



## **MACHINE LEARNING**

**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. Because it accounts for the number of predictors in the model, fostering comparison between models with different numbers of independent variables without being biased towards models.

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

Ans. They are important concepts to evaluate the goodness of fit of the regression model. These metrics are related to each other through the following equation:

$$TSS=ESS+RSS$$

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

Ans- It is a technique to prevent overfitting and improve the generalization performance of models.

4. What is Gini impurity index?

Ans. Gini impurity is a measure of how often a randomly chosen element from a dataset would be incorrectly labeled.

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

Ans. Yes, Due to High Model Complexity, Sensitive to Small Variations, No Constraints on Tree Growth.

6. What is an ensemble technique in machine learning?

Ans. It combines multiple models to improve accuracy and resilience in forecasting.

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

Ans. Bagging reduces variance by averaging predictions from models trained on different subsets of data. Boosting reduces bias by sequentially training models that focus on errors of previous models

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

Ans- It is a method for estimating the performance of the model without the need for a separate validation set.

9. What is K fold cross validation?

Ans- It is a machine learning technique that evaluates the performance of a model by dividing a dataset into k subsets, or folds, of roughly equal size.

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

Ans- Hyperparameter tuning is a key part of machine learning that involves finding the best hyperparameter values for a learning algorithm. These values maximize the model's performance and minimize a predefined loss function.

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

Ans- A large learning rate in gradient descent can cause a number of issues, including:

Overshooting: The algorithm may overshoot the minimum

Oscillating performance: Performance may oscillate over training epochs

Lower final performance: The final performance may be lower

Divergence: The loss function may diverge

Exploding gradients: Weights may increase exponentially, causing exploding gradients

Instabilities: Instabilities may occur

Overly high loss values: Loss values may be overly high

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

Ans- No, logistic regression cannot be used for classifying non-linear data because it assumes a linear relationship between the input features and the output.

13. Differentiate between Adaboost and Gradient Boosting.

Ans- AdaBoost (Adaptive Boosting) and Gradient Boosting are both machine learning ensemble learning techniques that improve prediction accuracy. However, they differ in how they approach building the ensemble:

**AdaBoost**

Combines multiple simple models into a strong one. It uses the stagewise addition method to combine multiple weak learners into strong learners. AdaBoost trains trees sequentially, with each new tree attempting to correct errors from the previous one. AdaBoost is generally slower than Gradient Boosting because it requires multiple iterations to build the sequence of models.

**Gradient Boosting**

Constructs models in a series, with each new model specifically targeting the errors of its predecessor. Gradient Boosting starts by making a single leaf, which represents an initial guess for weights of all the samples. It then builds a tree based on the initial guess. Gradient Boosting excels in managing diverse datasets, particularly those with non-linear patterns.

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

Ans- Bias-variance tradeoff is a fundamental concept in machine learning that describes the balance between a model's bias and variance. Bias is the model's simplifying assumptions to make the target function easier to approximate, while variance is how much the estimate of the target function changes when using different training data.

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

Ans-

**Linear Kernel:**

The linear kernel is the simplest kernel function used in SVM.

It computes the dot product between the feature vectors in the original feature space.

The linear kernel is often used as a baseline or when the data is linearly separable.

**Radial Basis Function (RBF) Kernel:**

The RBF kernel, also known as the Gaussian kernel, is a popular choice for non-linear classification tasks.

It maps the input features into a higher-dimensional space using a non-linear transformation.

The RBF kernel computes the similarity (or distance) between data points in the transformed space using the Gaussian similarity measure.

**Polynomial Kernel:**

The polynomial kernel is another non-linear kernel used in SVM.

It computes the dot product between the feature vectors in a higher-dimensional space using a polynomial function.

The polynomial kernel allows the SVM to capture non-linear decision boundaries by computing polynomial combinations of the input features.