ASSIGNMENT – 8

MACHINE LEARNING

In Q1 to Q7, only one option is correct, Choose the correct option:

1. What is the advantage of hierarchical clustering over K-means clustering?

   Answer- B) In hierarchical clustering you don't need to assign number of clusters in beginning

2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

   Answer-A) max_depth

3. Which of the following is the least preferable resampling method in handling imbalance datasets?

   Answer-C) RandomUnderSampler

4. Which of the following statements is/are true about "Type-1" and "Type-2" errors?

   1. Type1 is known as false positive and Type2 is known as false negative.

   2. Type1 is known as false negative and Type2 is known as false positive.

   3. Type1 error occurs when we reject a null hypothesis when it is actually true.

   Answer- C) 1 and 3

5. Arrange the steps of k-means algorithm in the order in which they occur:

   1. Randomly selecting the cluster centroids

   2. Updating the cluster centroids iteratively

   3. Assigning the cluster points to their nearest center

   Answer-D) 1-3-2

6. Which of the following algorithms is not advisable to use when you have limited CPU resources and time, and when the data set is relatively large``

   Answer-B) Support Vector Machines

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?

   Answer-C) CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node)

In Q8 to Q10, more than one options are correct, Choose all the correct options:

8. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?

   Answer-A) Ridge will lead to some of the coefficients to be very close to 0

   D) Lasso will cause some of the coefficients to become 0.

9. Which of the following methods can be used to treat two multi-collinear features?

   Answer- A) remove both features from the dataset

   B) remove only one of the features

10. After using linear regression, we find that the bias is very low, while the variance is very high. What are the possible reasons for this?

    Answer-A) Overfitting

    C) Underfitting

Q11 to Q15 are subjective answer type questions, Answer them briefly.

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?

Answer-we prefer one hot encoding when we require the values in binary form otherwise we must avoid the one hot encoding. Both the encoding technique perfom excellent as per there own way ,mostly prefer to use Label

encoding,when we have to deal with lots of data and we neet it in continues values.

12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.

Answer- In imabalanced dataset first we need to think what we require to oversampling or downsampling as per that we need to use the sampling technique .There are different sampling techniques to use the mostly people prefer SMOTE to balance the data.

13. What is the difference between SMOTE and ADASYN sampling techniques?

Answer-The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed data

14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?

Answer- Gridsearch cv tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method, Hence after using this function we get accuracy/loss for every combination of hyperparameters and we can choose the one with the best performance.

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief

Answer- Mean Absolute Error, Mean Squared Error , Root Mean Square Error , Root Mean square Log Error ,these are the list of evaluation metrics are mainly using this metrics to calculate the errors.

Mean Absolute Error => MAE is sum of absolute error if we have 100 as actual value and 130 as a predicted value here the absolute error is 30 it

doesn't consider the direction whether it is negative and positive values , we are calculating the all absolute error and finding the mean from it.

Mean Square Error => In mean square error we calculate the error ocuured in between actual value and predicted value and squaring them in wether it is negative value it will automatically become the positive value and finding the mean from that all.

Root Mean Square Error => In this the formula is very similar to mean square error it is just we need to add the square to sign in it here it is indicates the residual error it is always positive and lower value indicates the better performance ideal value would be 0 but it is never achived. Root Mean square

Log Error => it is calculated at log arithmetic scale ,RMSLE is added 1 as constant of actual and predicted value because they can be 0 but log od 0 will be undefined like we have actual value=100 and predicted value=130 we add 1 as constant in each 101 and 131 then we will find the log of actual and predicted value after that we will find the error from the value then squaring it after that we will calculate the mean