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 measures the proportion of variance explained by independent variables, due to which we can explain overall goodness of fit. A higher R-squared suggests a better fit. Residual Sum of Squares (RSS) measures the total squared difference between observed and predicted values, highlighting overall model accuracy. R-squared is preferred for assessing explanatory power, while RSS gauges the magnitude of residuals. Both are valuable, and the choice depends on specific analysis goals. Consider using both metrics alongside other evaluation techniques for a comprehensive assessment.

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 -: The relationship between TSS, ESS, and RSS can be expressed by the equation: TSS=ESS+RSS

TSS - TSS measures the total variability in the dependent variable. It is the sum of squared differences between each observed value and the mean of the dependent variable.

RSS - The value of R-square lies between 0 to 1. The RSS is a technique used to measure amount variance in a dataset not predicted by the regression model.

ESS - ESS represents the variability in the dependent variable that is explained by the regression model.

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

Answer -: Regularization in machine learning is necessary for the below mentioned reasons:

Preventing Overfitting: Discourages models from fitting noise in training data.

Handling Collinearity: Manages high correlation between predictor variables.

Feature Selection: Induces sparsity, automatically selecting relevant features.

Improving Model Stability: Prevents extreme parameter values for better generalization.

Enhancing Interpretability: Leads to simpler, more interpretable models.

Regularization methods like Ridge and Lasso regression provide a trade-off between model complexity and generalization, improving model performance on new data.

4. What is Gini–impurity index?

Answer -: The Gini impurity is a measure of how often a randomly chosen element in a set would be misclassified. It is used in decision trees to determine the best splits for creating nodes that lead to more accurate classifications. The Gini impurity ranges from 0 (pure set) to 1 (maximal impurity).

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

Answer -: Yes, unregularized decision trees are prone to overfitting because they can become too complex, capturing noise and specific details of the training data that may not generalize well to new data. Regularization techniques, such as pruning, are used to mitigate this issue by simplifying the tree structure.

6. What is an ensemble technique in machine learning?

Answer -: Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods in machine learning usually produce more accurate solutions than a single model would.

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

Answer -: Bagging reduces variance for stability, while Boosting reduces both variance and bias for accuracy. Pick Bagging for stability and simplicity, Boosting for higher accuracy but with more complexity.

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

Answer -: The out-of-bag (OOB) error in random forests measures the prediction accuracy on data points not used during the training of a particular tree. It provides an estimate of the model's generalization performance without requiring a separate validation set.

9. What is K-fold cross-validation?

Answer -: K-fold cross-validation is a technique for assessing a machine learning model's performance. The dataset is divided into K subsets, and the model is trained and evaluated K times, each time using a different subset as the validation set. This helps obtain a more reliable estimate of the model's performance.

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

Answer -: Hyperparameter tuning in machine learning involves optimizing settings external to the model, such as learning rates or regularization strengths, to enhance model performance. It is done to improve accuracy, prevent overfitting, increase robustness, optimize computational resources, and adapt the model to different datasets.

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

Answer -: A large learning rate in gradient descent can result in the algorithm taking overly large steps during optimization, leading to various problems. Divergence may occur, preventing the model from converging to an optimal solution. Overshooting the minimum can cause oscillations and instability, making the optimization process unpredictable. Additionally, the algorithm may fail to converge or take a long time to do so. To mitigate these issues, it's essential to choose an appropriate learning rate and consider techniques like learning rate annealing or adaptive learning rate methods. These approaches help strike a balance between the speed of convergence and the stability of the optimization process.

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

Answer -: Logistic Regression is suitable for linearly separable data in binary classification. However, it may not perform well for non-linear data patterns. In such cases, more complex models like Decision Trees, Random Forests, SVMs, or neural networks are often preferred, as they can capture non-linear relationships more effectively.

13. Differentiate between Adaboost and Gradient Boosting.

Answer -: Adaboost and Gradient Boosting are ensemble methods that differ in their approach. Adaboost sequentially corrects errors by assigning higher weights to misclassified instances, using a high learning rate. Gradient Boosting builds weak models sequentially, correcting errors by fitting to residuals with a lower learning rate. Adaboost is less prone to overfitting, employing a variety of weak learners, while Gradient Boosting often uses decision trees and can be more prone to overfitting, but regularization techniques are used to mitigate this.

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

Answer -: The bias-variance tradeoff in machine learning is a balance between a model's ability to fit the training data accurately (low bias) and its ability to generalize to new data (low variance). High bias models are too simplistic, leading to systematic errors, while high variance models are overly complex and capture noise. The goal is to find the optimal complexity that minimizes both bias and variance for effective model generalization. Techniques like cross-validation and regularization are employed to strike this balance.

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

i) Linear Kernel:

Description: Computes dot product, suitable for linearly separable data.Use Case: Appropriate when a linear decision boundary is desired.

ii) RBF Kernel:

Description: Measures similarity using Gaussian function, effective for non-linear relationships. Use Case: Suitable for data with complex, non-linear decision boundaries.

iii) Polynomial Kernel:

Description: Computes similarity based on polynomial combinations, introduces non-linearity. Use Case: Useful for capturing polynomial relationships in the data.

16. Using a goodness of fit,we can assess whether a set of obtained frequencies differ from a set of frequencies. a) Mean b) Actual c) Predicted d) Expected

Answer -: d) Expected

17. Chisquare is used to analyse a) Score b) Rank c) Frequencies d) All of these

Answer -: c) Frequencies

18. What is the mean of a Chi Square distribution with 6 degrees of freedom? a) 4 b) 12 c) 6 d) 8

Answer -: c) 6

19. Which of these distributions is used for a goodness of fit testing? a) Normal distribution b) Chisqared distribution c) Gamma distribution d) Poission distribution

Answer -: b) Chisqared distribution

20. Which of the following distributions is Continuous a) Binomial Distribution b) Hypergeometric Distribution c) F Distribution d) Poisson Distribution.

Answer -: c) F Distribution

21. A statement made about a population for testing purpose is called? a) Statistic b) Hypothesis c) Level of Significance d) TestStatistic

Asnwer -: b) Hypothesis

22. If the assumed hypothesis is tested for rejection considering it to be true is called? a) Null Hypothesis b) Statistical Hypothesis c) Simple Hypothesis d) Composite Hypothesis

Asnwer -: a) Null Hypothesis

23. If the Critical region is evenly distributed then the test is referred as? a) Two tailed b) One tailed c) Three tailed d) Zero tailed

Asnwer -: