MACHINE LEARNING ASSIGNMENT – 5

Q1 to Q15 are subjective answer type questions, Answer them briefly.

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 - R-squared is a better measure of goodness of fit in regression because it represents the proportion of variance in the dependent variable explained by the independent variables in the model.

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 - TSS (Total Sum of Squares) represents the total variance in the dependent variable, ESS (Explained Sum of Squares) represents the variance explained by the regression model, and RSS (Residual Sum of Squares) represents the unexplained variance. They are related by the equation: TSS = ESS + RSS.

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

Ans - Regularization in machine learning is needed to prevent overfitting and improve the generalization ability of models by penalizing complex models.

4. What is Gini–impurity index?

Ans - Gini impurity index is a measure of impurity or disorder used in decision tree algorithms to determine the split at each node.

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

Ans - Yes, unregularized decision trees are prone to overfitting because they can keep splitting nodes until they perfectly fit the training data.

6. What is an ensemble technique in machine learning?

Ans - Ensemble techniques in machine learning combine multiple models to improve performance, such as bagging, boosting, and stacking.

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

Ans - Bagging builds multiple models independently and combines their predictions, while boosting sequentially builds models, each correcting the errors of the previous ones.

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

Ans - Out-of-bag error in random forests is the error rate of the model on the samples not included in the bootstrap sample used to train each tree.

9. What is K-fold cross-validation?

Ans - K-fold cross-validation is a technique used to evaluate model performance by dividing the dataset into k subsets and iteratively training the model on k-1 subsets and testing it on the remaining subset.

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

Ans - Hyperparameter tuning in machine learning involves selecting the optimal values for parameters that are not learned during training to improve model performance.

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

Ans - Large learning rates in gradient descent can lead to overshooting the minimum and divergence, causing the algorithm to fail to converge or converge slowly.

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

Logistic Regression can handle linearly separable data but may not perform well on nonlinear data without appropriate feature transformations or kernel tricks.

13. Differentiate between Adaboost and Gradient Boosting.

Ans - Adaboost focuses on adjusting the weights of observations based on the previous classification, while Gradient Boosting builds trees sequentially, with each tree learning from the mistakes of the previous ones.

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

Bias-variance tradeoff refers to the balance between bias (error from overly simplistic models) and variance (error from overly complex models) in machine learning models.

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

- Ans - Linear kernel computes the dot product of feature vectors.
- RBF (Radial Basis Function) kernel uses a Gaussian function to map samples into higher-dimensional space.
- Polynomial kernel computes the polynomial terms of the original features.