MACHINE LEARNING

Q1 to Q15 are subjective answer type questions, Answer them briefly.

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 goodness of fit because it provides a normalized measure of how well the model explains the variability of the response data. It indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, whereas RSS is the sum of the squares of residuals and does not provide this proportion.

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** measures the total variance in the dependent variable. **ESS** measures the variance explained by the regression model. **RSS** measures the variance not explained by the model (the residuals).
Equation: TSS= ESS+RSS

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

Answer: • Regularization is needed to prevent overfitting by adding a penalty to the loss function for large coefficients. This encourages simpler models that generalize better to new, unseen data.

4. What is Gini–impurity index?

Answer: The Gini impurity index is a measure of the likelihood of an incorrect classification of a new instance if it was randomly classified according to the distribution of the target variable in a dataset. It is used to evaluate splits in decision trees.

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

Answer: Yes, unregularized decision trees are prone to overfitting because they can create very complex trees that perfectly fit the training data, including noise, but do not generalize well to new data.

6. What is an ensemble technique in machine learning?

Answer: An ensemble technique in machine learning combines multiple models to produce a single model that is typically more robust and performs better than any individual model.

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

- **Answer**: Bagging (Bootstrap Aggregating) involves training multiple models independently on different random subsets of the data and averaging their predictions. Boosting involves training models sequentially, where each model focuses on correcting the errors of the previous ones.

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

Answer: The out-of-bag error is an estimate of the prediction error for a random forest model. It is calculated using the predictions for each observation from the trees that did not use that observation in their training set.

9. What is K-fold cross-validation?

- **Answer**: K-fold cross-validation is a technique where the dataset is randomly partitioned into K equal-sized subsets. Each subset is used as a test set while the remaining K-1 subsets are used for training. This process is repeated K times, and the results are averaged to get the final performance estimate.

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

- **Answer**: Hyperparameter tuning is the process of selecting the best set of hyperparameters for a learning algorithm. It is done to optimize the model's performance on unseen data.

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

- **Answer**: A large learning rate can cause the gradient descent algorithm to overshoot the optimal solution, resulting in divergence or oscillation rather than convergence to the minimum loss.

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

- **Answer**: Logistic Regression is a linear model and cannot capture non-linear relationships in the data. However, by using polynomial features or kernel methods, it can be adapted to handle non-linear data.

13. Differentiate between Adaboost and Gradient Boosting.

- **Answer**: Adaboost adjusts the weights of incorrectly classified instances, making them more significant for the next model. Gradient Boosting builds models sequentially where each new model corrects the residual errors of the previous models by optimizing a loss function.

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

- **Answer**: The bias-variance trade-off is the balance between the error introduced by approximating a real-world problem (bias) and the error introduced by the model's sensitivity to fluctuations in the training data (variance). Ideally, one aims to find a model with low bias and low variance.

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

- **Answer:**
    - **Linear kernel:** Computes a linear boundary for classification, suitable for linearly separable data.
    - **RBF (Radial Basis Function) kernel:** Computes a non-linear boundary using a Gaussian function, suitable for non-linearly separable data.
    - **Polynomial kernel:** Computes a polynomial boundary by considering not only the input features but also their interactions up to a certain degree, allowing for more complex boundaries.