Assignment : Machine Learning **Internship No. : DSC2311**

Question 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?

Ans :- R-squared will give better measure of goodness of fit model in regression , because when Residual sum of Square less and total sum of square being large then 2nd term of R-Square being near to 0 and R-Square being near t0 1, So it will give best fit of line compare to Residual sum of Square.

Question 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.

Ans :-

TSS :- TSS tells us how much variation there is in the dependent variable.

ESS :- ESS tells us how much variation in dependent variable explained by our model.

RSS :- RSS tells us how much variation in dependent variable which is not explained by our model.

TSS=ESS+RSS

Question 3 :-What is the need of regularization in machine learning?

Ans :- While training machine learning, the model can easily be over-fitted or under-fitted.

To avoid this, we use Regularization in ML to properly fit a model into our test data.

Question 4 :- What is Gini–impurity index?

Ans :- Gini-Impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in subset.

Question 5 :- Are un-regularized decision-trees prone to overfitting? If yes, why?

Ans :- Un regularized Decision-Tree prone to over-fitting. Over-fitting occurs when the tree becomes too complex and capture the noise in training data, rather than underlying pattern.

Question 6 :- What is an ensemble technique in machine learning?

Ans :- Ensemble is measure of homogeneity of sample in node, if sample is completely homogenous the entropy is 0, if sample is equally divided it has entropy is 0.

Question 7 :- What is the difference between Bagging and Boosting techniques?

Ans :- Bagging and Boosting are ensemble technique.

Bagging involves fitting many models on different samples of dataset and averaging the prediction.

Boosting involves adding ensemble member sequentially to correct the prediction made by prior models and outputs weighted average of prediction.

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

Ans :- Out-of-bag errors are an estimate of the performance of random forest classifier or regression on unseen data.

Question 9 :- What is K-fold cross-validation?

Ans :- In K-Fold cross-validation, we split the dataset into k number of subsets (known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

Question 10 :- What is hyper-parameter tuning in machine learning and why it is done?

Ans :- Hyper-parameter are parameters whose values control the learning process.

These are adjustable parameters used to obtain an optimal model.

These can be considered as External parameters.

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

Ans :- If learning rate is too high, the algorithm will take too big of steps and continuously miss the optimal solution. So, it is important to choose an appropriate learning rate for the problem.

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

Ans :- Yes, it is possible to use logistic Regression as a non-linear classifier by formulating with a non linear model.

Question 13 :- Differentiate between Adaboost and Gradient Boosting.

Ans :-

- AdaBoost is the first designed boosting algorithm with a particular loss function, while Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem.
- Gradient Boosting is much more flexible than AdaBoost.
- The adaptive boosting method minimizes the exponential loss function which changes the algorithm more profound to its outliers, while in gradient boosting, the differentiable loss function makes it more sensitive to outliers when compared to AdaBoost.
- AdaBoost is computed with a specific loss function and becomes more rigid when comes to few iterations, while Gradient Boosting tries to fit the new predictor to the residual errors made by the previous predictor.
- Gradient boosting has a fixed base estimator i.e., Decision Trees whereas in AdaBoost we can change the base estimator according to our needs.

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

Ans :- If algorithms too simple then it may be on high bias and low variance and thus is error prone.
If algorithm fit too complex then it may be high variance and low bias.
In better condition, the new entries will not perform, there is something between both of these condition known as Trade-off or Bias-variance Trade-off.

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

Ans :-

Linear Kernel :- A linear kernel SVM is a type of support vector machine (SVM) that is used to classify data that is linearly separable. It is one of the most common kernels to be used and is mostly used when there are a large number of features in a particular dataset.

RBF Kernel :- The Radial Basis Function (RBF) kernel is a popular kernel function used in Support Vector Machines (SVMs) for classification, regression, and outlier detection.

Polynomial Kernel :- A polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.