MACHINE LEARNING

Assignment by: Divya Chowdaiah

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 ofgoodness of fit model in regression and why?**
   R-squared or Residual Sum of Squares (RSS): R-squared is a better measure of the goodness of fit model in regression. R-squared represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, where 1 indicates a perfect fit. RSS, on the other hand, measures the sum of squared differences between the observed and predicted values. While RSS provides information on the overall fit, R-squared gives the proportion of variability explained by the model.

2. **What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sumof Squares) in regression. Also mention the equation relating these three metrics with each other.**

   TSS: Total Sum of Squares is the sum of squared differences between the observed dependent variable and its mean.
   ESS: Explained Sum of Squares is the sum of squared differences between the predicted values and the mean of the dependent variable.
   RSS: Residual Sum of Squares is the sum of squared differences between the observed and predicted values.
   Equation: TSS = ESS + RSS

3. **What is the need of regularization in machine learning?**
   Regularization is needed to prevent overfitting in machine learning models. It adds a penalty term to the cost function, discouraging overly complex models. This helps in generalizing the model to unseen data and avoids fitting the noise in the training data.

4. **What is Gini–impurity index?**
   Gini impurity is a measure of how often a randomly chosen element would be incorrectly classified. It is used in decision tree algorithms to evaluate the impurity or disorder in a set of data.

5. **Are unregularized decision-trees prone to overfitting? If yes, why?**
   Yes, unregularized decision-trees are prone to overfitting. They can become highly specialized to the training data, capturing noise and outliers, leading to poor performance on new, unseen data.

6. **What is an ensemble technique in machine learning?**
   An ensemble technique combines multiple individual models to create a stronger and more accurate model. Common ensemble techniques include Random Forests, Bagging, and Boosting.

7. **What is the difference between Bagging and Boosting techniques?**
   Bagging vs. Boosting:
   Bagging (Bootstrap Aggregating) involves training multiple instances of the same model on different subsets of the training data and combining their predictions.
   Boosting focuses on training weak learners sequentially, where each new model corrects errors made by the previous ones.

8. **What is out-of-bag error in random forests?**
   Out-of-bag error is the error rate of a machine learning model on the samples that were not used in training but were left out during the bootstrap sampling. It provides a built-in estimate of a model's performance without the need for a separate validation set.

9. **What is K-fold cross-validation?**
   K-fold cross-validation involves dividing the dataset into K subsets, using K-1 subsets for training,

and the remaining one for validation. This process is repeated K times, and performance metrics are averaged.

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

Hyperparameter tuning involves adjusting the hyperparameters of a machine learning model to optimize its performance. It is done to find the best configuration for hyperparameters that control the learning process.

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

Issues with Large Learning Rate in Gradient Descent: A large learning rate in gradient descent may lead to overshooting the minimum point, causing the algorithm to diverge and fail to converge to the optimal solution.

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

Logistic Regression for Non-Linear Data: Logistic Regression is a linear model and may not perform well on highly non-linear data. For non-linear data, more complex models like support vector machines or decision trees are often more suitable.

13. Differentiate between Adaboost and Gradient Boosting.

Adaboost focuses on correcting errors of weak learners by assigning more weight to misclassified instances.

Gradient Boosting builds trees sequentially, with each tree correcting errors made by the previous ones using gradients.

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

The bias-variance trade-off refers to the trade-off between a model's ability to fit the training data (bias) and its ability to generalize to new, unseen data (variance). Increasing model complexity reduces bias but may increase variance and vice versa.

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

Linear Kernel: Suitable for linearly separable data.

RBF (Radial Basis Function) Kernel: Effective for non-linear data, uses a Gaussian function.

Polynomial Kernel: Useful for data with polynomial decision boundaries, allows modeling of complex relationships.