Assignment 5

Machine Learning

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 or RSS both can be measure for goodness of fit model in regression, but there is little difference in both the model. The difference is R -Squared is measure on the basis of statistically, which measure the changes in range of dependent variable and independent variable. It measures the overall proportion of variance in dependent value. On the other hand, Residual Sum of Squares (RSS)measure difference between actual value or predicated of the dependent variable. RSS Value is mainly use to reduce the sum of square.

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

   1 TSS (Total Sum of Squares) -Total sum of square is a dependent variable which is differ from the dependent variable mean.

$$TSS = \sum_{i-1}^{n} (yi - \bar{y})\,2$$

   2 ESS (Explained Sum of Squares)-Explained Sum of Squares explain about the modelled data.

$$ESS = \sum_{i-1}^{n} (\hat{y}\imath - \bar{y})\,2$$

   3 RSS (Residual Sum of Squares)-Residual Sum of Squares is measure the mistaken in variation**.**

$$RSS = \sum_{i-1}^{n} (yi - \hat{y}\imath)\,2$$

3. What is the need of regularization in machine learning?
   Answer3
   Regularization in machine learning is required to solve the problem of overfitting of extra or add data, and also simplify the complexity of training data. Overfitting occur when the training error decreases

4. What is Gini–impurity index?
   Answer 4
   In this we measure the impurity of set of data, Impurity defines in this by mixing in various unwanted set of data. The higher rate of impurity means a higher gini impurity.it is used in Machine learning decision tree to measure how to divide in to small - small data set to sort the data set.

5. Are unregularized decision-trees prone to overfitting? If yes, why?
6. What is an ensemble technique in machine learning?
   Answer 6
   Ensemble technique in machine learning says kind of server happen before purchasing an product as there are various example for the same like want an car people go for car daekho, want buy a mobile go for comparing website which tell about the feature or mobile and we can take also an example of swiggy

Zomato etc are some example which tell about the product and suggest the best. Similarly, the following technique perceive from various machine learning alogrithms to get more appropriate result to analysis. There are various types Ensemble machine learning-
Bagging
Bosting
Stacking

7. What is the difference between Bagging and Boosting techniques?
Answer 7
Bagging – In Bagging techniques parallel trained various week learners. The input data is trained for week learning by randomly select a data set from original data set with replacement. These models are interdependent each other.

Boosting – Boosting techniques is an ongoing process in which new models try to correct the mistake made by the Bagging model or stacking model.

Difference between Bagging and Boosting

SNo.	Bagging	Boosting
1	Combine various trained model on different subset data.	Focuses on error made by pervious model
2	Improving accuracy	Higher Accuracy
3	Handle the Unitability	Reduce Bias
4	Model Serve equal weightage in final output	Model Based on Accuracy and fillable

8. What is out-of-bag error in random forests?
Answer 8
Out -of -Bag error is used to measure the error in precedence of tree-based model in machine learning Bagging method. Number of wrong classifications is a out- of -bag error. It serves better precedence and less overfitting problem. Calculation of Out of bag error reach out both trees-based model and also in machine learning.

9. What is K-fold cross-validation?

Answer 9

The data sample has to divided into K number of smaller sample is a K fold cross. Value of K data is depended on size of the data. It uses to evaluate in limited data sample. Mostly it evaluates the moder design. 1st subset data consider as test data and other data consider as a training data.

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

Answer 10

Hyper parameter turning into a machine learning because maximizing the performance for optimum result.

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

Answer 11

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

Answer 12 Yes we can use logistic regression for classification o non linear data

13. Differentiate between Adaboost and Gradient Boosting.

Answer 13

S No.	Adaboost	Gradient Boosting
1	Adaboost is voting weight	Gradient is adding gradient optimizing
2	Gave more accuracy on targeted model	It increases accuracy by loss function take it as target for iteration

Thus, both is bossting alorithms which is combination of various week learner in decision stump to generate the strong learner

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

Answer 14

Making a relation between bias and variance is a bias variance

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

Answer 15

Linear- Linear is a highly used statistical technique for analysis.it explain the relation between depending variable and explanatory variable

RBF- Radial Base Function is a machine learning algorithm is use for classification regression work. It works om non liner and high dimensional data, it is a popular Kernels function

Polynoial Kernels- Polynoial Kernels is a function to transfer the input data into higher dimensional space.it represent data in form of dots in the graphical pictures.