ASSIGNMENT - 5

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?
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.
3. What is the need of regularization in machine learning?
4. What is Gini impurity index?
5. Are unregularized decision trees prone to overfitting? If yes, why?
6. What is an ensemble technique in machine learning?
7. What is the difference between Bagging and Boosting techniques?
8. What is out of bag error in random forests?
9. What is K-fold cross validation?
10. What is hyper parameter tuning in machine learning and why it is done?
11. What issues can occur if we have a large learning rate in Gradient Descent?
12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
13. Differentiate between Adaboost and Gradient Boosting.
14. What is bias-variance trade off in machine learning?
15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

1Ans) R-squared is a better measure of goodness of fit because it represents the proportion of variance explained by the model. In contrast, root square residual (RSS) measures unexplained variance and provides additional information about the functional model.

2Ans) TSS (Total Sum of Squares) represents the total variability of the dependent variable. ESS (Sum of Squares of Explanatory Variables) measures the variability explained by the model and RSS (Residual Sum of Squares) measures the unexplained variability. The equation connecting them is TSS = ESS + RSS, which represents how all variables are divided into explained and unexplained variables.

3 Ans) Regularization in machine learning is necessary to prevent overfitting by penalizing models that are too complex. It prevents the model from getting too noisy in the training data by adding a time penalty to the model's objective function.

4Ans) Gini Impurity Index is a metric used in decision trees to measure the level of impurities or problems in the material. Data points. It is reduced when a node contains data points from only one group, indicating a pure distribution.

5Ans) Continuous decision trees can become complicated by keeping noise in the training data as if it were the real model, which can lead to overfitting. . This inhibits their ability to effectively expand on new, unseen information.

6Ans) Integration in machine learning involves combining predictions from multiple models to improve overall performance and robustness. They tap into the wisdom of the crowd by combining different models.

7 Ans) Bagging and Supplementation are related but their methods are different. Bagging builds multiple models independently and averages their predictions, while Boosting builds models sequentially, with each model correcting the error of the previous model.

8 Ans) The out-of-bag error in a random forest is a measure of the model's prediction error for all tree data without using points during training. It is used as a validation metric and does not require a separate validation set.

9 Ans) K-fold cross-validation is a process that divides the dataset into K subsets and trains and tests the sample K times, using a different test each time. This helps achieve better performance metrics and reduce the impact of dataset variability.

10 Ans) Hyperparameter tuning in machine learning will tune parameters that are not learned from data. This is done to improve the performance of the model by finding the best combination of hyperparameters (such as learning rate or power constant).

11) The main learning in gradient descent will cause the minimum to be exceeded, causing the algorithm to move away from the optimal solution rather than getting closer to it. This can lead to instability and poor performance of the model.

12 Ans) Logistic regression is often used for linear distributions. Although it can handle some nonlinear models, it will be difficult to handle nonlinear relationships. In this case, more models such as support vector machines or decision trees are often preferred.

13 Ans) Adaboost and gradient boosting are both boosting algorithms, but their purposes are different. Gradient Boosting creates a series of samples, with each sample correcting the error of the previous sample, while Adaboost adjusts the weights of the data points.

14 And) Differences in machine learning mean how well the model fits the training data (less bias) and its ability to expand to new, unobserved data (lower variance). Ensuring parallelism is critical for modeling performance on both training and test data.

15 Ans) The linear kernel in SVM calculates the point features of the input, the RBF (Radial Basis Function) kernel uses a Gaussian function to measure the similarity of data points, and the polynomial kernel includes many functions to capture non-existent data. -linear relationships. Each core has a specific purpose, such as transferring input data to a high-resolution cluster.