MACHINE LEARNING

Answer to all questions in 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 1. **R-squared vs. Residual Sum of Squares (RSS):**

- **R-squared** is generally a better measure of the goodness of fit in regression because it represents the proportion of variance explained by the model. It is easier to interpret as it provides a relative measure (0 to 1) of how well the model explains the variability of the response data.
- **RSS** represents the total squared difference between the observed and predicted values, which can be less intuitive and harder to compare across different models.

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 2. **TSS, ESS, and RSS in Regression:**

- **TSS (Total Sum of Squares)**: Measures the total variance in the response 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 (errors).
- **Equation**: TSS=ESS+RSSTSS = ESS + RSSTSS=ESS+RSS

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

Answer 3. **Need for Regularization:**

- Regularization is needed to prevent overfitting by adding a penalty for large coefficients in the model, thereby simplifying the model and improving its generalization to new data.

4. What is Gini–impurity index?

Answer 4. **Gini-Impurity Index:**

- The Gini impurity index measures the likelihood of an incorrect classification of a randomly chosen element if it was randomly labeled according to the distribution of labels in the dataset. It is used in decision trees to select the best splits.

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

Answer 5. **Unregularized Decision-Trees and Overfitting:**

- Yes, unregularized decision trees are prone to overfitting because they can create very complex models that fit the training data perfectly, including the noise, which does not generalize well to unseen data.

6. What is an ensemble technique in machine learning?

Answer 6. **Ensemble Technique:**

- Ensemble techniques combine multiple machine learning models to improve the overall performance by reducing variance, bias, or improving predictions.

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

Answer 7. **Bagging vs. Boosting:**

- **Bagging (Bootstrap Aggregating)**: Reduces variance by training multiple models independently on random subsets of the data and averaging their predictions.
- **Boosting**: Reduces bias by sequentially training models, where each model corrects the errors of the previous ones, combining their outputs for the final prediction.

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

Answer 8. **Out-of-Bag Error in Random Forests:**

- Out-of-bag error is an estimate of the model's prediction error on new data, calculated using the samples that were not included in the bootstrap sample for training each tree in the random forest.

9. What is K-fold cross-validation?

Answer 9. **K-Fold Cross-Validation:**

- K-fold cross-validation involves splitting the dataset into K subsets, training the model on K-1 subsets, and validating it on the remaining subset. This process is repeated K times, and the results are averaged to estimate model performance.

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

Answer 10. **Hyperparameter Tuning:**

- Hyperparameter tuning is the process of finding the optimal set of hyperparameters for a machine learning model to improve its performance. It is done to ensure the model generalizes well to unseen data.

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

Answer 11. **Issues with Large Learning Rate in Gradient Descent:**

- A large learning rate can cause the model to overshoot the optimal solution, leading to divergence rather than convergence, and potentially making the training process unstable.

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

Answer 12. **Logistic Regression for Non-Linear Data:**

- Logistic regression is not suitable for non-linear data as it assumes a linear relationship between the input features and the log-odds of the output. For non-linear data, more complex models or transformations are required.

13. Differentiate between Adaboost and Gradient Boosting.

Answer 13. **Adaboost vs. Gradient Boosting:**

- **Adaboost**: Adjusts the weights of incorrectly classified instances and trains new models sequentially to correct these errors.
- **Gradient Boosting**: Sequentially fits new models to the residual errors made by previous models, using gradient descent to minimize the loss function.

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

Answer 14. **Bias-Variance Trade-Off:**

- The bias-variance trade-off is the balance between model complexity and the ability to generalize. High bias leads to underfitting, while high variance leads to overfitting. The goal is to find a model with optimal complexity that minimizes both bias and variance.

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

Answer 15. **Linear, RBF, Polynomial Kernels in SVM:**

- **Linear Kernel**: Suitable for linearly separable data, computes the dot product of the input features.
- **RBF (Radial Basis Function) Kernel**: Suitable for non-linear data, computes the similarity based on the distance between points.
- **Polynomial Kernel**: Suitable for polynomial relationships, computes the similarity using polynomial functions of the input features.