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?

Answer: 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.

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) is the sum of squared differences between the observed dependent variables and the overall mean.

(ESS) is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, in a standard regression model

(RSS) is a statistical method that calculates the variance between two variables that a regression model doesn't explain. The higher the RSS, the worse a model is performing

TSS=ESS+RSS

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

Answer: 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.

4. What is Gini–impurity index?

Answer: Gini Index is a powerful measure of the randomness or the impurity or entropy in the values of a dataset. Gini Index aims to decrease the impurities from the root to the leaf of a decision tree model.

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

Answer: Overfitting in decision tree models occurs when the tree becomes too complex and captures noise or random fluctuations in the training data, rather than learning the underlying patterns that generalize well to unseen data. Other reasons for overfitting include:

1. Complexity: Decision trees become overly complex, fitting training data perfectly but struggling to generalize to new data.
2. Memorizing Noise: It can focus too much on specific data points or noise in the training data, hindering generalization.
3. Overly Specific Rules: Might create rules that are too specific to the training data, leading to poor performance on new data.
4. Feature Importance Bias: Certain features may be given too much importance by decision trees, even if they are irrelevant, contributing to overfitting.
5. Sample Bias: If the training dataset is not representative, decision trees may overfit to the training data's idiosyncrasies, resulting in poor generalization.
6. Lack of Early Stopping: Without proper stopping rules, decision trees may grow excessively, perfectly fitting the training data but failing to generalize well.

6. What is an ensemble technique in machine learning?

MACHINE LEARNING

Answer: 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?

Answer: Bagging is the simplest way of combining predictions that belong to the same type while Boosting is a way of combining predictions that belong to the different types. Bagging aims to decrease variance, not bias while boosting aims to decrease bias, not variance.

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

Answer: 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). Bagging uses subsampling with replacement to create training samples for the model to learn from.

9. What is K-fold cross-validation?

Answer: In K-fold cross-validation, the data set is divided into a number of K-folds and used to assess the model's ability as new data become available. K represents the number of groups into which the data sample is divided.

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

Answer: Hyperparameter tuning is the process of selecting the optimal values for a machine learning model's hyperparameters. Hyperparameters are settings that control the learning process of the model, such as the learning rate, the number of neurons in a neural network, or the kernel size in a support vector machine. The goal of hyperparameter tuning is to find the values that lead to the best performance on a given task.

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

Answer: Gradient descent can overfit the training data if the model is too complex or the learning rate is too high.

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

Answer: Logistic regression is a simple and more efficient method for binary and linear classification problems. Drawback of logistic regression is that it assumes a linear relationship between the input features and the output. This means that it cannot capture the complexity and non-linearity of the data.

13. Differentiate between Adaboost and Gradient Boosting.

Answer: AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

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

Answer: The bias-variance trade off is about finding the right balance between simplicity and complexity in a machine learning model. High bias means the model is too simple and consistently misses the target, while high variance means the model is too complex and shoots all over the place.

MACHINE LEARNING

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

Answer: The linear kernel produces a decision boundary that is a hyperplane in the feature space. This hyperplane separates data points from different classes in a linear fashion. Assumption: It assumes that the relationship between the features and the target variable is linear.

the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.