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?

   Answer :- R-squared is a better measure of goodness of fit in regression as it provides a proportion of the total variation in the response variable that is explained by the predictors. Residual Sum of Squares (RSS) on the other hand, only gives the magnitude of the error in the predictions.

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.

   Answer :- TSS (Total Sum of Squares) is the sum of the squares of the differences between the actual response variable values and the mean of the response variable. ESS (Explained Sum of Squares) is the sum of the squares of the differences between the predicted response variable values and the mean of the response variable. RSS (Residual Sum of Squares) is the sum of the squares of the differences between the actual and predicted response variable values. These three metrics are related by the equation: TSS = ESS + RSS.

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

   Answer :- Regularization is used in machine learning to prevent overfitting of models to the training data by adding a penalty term to the loss function that discourages models from having too many parameters or coefficients.

4. What is Gini–impurity index?

   Answer :- Gini-impurity index is a measure of the impurity or disorder in a set of observations. It is used as a criterion for deciding how to split a node in decision tree algorithms.

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

   Answer :- Yes, unregularized decision trees are prone to overfitting due to their tendency to grow too deep and have too many branches. This makes the trees highly sensitive to the training data, leading to poor performance on new data.

6. What is an ensemble technique in machine learning?
   Answer :- An ensemble technique in machine learning is a method of combining multiple models to improve the performance of the overall system.

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

   Answer :- Bagging (Bootstrap Aggregation) is an ensemble technique that involves training multiple models independently on random samples of the training data and combining their predictions through a majority voting or averaging process. Boosting, on the other hand, trains models sequentially, giving more weight to samples that were misclassified by previous models in the sequence.

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

   Answer :- Out-of-bag error in random forests is the average error of predictions made by individual trees on samples not included in their bootstrapped training sets. It provides a measure of the generalization performance of the model without the need for cross-validation.

9. What is K-fold cross-validation?

   Answer :- K-fold cross-validation is a technique for evaluating the performance of a model by dividing the data into k folds and using k-1 folds for training and the remaining fold for testing. This process is repeated k times with each fold used as the test set once. The average performance across all k iterations is used as the overall performance metric.

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

    Answer :- Hyperparameter tuning in machine learning involves adjusting the parameters of a model that are not learned from the data, such as the regularization strength or learning rate. This is done to improve the performance of the model on new data by finding the optimal combination of hyperparameters.

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

    Answer :- If we have a large learning rate in gradient descent, the optimization process may converge too quickly to a sub-optimal solution, oscillate or diverge instead of converging.

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

    Answer :- No, Logistic Regression is limited to linear classification problems and cannot be used for non-linear data. For non-linear data, non-linear models such as Support Vector Machines (SVMs) with non-linear kernels or decision trees/random forests are more suitable.

13. Differentiate between Adaboost and Gradient Boosting.

    Answer :- Adaboost is a boosting technique that trains weak classifiers (such as decision trees with a small number of splits) sequentially, giving more weight to samples that were misclassified by previous classifiers. Gradient Boosting, on the other hand, uses gradient descent to optimize the loss function of a set of weak

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

    Answer :- Bias-variance tradeoff in machine learning refers to the balance between the error introduced by approximating the underlying relationship between inputs and outputs (bias) and the error due to the inherent noise and variability in the data (variance). A model with high bias will have high error in estimating the underlying relationship, leading to underfitting, while a model with high variance will have high error due to its sensitivity to the noise in the data, leading to overfitting. The goal is to find a model that has the right balance between bias and variance.

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

Answer :- Linear Kernel: A linear kernel in Support Vector Machines (SVMs) is used for linear classification problems where the decision boundary is a straight line.

RBF (Radial Basis Function) Kernel: An RBF kernel is used for non-linear classification problems where the decision boundary is a curve. The RBF kernel maps the input data into a high-dimensional feature space where a linear boundary can separate the classes.

Polynomial Kernel: A polynomial kernel is used for polynomial classification problems where the decision boundary is a polynomial curve. It transforms the input data into a higher dimensional feature space, where a linear boundary can separate the classes. The degree of the polynomial can be specified as a parameter.