MACHINE LEARNING ASSIGNMENT

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 because it gives you a simple, clear way to see how well your model fits the data, while RSS is harder to interpret and compare.

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 represents the total variability in the dependent variable (the data you're trying to predict). It measures how much the actual data points differ from their overall mean

ESS represents the portion of TSS that is explained by the regression model. It measures how much of the variation in the data is captured by the model's predictions.

RSS represents the portion of TSS that is not explained by the regression model. It measures the leftover error between the actual data points and the model's predictions.

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

Ans. Regularization is needed in machine learning to:

- Prevent overfitting by penalizing complexity.
- Ensure the model generalizes well to unseen data.
- Handle high-dimensional datasets by simplifying the model.
- Improve model stability and reliability, especially in the presence of multicollinearity.

4. What is Gini–impurity index?

Ans. **Gini Impurity** is a metric used to measure how often a randomly chosen element from a set would be incorrectly classified. It is commonly used in decision trees (like in **classification tasks**) to decide how to split data at each node. The lower the Gini Impurity, the purer (more homogeneous) the node is.

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

Ans. Unregularized decision trees are prone to overfitting because they can grow too deep and fit even random noise in the data. Regularization techniques, like limiting the tree's depth or pruning, are essential to prevent the tree from becoming too complex and help it generalize better to unseen data.

6. What is an ensemble technique in machine learning?

Ans**.** Ensemble techniques in machine learning refer to methods that combine the predictions of multiple individual models (called "base models" or "weak learners") to create a more accurate and robust model. The primary goal of ensemble methods is to improve the overall predictive performance by leveraging the strengths of different models and reducing the weaknesses of any single model.

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

Ans. Bagging reduces variance by averaging predictions from independent models trained on different data subsets.

Boosting reduces both bias and variance by sequentially training models, focusing on correcting errors from previous models.

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

Ans. Out-of-bag (OOB) error is a concept used in Random Forests to estimate the model's performance without needing a separate validation dataset.

9. What is K-fold cross-validation?

Ans. K-fold cross-validation is a technique used to evaluate the performance of a machine learning model by dividing the dataset into several subsets.

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

Ans. Hyperparameter tuning is the process of optimizing the hyperparameters of a machine learning model to improve its performance on a specific task. Hyperparameters are the configuration settings used to control the learning process but are not learned from the data during training.

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

Ans. Using large learning rate in gradient descent can hinder the training process by causing divergence, oscillation, and poor generalization, making it crucial to choose an appropriate learning rate for effective model training.

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

Ans. Logistic Regression is not suitable for directly classifying non-linear data due to its linear nature, you can either transform the data through feature engineering or opt for other machine learning algorithms that can effectively capture non-linear relationships.

13. Differentiate between Adaboost and Gradient Boosting.

Ans. Both AdaBoost and Gradient Boosting are powerful boosting techniques, they differ in their approach to model updates, loss functions, sensitivity to noise, and overall performance characteristics. AdaBoost focuses on reweighting misclassified instances, while Gradient Boosting aims to optimize the residual errors of the ensemble.

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

Ans. The bias-variance trade-off is a fundamental concept in machine learning that describes the balance between two sources of error that affect the performance of predictive models: bias and variance.

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

Ans. Support Vector Machines (SVM) use various kernel functions to enable the model to classify data in higher-dimensional spaces. Short descriptions of three commonly used kernels: Linear, RBF (Radial Basis Function), and Polynomial.

1. The linear kernel is the simplest kernel function. It computes the dot product of the input features directly. It is used when the data is linearly separable.

2. The RBF kernel is a popular choice for non-linear data. It maps the input space into an infinite-dimensional space, allowing for more complex decision boundaries.

3. The polynomial kernel computes the dot product of the input features raised to a specified power, allowing for a polynomial decision boundary.