Machine Learning Assignment-5

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 generally considered a better measure of the goodness of fit in regression as it represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It is a standardized measure ranging from 0 to 1, which makes it easier to interpret and compare across different models. RSS, on the other hand, measures the total deviation of the predicted values from the actual values but does not provide a normalized metric, making it less intuitive for assessing model performance.

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: In regression analysis:

TSS (Total Sum of Squares)- measures the total variance in the observed data.

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

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

The relationship between these metrics is given by the equation:

$$\mathbf{TSS = ESS + RSS}$$

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

Answer: Regularization is needed in machine learning to prevent overfitting by adding a penalty to the model's complexity. It helps to improve the generalization of the model to unseen data by discouraging excessively complex models that fit the training data too closely, capturing noise instead of the underlying patterns.

4. What is the Gini–impurity index?

Answer: The Gini-impurity index is a measure of the impurity or disorder used in decision tree algorithms to determine the best split at each node. It calculates the probability of a randomly chosen

element being misclassified if it was randomly labeled according to the distribution of labels in the dataset. The Gini impurity for a binary classification is given by:

$$\textbf{Gini} = 1 - (p^2 + q^2)$$

where p and q are the probabilities of the two classes.

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

Answer: Yes, unregularized decision trees are prone to overfitting because they can create overly complex trees that perfectly fit the training data, including noise and outliers. This results in a model that performs well on training data but poorly on unseen data due to its high variance.

6. What is an ensemble technique in machine learning?

Answer: An ensemble technique in machine learning combines multiple models to improve the overall performance and robustness of the prediction. The idea is to leverage the strengths of different models to produce a more accurate and stable output. Common ensemble techniques include Bagging, Boosting, and Stacking.

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

Answer:

Bagging (Bootstrap Aggregating): Involves training multiple models independently on different subsets of the training data (created through bootstrapping) and averaging their predictions. It helps reduce variance and prevent overfitting.

Boosting: Involves training multiple models sequentially, where each model focuses on correcting the errors made by the previous ones. It aims to reduce both bias and variance, leading to a strong predictive model by combining weak learners.

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

Answer: The out-of-bag (OOB) error in random forests is an estimate of the model's prediction error. It is calculated using the data not included in the bootstrap sample for each decision tree. Since each tree is trained on a subset of the data, the OOB samples provide a way to validate the model without needing a separate validation set.

9. What is K-fold cross-validation?

Answer: K-fold cross-validation is a technique used to evaluate the performance of a machine learning model. The dataset is divided into K equally sized folds. The model is trained on K-1 folds and tested on the remaining fold. This process is repeated K times, with each fold serving as the test set once. The results are averaged to provide a more robust estimate of model performance.

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

Answer: Hyperparameter tuning is the process of finding the optimal set of hyperparameters for a machine learning model. Hyperparameters are the configuration settings used to control the learning process. Tuning is done to improve the model's performance by selecting the best combination of hyperparameters that minimize prediction error on validation data.

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

Answer: A large learning rate in Gradient Descent can cause the optimization process to overshoot the minimum of the loss function, leading to divergence or oscillation around the minimum rather than convergence. This results in poor model performance and instability during training.

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

Answer: Logistic Regression is not suitable for non-linear data classification because it assumes a linear relationship between the input features and the log-odds of the outcome. For non-linear data, the decision boundary created by Logistic Regression will be insufficient, leading to poor performance. Techniques such as kernel methods or more complex models like Neural Networks are better suited for non-linear data.

13. Differentiate between Adaboost and Gradient Boosting.

Answer:

Adaboost (Adaptive Boosting): Focuses on training weak learners sequentially, where each subsequent learner focuses more on the errors made by the previous ones. It assigns weights to incorrectly classified instances, making them more important in the next iteration.

Gradient Boosting: Builds models sequentially by optimizing the residual errors of the previous models. Each new model is trained to correct the errors made by the combined previous models using gradient descent on the loss function.

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

Answer: The bias-variance trade-off refers to the balance between two sources of error that affect model performance:

Bias: Error due to oversimplified models that cannot capture the underlying patterns (underfitting).

Variance: Error due to overly complex models that capture noise along with the underlying patterns (overfitting).

The goal is to find a model with low bias and low variance, providing good generalization to unseen data.

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

Answer:

Linear Kernel: Uses the dot product of the input features. It is suitable for linearly separable data.

RBF (Radial Basis Function) Kernel: Uses the exponential function of the squared Euclidean distance between points. It is suitable for non-linear data and can handle complex decision boundaries.

Polynomial Kernel: Uses the polynomial combination of the input features. It can create more flexible decision boundaries by increasing the degree of the polynomial.