Machine Learning Worksheet 1

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

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?

A) max_depth

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

D) ADASYN

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.

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

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?

B) Support Vector Machines

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

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?

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?

B) remove only one of the features

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?

A) Overfitting

B) Multicollinearity

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

One-hot Encoding is a feature encoding strategy to convert categorical features into a numerical vector. For each feature value, the one-hot transformation creates a new feature limiting the presence or absence of feature value.

One-hot encoding creates n-dimensional vectors for each instance where n is the unique number of feature values in the dataset. For a feature having a large number of unique feature values or categories, one-hot encoding is not a great choice.

To overcome this, there are other encoding techniques that can be used in such situations such as Label encoder or Binary encoder.

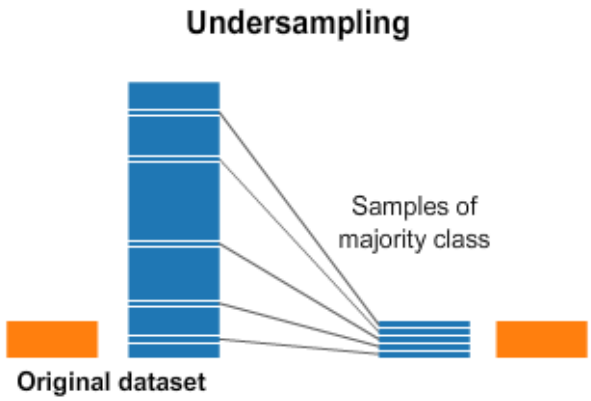
12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.

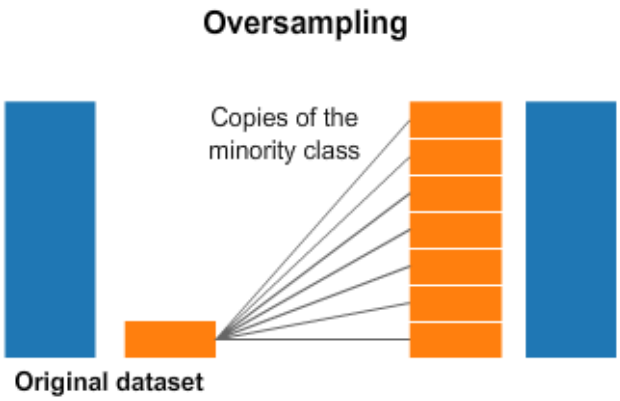
While working with the dataset we may encounter cases where the target variable contains classes that are imbalanced. Feeding such dataset directly to the machine results in inaccurate results. To tackle imbalanced data Resampling technique is used.

Resampling consists of removing samples from the majority class and/or adding more examples from the minority class.

There are two types of resampling techniques- Undersampling and Oversampling.

When samples are removed from the majority class in order to match with the minority class is called Undersampling whereas when more no. of samples are added to the minority class to match with majority class, it is known as Oversampling.





Further, different types of Undersampling and Oversampling are-

1. ***Random Undersampling***
2. ***Random Oversampling***
3. ***Random Undersampling with imblearn***
4. ***Random Oversampling with imblearn***
5. ***Undersampling: Tomek links***
6. ***Synthetic Minority Oversampling Technique (SMOTE)***
7. ***NearMiss***

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

Concept of SMOTE is simple, it first finds n-nearest neighbors in the minority class for each of the samples in the class. Then it draws a line between the neighbors and generates random points on the line.

ADASYN is an improved method of SMOTE, it's concept is very similar to SMOTE but with a minor improvement which is that after randomly creating points on the lines, it adds a random small values to the points. Hence instead of the points being linearly correlated to their parent points as in SMOTE, they are little scattered i.e., they have little variance in them.

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

In Machine Learning, we train different models on the dataset and select the best performing model among them. However, there can be a scope of improvement. This can be accomplished by using hyperparameter tunning. In hyperparameter we need to select appropriate values which can improve the performance of the model. Hence, it is necessary to find the optimal values for the hyperparameters of a model which cannot be done manually as it is very much time consuming. This can be done by GridSearchCV.

GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. Here, we pass predefined values for hyperparameters to the GridsearchCV function by defining a dictionary in which the hyperparameter along with the values is passed.

Incase of large datasets it is not preferable to use GridSearchCV as this is an exhaustive function. As we know, we pass certain values for hyperparameter tunning. GridSearchCV tries all the combinations and possibilities with every value provided which can result in millions of possibilities and it will take much long time to finish the task. It alse follows certain sequence without considering the past experience. Every

point in GridSearchCV needs k-fold cross-validation which requires k training steps. So, tuning of hyperparameters in this way can be quite complex, time consuming and expensive.

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

There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model:

- ***Mean Squared Error (MSE).***
- ***Root Mean Squared Error (RMSE).***
- ***Mean Absolute Error (MAE)***

Mean Squared Error (MSE):

The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.

Root Mean Squared Error (RMSE):
It is an extension of the mean squared error. the square root of the Mean Squared Error error is calculated.

Mean Absolute Error (MAE):
MAE score is calculated as the average of the absolute error values. Absolute or abs() ***is a mathematical function that simply makes a number positive. Therefore, the difference between an expected and predicted value may be positive or negative and is forced to be positive when calculating the MAE.***