

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?

Ans 1. R-squared is generally a better measure of goodness of fit in regression because it indicates the proportion of variance in the dependent variable that is predictable from the independent variables. RSS simply measures the sum of the squares of residuals and doesn't provide a normalized measure 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?

Ans 2. TSS (Total Sum of Squares) measures the total variance in the dependent variable.

ESS (Explained Sum of Squares) measures the variance explained by the regression model.

RSS (Residual Sum of Squares) measures the variance not explained by the model.

The relationship is:  $TSS = ESS + RSS$ .

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

Ans 3. Regularization is needed in machine learning to prevent overfitting by penalizing large coefficients, ensuring the model generalizes well to unseen data.

4. What is Gini-impurity index?

Ans 4. Gini impurity index is a measure of the probability of a randomly chosen element being incorrectly classified if it was randomly labeled according to the distribution of labels in the dataset. It is used in decision trees to determine the quality of splits.

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

Ans 5. Yes, unregularized decision trees are prone to overfitting because they can create overly complex models that capture noise in the training data rather than the underlying data distribution.

6. What is an ensemble technique in machine learning?

Ans 6. An ensemble technique in machine learning combines multiple models to produce a more robust and accurate prediction than any single model.

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

Ans 7. Bagging (Bootstrap Aggregating) involves training multiple models on different subsets of the training data and averaging their predictions. Boosting sequentially trains models, each trying to correct the errors of the previous model.

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

Ans 8. Out-of-bag error in random forests is the error rate estimated on the data that was not used for training each tree (the out-of-bag samples), providing an unbiased estimate of the model's performance.

9. What is K-fold cross-validation?

Ans 9. K-fold cross-validation involves dividing the dataset into K subsets, training the model on K-1 subsets, and testing it on the remaining subset. This process is repeated K times with each subset used exactly once for testing.

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

Ans 10. Hyperparameter tuning in machine learning is the process of selecting the best set of hyperparameters for a learning algorithm to improve its performance on a given dataset.

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

Ans 11. If the learning rate in Gradient Descent is too large, it can cause the algorithm to overshoot the minimum, leading to divergence or oscillation around the minimum.

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

Ans 12. Logistic Regression can struggle with non-linear data because it assumes a linear relationship between the independent variables and the log-odds of the dependent variable. Non-linear relationships require transformations or different algorithms.

13. Differentiate between Adaboost and Gradient Boosting?

Ans 13. AdaBoost focuses on reweighting misclassified instances and adjusting the model accordingly. Gradient Boosting builds models sequentially to minimize a specified loss function using gradient descent techniques.

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

Ans 14. The bias-variance trade-off in machine learning refers to the balance between a model's complexity and its ability to generalize to new data. High bias means underfitting, while high variance means overfitting.

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

Ans 15. Linear kernel computes the dot product between vectors, suitable for linearly separable data.

RBF (Radial Basis Function) kernel uses a Gaussian function to handle non-linear data by considering the distance between data points.

Polynomial kernel computes the similarity of vectors in a feature space over polynomials of the original variables, useful for non-linear data with polynomial relationships.