- ESS (Explained Sum of Squares): Variation explained by the regression model.
- RSS (Residual Sum of Squares): Variation not explained by the model.

Equation: TSS = ESS + RSS

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

Regularization is needed to prevent overfitting by adding a penalty to the loss function, which encourages the model to have simpler weights.

4. What is Gini–impurity index?

The Gini impurity index measures the frequency at which a randomly chosen element would be incorrectly classified.

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

Yes, unregularized decision trees are prone to overfitting because they can create overly complex trees that fit the training data very closely but fail to generalize to new data.

6. What is an ensemble technique in machine learning?

An ensemble technique combines multiple models to produce a better performance than a single model.

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

- Bagging: Combines multiple models to reduce variance.

- Boosting: Combines models sequentially to reduce bias.

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

Out-of-bag error is the average prediction error on each training sample using only the trees that did not have the sample in their bootstrap sample.

9. What is K-fold cross-validation?

K-fold cross-validation is a technique to assess the performance of a model by dividing the dataset into K subsets and training K times, each time using a different subset as the test set and the remaining as the training set.

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

Hyperparameter tuning is the process of finding the optimal set of hyperparameters for a learning algorithm to improve its performance.

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

A large learning rate can cause the model to converge too quickly to a suboptimal solution or even diverge, missing the optimal point altogether.

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

Logistic Regression cannot directly handle non-linear data, but it can be extended using techniques like polynomial features or kernel methods to capture non-linearity.

13. Differentiate between Adaboost and Gradient Boosting.

- Adaboost: Focuses on misclassified instances, adjusting the weights of classifiers.
- Gradient Boosting: Optimizes the loss function by fitting new models to the residuals of previous models.

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

The bias-variance tradeoff refers to the balance between a model's ability to minimize bias (error from erroneous assumptions) and variance (error from sensitivity to fluctuations in the training set).

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

- Linear Kernel: Computes the dot product between two vectors.
- RBF Kernel: Measures similarity based on the distance between points.
- Polynomial Kernel: Represents the similarity of vectors in a polynomial space.