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 or Residual Sum of Squares (RSS) - R-squared is generally considered a better measure of goodness of fit in regression models. It represents the proportion of the variance in the dependent variable that is predictable from the independent variables. On the other hand, RSS measures the total errors in the model. R-squared gives an indication of how well the model explains the variance, while RSS focuses on the magnitude of errors.

1. 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) ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in Regression:** TSS represents the total variability in the dependent variable, ESS measures the variability explained by the regression model, and RSS quantifies the unexplained variability or errors in the model. The equation relating these is TSS = ESS + RSS, indicating that the total sum of squares can be decomposed into the explained sum of squares and the residual sum of squares.

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

Answer- Need for Regularization in Machine Learning: Regularization is needed in machine learning to prevent overfitting and improve the generalization of models. It adds a penalty term to the cost function, discouraging the model from fitting the training data too closely, which can result in poor performance on new, unseen data.

1. What is Gini–impurity index?

Answer- Gini-Impurity Index - The Gini impurity index is a measure of how often a randomly chosen element would be incorrectly classified in a dataset. In the context of decision trees, it is used to evaluate the quality of a split. A lower Gini impurity indicates a better split.

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

Answer- Unregularized Decision Trees and Overfitting- Yes, unregularized decision trees are prone to overfitting. They can become too complex and fit the noise in the training data, leading to poor generalization on new data.

1. What is an ensemble technique in machine learning?

Answer- Ensemble Technique in Machine Learning - Ensemble techniques involve combining the predictions of multiple models to improve overall performance. Examples include Random Forests and Gradient Boosting.

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

Answer- Difference between Bagging and Boosting- Bagging (Bootstrap Aggregating) involves training multiple models independently on different subsets of the data. Boosting, on the other hand, focuses on training models sequentially, giving more weight to misclassified instances, and combining their predictions.

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

Answer-Out-of-Bag Error in Random Forests- Out-of-bag error is the error rate of a model on the instances that were not used during its training. In random forests, it provides an unbiased estimate of the model's performance without the need for a separate validation set.

1. What is K-fold cross-validation?

Answer- K-fold Cross-Validation- K-fold cross-validation is a technique to assess a model's performance by dividing the dataset into K subsets. The model is trained and tested K times, using a different subset for testing each time.

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

Answer- Hyperparameter Tuning in Machine Learning- Hyperparameter tuning involves optimizing the parameters that are not learned during the training process. It is done to find the best configuration for a model, enhancing its performance.

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

Answer- Issues with Large Learning Rate in Gradient Descent- A large learning rate in gradient descent can lead to overshooting the minimum of the cost function, causing the algorithm to diverge and fail to converge to an optimal solution.

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

Answer- Use of Logistic Regression for Non-Linear Data- Logistic Regression is linear and may not handle non-linear data well. For non-linear data, more complex models like decision trees or support vector machines with non-linear kernels are often more suitable.

1. Differentiate between Adaboost and Gradient Boosting.

Answer- Difference between Adaboost and Gradient Boosting- Both are boosting algorithms, but Adaboost focuses on adjusting weights of instances, while Gradient Boosting builds trees sequentially, minimizing errors of previous trees.

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

Answer- Bias-Variance Trade-off in Machine Learning- The bias-variance trade-off involves finding the right balance between bias (error from overly simple models) and variance (error from overly complex models) to achieve optimal model performance.

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

Answer- Linear, RBF, Polynomial Kernels in SVM- In Support Vector Machines (SVM), kernels define the type of decision boundary. Linear kernels create linear decision boundaries, RBF (Radial Basis Function) kernels create non-linear boundaries, and Polynomial kernels introduce polynomial decision boundaries of a specified degree. Each is suited to different types of data.