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 a better measure of goodness of fit model in regression because it measures the proportion of variance in the dependent variable that is explained by the independent variables. RSS, on the other hand, measures the difference between the observed values and the predicted values. R-squared is a more comprehensive measure of how well a model fits the data.

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): TSS is a measure of the total variance in a data set. It is calculated by summing the squared differences between each data point and the mean of the data set.

ESS (Explained Sum of Squares): ESS is a measure of the amount of variance in a data set that can be explained by a regression model. It is calculated by summing the squared differences between each data point and its predicted value from the regression model.

RSS (Residual Sum of Squares): RSS is a measure of the amount of variance in a data set that cannot be explained by a regression model. It is calculated by summing the squared differences between each data point and its actual value.

The equation relating these three metrics is: TSS = ESS + RSS

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

Answer: Regularization is a technique used to prevent overfitting in machine learning models. It adds a penalty to the loss function of the model which makes the model more robust and prevents it from memorizing the training data. Regularization helps to reduce the complexity of the model and makes it more generalizable to unseen data.

4. What is Gini–impurity index?

Answer: The Gini–impurity index is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the set. It is used in decision tree algorithms to determine the best split point within a given set of data. The Gini index is calculated by subtracting the sum of the squared probabilities of each class from one. A lower Gini index indicates a better split.

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

Answer: Yes, unregularized decision trees are prone to overfitting. This is because they are not limited in the number of splits they can make, so they can continue to split until each leaf node contains only one data point. This can lead to a model that is overly complex and does not generalize well to unseen data.

6. What is an ensemble technique in machine learning?

Answer: Ensemble techniques in machine learning are methods that combine multiple models to create a more powerful and accurate model. These techniques are used to improve the performance of a single model by combining the predictions of multiple models. Examples of ensemble techniques include bagging, boosting, stacking, and blending.

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

Answer: Bagging and Boosting are both ensemble methods used to improve the accuracy of machine learning models. Bagging is a technique that combines multiple models to reduce the variance of a single model. It works by training multiple models on different subsets of the data and then combining the results. Boosting is a technique that combines multiple models to reduce the bias of a single model. It works by training multiple models sequentially, each one focusing on the errors made by the previous model. The final result

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

Answer: Out-of-bag (OOB) error is an estimate of the generalization error of a random forest model. It is calculated by taking the average prediction error on each training sample, using only the trees in the forest that did not have that particular sample in their bootstrap sample. This allows for an unbiased estimate of the model's generalization error,

since the OOB samples are not used in any way to create or modify the model.

9. What is K-fold cross-validation?

Answer:K-fold cross-validation is a technique used to evaluate a machine learning model's performance. It involves randomly splitting the dataset into k subsets, or folds, of equal size. The model is then trained on k-1 of the folds and tested on the remaining fold. This process is repeated k times, with each fold used as the testing set once. The average of the k testing scores is used as the overall performance score for the model.

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

Answer: Hyperparameter tuning is the process of optimizing the hyperparameters of a machine learning model to improve its performance. Hyperparameters are parameters that are set before the learning process begins and are used to control the learning process. Examples of hyperparameters include the learning rate, number of layers, number of neurons, and activation functions. Hyperparameter tuning is done to find the best combination of hyperparameters that will result in the best performance for a given model.

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

Answer: A large learning rate can cause the gradient descent algorithm to overshoot the minimum value and fail to converge. It can also cause the algorithm to oscillate around the minimum value, never actually reaching it. Additionally, a large learning rate can cause the algorithm to diverge, meaning it will continue to increase or decrease in value without ever converging.

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

Answer: No, we cannot use logistic regression for classification of non-linear data because logistic regression is a linear model and is not suitable for non-linear data. Logistic regression assumes that the relationship between the dependent and independent variables is linear, which is not the case for non-linear data.

13. Differentiate between Adaboost and Gradient Boosting.

Answer: Adaboost and Gradient Boosting are both ensemble learning methods used to improve the accuracy of machine learning models.

Adaboost:
- Adaboost is a boosting algorithm that uses a weighted average of weak learners to create a strong learner.
- It works by sequentially adding weak learners to the ensemble, each one correcting the errors of the previous one.
- Adaboost is sensitive to noisy data and outliers.

Gradient Boosting:
- Gradient Boosting is an iterative algorithm that uses a gradient descent optimization technique to minimize the loss function.
- It works by sequentially adding weak learners to the ensemble, each one focusing on the errors of the previous one.
- Gradient Boosting is less sensitive to noisy data and outliers than Adaboost.

14. What is bias-variance trade off in machine learning?
Answer: Bias-variance trade off is a fundamental concept in machine learning. It is the balance between overfitting and underfitting. Over fitting occurs when a model is too complex and captures the noise in the data, while underfitting occurs when the model is too simple and fails to capture the underlying trend of the data. The bias-variance tradeoff is the process of balancing these two competing objectives in order to create an optimal model.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.
Answer: **Linear Kernel:** The linear kernel is the simplest kernel function. It is used to separate linearly separable data. It is defined as the dot product of two feature vectors.
RBF Kernel: The radial basis function (RBF) kernel is a non-linear kernel used for classifying data that is not linearly separable. It is defined as the Euclidean distance between two feature vectors.
Polynomial Kernel: The polynomial kernel is a non-linear kernel used for classifying data that is not linearly separable. It is defined as the product of two feature vectors raised to a power.