MACHINE LEARNING

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 and Residual Sum of Squares (RSS) are both measures of the goodness of fit of a regression model, but they serve slightly different purposes.

R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. In other words, R-squared measures the extent to which changes in the dependent variable can be predicted by changes in the independent variable(s). Higher R-squared values indicate a better fit of the regression model to the data. Therefore, R-squared is often used to compare different models and select the best one.

On the other hand, Residual Sum of Squares (RSS) measures the difference between the observed values of the dependent variable and the predicted values by the model. It represents the sum of the squared differences between the actual and predicted values of the dependent variable. The goal is to minimize the residual sum of squares to obtain a better model fit.

In terms of determining the goodness of fit of a model, R-squared is generally considered a better measure than RSS. This is because R-squared provides an overall measure of the proportion of variance in the dependent variable that is explained by the model, whereas RSS only measures the magnitude of the residuals. Additionally, R-squared is a standardized measure and ranges from 0 to 1, making it easy to compare the fit of different models. In contrast, the magnitude of the RSS value depends on the scale of the dependent variable and can't be easily compared across models.

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 is the sum of square of difference of each data point from the mean value of all the values of target variable.

ESS is the sum of the differences between the predicted value and the mean of the dependent variable.

RSS measures the level of variance in the error term, or residuals, of a regression model.

TSS = ESS + RSS, where TSS is Total Sum of Squares, ESS is Explained Sum of Squares and RSS is Residual Sum of Suqares. The aim of Regression Analysis is explain the variation of dependent variable Y.

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

Ans- It is a technique to prevent the model from overfitting by adding extra information to it.

Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique.

This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

4. What is Gini–impurity index?

Ans- Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree. More precisely, the Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

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

Ans- Yes, reason is Complexity: Decision trees become overly complex, fitting training data perfectly but struggling to generalize to new data. Memorizing Noise: It can focus too much on specific data points or noise in the training data, hindering generalization.

6. What is an ensemble technique in machine learning?

Ans- Ensemble learning refers to a machine learning approach where several models are trained to address a common problem, and their predictions are combined to enhance the overall performance.

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

Ans- Bagging is a learning approach that aids in enhancing the performance, execution, and precision of machine learning algorithms. Boosting is an approach that iteratively modifies the weight of observation based on the last classification. 2. It is the easiest method of merging predictions that belong to the same type.

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

Ans- Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging).

9. What is K-fold cross-validation?

Ans- K-fold cross validation in machine learning cross-validation is a powerful technique for evaluating predictive models in data science. It involves splitting the dataset into k subsets or folds, where each fold is used as the validation set in turn while the remaining k-1 folds are used for training.

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

Ans- Hyperparameters directly control model structure, function, and performance. Hyperparameter tuning allows data scientists to tweak model performance for optimal results. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success.

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

Ans- Overfitting: Gradient descent can overfit the training data if the model is too complex or the learning rate is too high. This can lead to poor generalization performance on new data.

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

Ans-

13. Differentiate between Adaboost and Gradient Boosting.

Ans-

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

Ans- The bias-variance tradeoff is about finding the right balance between simplicity and complexity in a machine learning model.

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

Ans-