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?
A) Hierarchical clustering is computationally less expensive
B) In hierarchical clustering you don't need to assign number of clusters in beginning
C) Both are equally proficient
D) None of these

Ans: B

2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?
A) max_depth
B) n_estimators
C) min_samples_leaf
D) min_samples_splits

Ans: A

3. Which of the following is the least preferable resampling method in handling imbalance datasets?
A) SMOTE
B) RandomOverSampler
C) RandomUnderSampler
D) ADASYN

Ans: C

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.

A) 1 and 2
B) 1 only
C) 1 and 3
D) 2 and 3

Ans: C

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

A) 3-1-2
B) 2-1-3
C) 3-2-1
D) 1-3-2

Ans: D

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?

A) Decision Trees
B) Support Vector Machines
C) K-Nearest Neighbors
D) Logistic Regression

Ans: B

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

A) CART is used for classification, and CHAID is used for regression.
B) CART can create multiway trees (more than two children for a node), and CHAID can only create binary trees (a maximum of two children for a node).
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)
D) None of the above

Ans: C

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?
A) Ridge will lead to some of the coefficients to be very close to 0
B) Lasso will lead to some of the coefficients to be very close to 0
C) Ridge will cause some of the coefficients to become 0
D) Lasso will cause some of the coefficients to become 0.

Ans: A & D

9. Which of the following methods can be used to treat two multi-collinear features?
A) remove both features from the dataset
B) remove only one of the features
C) Use ridge regularization
D) use Lasso regularization

Ans: B, C & D

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?
A) Overfitting
B) Multicollinearity
C) Underfitting
D) Outliers

Ans: A & C

Q10 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?

Ans: The Machine Learning model takes only numerical data to train but sometimes dataset contains numerical as well as categorical column also. Before sending these types of data for training we must have to deal with the categorical data and change it into numerical data. For this we have some encoding techniques like One-hot encoder and Label-encoder. These techniques convert the unique categories of column into different numerical values.

The categorical column can contain only two unique features as well as more than that. So, at that time when there are more than two unique features contains in categorical column, we can't use One-hot encoder it must be avoided at that time because it has some limitations. It can be use only for two unique categorical data.

When there are more than two unique categorical data contains in column we use Label-encoder to convert these data into unique numerical data.

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

Ans: Sometimes we have a dataset in a classification problem that is biased toward a single output like if we have two categories of output data i.e., 0 & 1 then we have more 1 than 0 or vice versa. So, at that time the problem of imbalance occurs in data due to that the model gets biased towards one type of

data i.e., it learns only one type of data and doesn't be able to learn the 2nd type properly.

If we don't handle this situation our model won't be able to give correct outputs.

We can handle the imbalance of data by the following two types of techniques:

- Under Sampling
- Over Sampling

Further, we have different methods for these techniques.

Under Sampling: - We use this technique when we have a very huge dataset like lakhs of rows. This chooses some random data from the biased side to make it equal to the small data and left the rest data then trains the model with that fewer data.

Let's understand it with an example, suppose we have a fraud dataset where it gives us information about fraud transactions and normal transactions from credit cards. Our dataset contains 2L rows and 15 columns. It is quite huge. Our target column contains two types of variables i.e., 0 and 1 where 0 represents a normal transaction and 1 represents a fraud transaction. When we have seen it, we found out of 2L rows we have only 2000 rows that contain the data of fraud transactions i.e., 1. It means we have more than 90% of the same type of data so it is more obvious that our model will learn only one type of data.

Here when we apply under-sampling methods this will randomly select some data from the higher side and try to make it equivalent to the less side and left out the rest data.

As the under-sampling technique deletes the data so we don't use it more. Otherwise, we will face data loss and the model will not be able to train properly.

Following are the methods which is used for Under-sampling:

- Cluster Centroids
- Condensed Nearest Neighbor
- Edited Nearest Neighbors
- Repeated Edited Nearest Neighbors
- All KNN
- Instance Hardness Threshold
- Near Miss
- Neighborhood Cleaning Rule
- One Sided Selection
- Random Under Sampler
- Tomek Links

Over Sampling: - This technique makes some elastic samples from its own by doing very little change in the input data to balance the dataset. It makes the elastic sample of less data to make it equivalent to the biased data so our dataset can contain both output data types equivalent so that the model can learn effectively.

Let’s understand it with example:

Suppose we have a diabetes dataset which contains two different types of output variable i.e., 0 & 1 where 1 represents diabetic and 0 represents non-diabetic. The dataset contains 2000 rows and 12 columns. When we have checked the output count, we found that there are 1780 people are diabetic and only 220 people are non-diabetic.

So, here we can see that the data is biased towards diabetic and we have to make it balance so that the model can learn effectively both types of data.

Here when we apply over-sampling technique this will make small changes in one of the input column and make an elastic sample of non-diabetic this process will keep repeating until it does not gets equivalent to the diabetic data. i.e., until the dataset gets balanced.

Most of the time we use over-sampling technique to balance the dataset so that we don't face any data loss and train the model effectively.

The following methods are used under the Over-sampling technique:

- SMOTE
- ADASYN
- Random Over Sampler

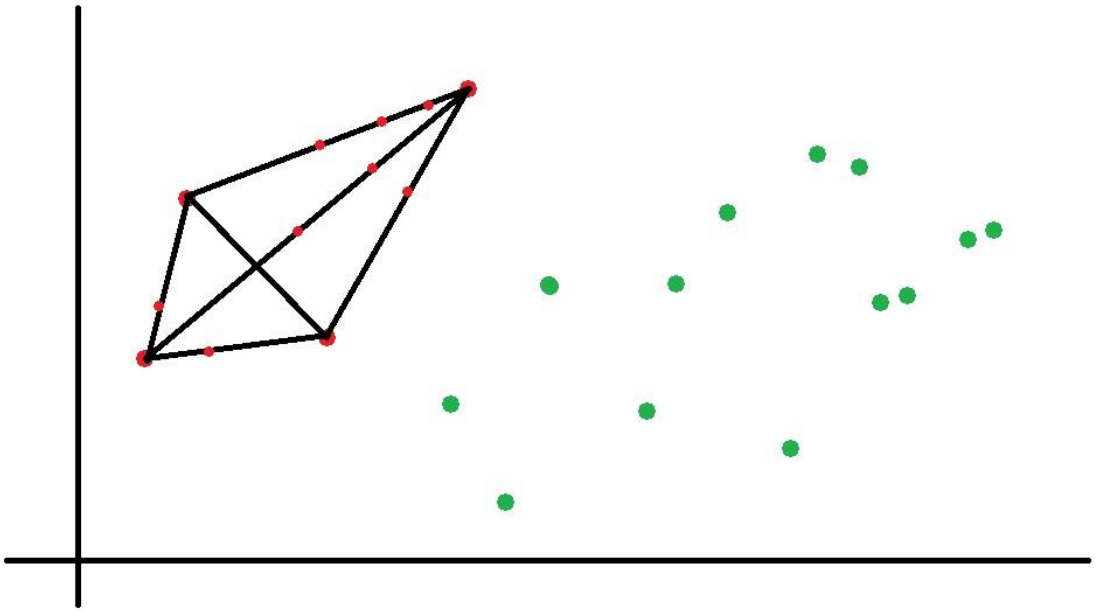
13. What is the difference between SMOTE and ADASYN sampling techniques?

Ans: SMOTE stands for Synthetic Minority Over-Sampling Technique and ADASYN stands for Adaptive Synthetic Sampling. Both are Over-sampling techniques for balancing the dataset. Both make synthetic samples of the minority class to make it equivalent to the majority class. But there is a very small difference in working of both sampling techniques.

SMOTE creates lines between the data of the minority class and generates the samples which lie on those lines.

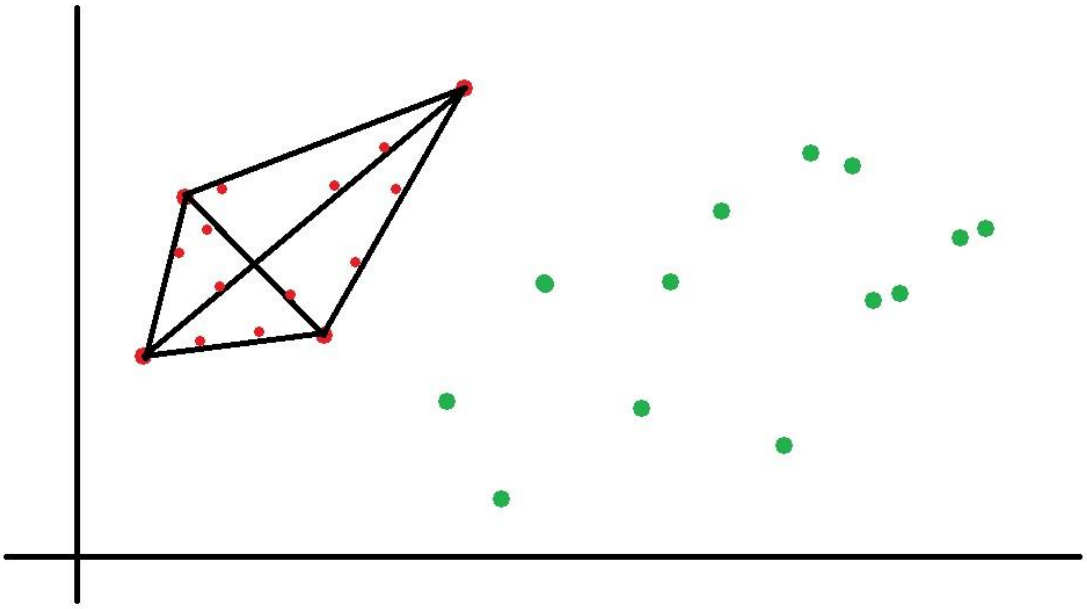
While the samples generated by ADASYN are little shifted from those lines.

Let's understand both with the help of figures:



SMOTE

Here we can see that the lines are drawn between the data of the minority class and the elastic samples generated by SMOTE lies on those lines.



ADASYN

The same is happening in the case of ADASYN but the elastic generated sample data are inside the area and little shifted from those lines.

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

Ans: There are lot of different parameter works behind all the Machine Learning algorithms. Form them some works by default. But sometimes the default parameter don’t tunes properly with the dataset and don’t train properly so at that time we have to change the parameter and try again by training the model.

As there are a lot of parameters behind an algorithm it is complex and very time-consuming if we check by changing every parameter manually one by one. To get this task easily done there is a hyperparameter tuning method called GridSearchCV is used. It takes parameters in dictionary key and value format and then it checks all the parameters one by one of a model. Then it provides the best fit parameter of the model.

The purpose of using GridSearchCV is to reduce the manual parameter tuning operation as well as the Jupyter notebook sheet and to reduce the complexity and confusion between best fit parameters.

In case of large dataset this method is not preferable to use because as it checks every parameter of model for training so it takes its own time to check all. i.e., using this method with large datasets time complexity will increase exponentially, so it is not practically feasible.

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

Ans: There are following evaluation metrics are used to evaluate a regression model:

- Mean absolute error
- Mean squared error
- Root mean squared error
- R2 score

Mean absolute Error: This metric calculates the absolute difference between each absolute and predicted value.

The mean of these absolute differences is represented by mean absolute error.

Mean squared error: This works like mean absolute error with a little bit of change, it calculates the squared difference between actual and predicted values. And then find their mean. It performs the square to avoid the cancellation of negative terms.

For the best fit of model, this should be minimum.

Root mean squared error: It is square root of mean squared error.

R2 score: This score gives information about how close the generated regression line from the actual data points. This is also called Coefficient of Determination. It is find using the following formula.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Where, y in numerator is predicted value – actual value

And y in denominator is actual value – mean of y

The value of R2 score lies between 0 to 1. And it should be closer to 1 for the best performance of regression model.