Assignment 8 -Machine Learning

By- Rakesh Shinde

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

Ans: 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?

Ans: A) max_depth

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

Ans:

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.

Ans: 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

Ans: 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?

Ans: B) Support Vector Machines

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

Ans: 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)

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

Ans:

A) Ridge will lead to some of the coefficients to be very close to 0
B) 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?

Ans:

B) remove only one of the features

C) Use ridge regularization

D) use Lasso regularization

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?

Ans:

A) Overfitting, B) Multicollinearity, D) Outliers

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

Ans:

If the categorical values are more then the one hot encoder should be avoided, because one hot encoder increases the dimension of the dataset leading to curse of dimensionality. We can use Label Encoder, Ordinal Encoder,

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

Ans: Resampling is the technique to be used to make the dataset balance by oversampling and undersampling

Oversampling: This means the minority class problems to be created to go up to majority class. The most common technique is called SMOTE (Synthetic Minority Over-sampling Technique)

Undersampling: This means majority class data to be deleted randomly to come up to minority class can be done by tomek link

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

Ans: ADASYN are basically improved version of the SMOTE, in adasyn generates samples next to original samples which are wrongly classified using k nearest neighbors. The generated samples are giving different no. of samples depending on data distribution of a class to be oversampled

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

Ans: GridSearchCV is used for Hyper Parameter tuning in order to determine the optimal values of parameters of best performing model.

GridSearchCV is not recommended for the large dataset, as it requires more computational time and becomes expensive, also complexity also increases

We can use RandomizedSearch CV for quick results

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

Ans:

Mean Absolute error: It is absolute difference bet actual value and predicted, it is same unit of output variable

Mean Squared error: it is squared difference between actual and predicted, which avoids the negative terms

Root mean squared error: It is square root of mean squared error, low RMSE means model is predicting well

R2score: r2 score is also known as goodness of fit. It is the ratio of sum of square of regression line to sum of square of mean line. R2 score moves towards 1 means the model is fitting well