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: Both measures have their strengths and weaknesses. R-squared provides a quick, intuitive assessment of model performance, but it may not always accurately capture the actual fit of the model, particularly if the data does not follow a linear relationship. On the other hand, RSS provides a more objective measure of model fit, but it is less interpretable and can be misleading if the residuals are not homoscedastic. Therefore, when comparing models, it is recommended to use both R-squared and RSS to evaluate the goodness of fit.

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): It is the sum of the squared differences between each observation and the mean of the dependent variable.
ESS (Explained Sum of Squares): It is the sum of the squared differences between each observation and the predicted value of the dependent variable.
RSS (Residual Sum of Squares) : It is the sum of the squared differences between each observation and the predicted value of the dependent variable.

The relationship between these three metrics can be seen from the following equation:

RSS = TSS – ESS

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

Answer: Regularization in machine learning is a technique used to prevent overfitting in the model. Overfitting occurs when a model learns to well and includes noise in the training data, leading to poor generalization to new data. The primary reason for using regularization is to reduce the risk of overfitting and to make the model more robust. This can be achieved by adding a penalty term to the loss function, which is the sum of the squared differences between the predicted values and the actual values, and a regularization term. There are two main types of regularization techniques: L1 regularization and L2 regularization.

4. **What is Gini–impurity index?**

Answer: The Gini-impurity index, also known as the Gini index, is a measure of statistical dispersion used in decision tree learning. It calculates the likelihood of misclassifying an instance. Gini-impurity can be used for both classification and regression trees, although its use in regression trees is less common. The value of Gini-impurity ranges from 0 to 1. Lower values indicate a better split.

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

Answer: Decision trees are powerful, but prone to memorizing noise. Unchecked, they become overgrown forests, unable to adapt to new landscapes. Pruning, depth limits, and data-hungry leaves help tame these leafy giants, boosting their ability to navigate unseen terrain

6. **What is an ensemble technique in machine learning?**

Answer : An ensemble technique in machine learning is a method that combines multiple machine learning models to improve the performance of the final model. By combining multiple models, an ensemble can improve the overall accuracy, robustness, and reliability of the predictions.

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

Answer: The main difference between Bagging and Boosting techniques lies in their approach to model training and combining. While Bagging involves training models on different subsets of the original data, Boosting involves training models sequentially, with each new model aiming to correct the errors made by the previous models.

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

Answer : The out-of-bag (OOB) error is an important concept in random forests, a popular type of ensemble learning model. It refers to the error made by the forest when it makes predictions for samples not used in the construction of the forest.

9. **What is K-fold cross-validation?**

Answer: A method that divides data into K folds, trains models on K-1 folds, tests on the held-out fold, and repeats to evaluate performance comprehensively.

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

Answer : Hyperparameter tuning is the process of choosing the best combination of hyperparameters for a specific machine learning model. It is done as it improves model performance by optimizing the learning process and selecting the best configuration settings.

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

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

Answer : Logistic regression is not ideal for classifying non-linear data due to its linear decision boundary. Logistic regression assumes a linear relationship between the independent variables and the log odds of the dependent variable.

13. **Differentiate between Adaboost and Gradient Boosting.**

Answer : Adaboost focuses on hard learners, boosting misclassified examples, while Gradient Boosting builds sequentially on weak learners, minimizing prediction errors. Both create diverse ensembles for robust predictions.

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

Answer: Bias-variance trade-off is a fundamental concept in machine learning. It describes the inherent tension between fitting the training data perfectly (low bias) and minimizing the model's variance (high bias).