Machine Learning Assignment

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?

   R-squared is a better measure of goodness of fit because it provides a relative measure of how well the regression predictions approximate the real data points. It is a ratio that gives you a sense of the proportion of the variance in the dependent variable that is predictable from the independent variables.

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.

   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 residuals)
   Equation: TSS=ESS+RSSTSS = ESS + RSSTSS=ESS+RSS

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

   Regularization is needed to prevent overfitting by adding a penalty to the model complexity. This helps ensure that the model generalizes well to new, unseen data rather than just memorizing the training data.

4. What is Gini–impurity index?

   The Gini impurity is a measure used in decision trees to determine how often a randomly chosen element would be incorrectly labelled. It ranges from 0 (perfectly pure) to 0.5 (maximum impurity).

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

   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 pattern.

6. What is an ensemble technique in machine learning?

   Ensemble techniques in machine learning combine multiple models to improve the overall performance. The idea is that multiple models, when combined, can reduce errors compared to individual models.

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

   Bagging: Involves training multiple models in parallel on different subsets of the data and averaging their predictions.

   Boosting: Involves training models sequentially, where each new model focuses on correcting the errors made by the previous models.

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

   Out-of-bag error is an estimate of the model error based on the predictions for the training instances that were not used in training each individual tree (the out-of-bag samples).

9. What is K-fold cross-validation?

   K-fold cross-validation is a technique for evaluating the performance of a model by dividing the data into K subsets and training the model K times, each time using a different subset as the test set and the remaining data as the training set.

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

   Hyperparameter tuning involves finding the best set of hyperparameters for a learning algorithm to improve its performance. This is done because the optimal hyperparameters can significantly affect the model's accuracy and generalizability.

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

   A large learning rate can cause the gradient descent algorithm to overshoot the minimum, leading to divergence or poor convergence of the model.

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

   No, because Logistic regression is inherently a linear classifier. It cannot handle non-linear relationships between the features and the target variable unless you use non-linear transformations or polynomial features.

13. Differentiate between Adaboost and Gradient Boosting.

   Adaboost: Focuses on correcting the errors of the previous models by giving more weight to misclassified instances.

   Gradient Boosting: Builds models sequentially, where each new model is trained to correct the residual errors of the previous models using gradient descent.

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

   The bias-variance trade-off refers to the balance between a model's ability to minimize bias (error due to overly simplistic assumptions) and variance (error due to sensitivity to small fluctuations in the training set). High bias can cause underfitting, while high variance can cause overfitting.

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

   Linear Kernel: Uses a linear function to separate the data. Suitable for linearly separable data.

   RBF (Radial Basis Function) Kernel: Uses a Gaussian function to create a non-linear boundary. Suitable for non-linearly separable data.

   Polynomial Kernel: Uses polynomial functions to separate the data. Allows for more complex decision boundaries based on the polynomial degree.