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?

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?

A) SMOTE

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

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

B) 2-1-3

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)

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?

C) Ridge will cause some of the coefficients to become 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?

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?

A) Overfitting

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?

When the categorical features present in the dataset are ordinal i.e for the data being like Junior, Senior, Executive, Owner. When the number of categories in the dataset is quite large. One Hot Encoding should be avoided in this case as it can lead to high memory consumption

One-Hot encoding technique is used when the features are nominal(do not have any order). In one hot encoding, for every categorical feature, a new variable is created. Categorical features are mapped with a binary variable containing either 0 or 1

To fight the curse of dimensionality, **binary encoding** might be a good alternative to one-hot encoding because it creates fewer columns when encoding categorical variables. Ordinal encoding is a good choice if the order of the categorical variables matters

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

Resampling (Oversampling and Undersampling)

This technique is used to upsample or downsample the minority or majority class. When we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling

Oversampling methods duplicate or create new synthetic examples in the minority class, whereas undersampling methods delete or merge examples in the majority class. Both types of resampling can be effective when used in isolation, although can be more effective when both types of methods are used together.

Oversampling — Duplicating samples from the minority class. Undersampling — Deleting samples from the majority class.

Undersampling and oversampling are techniques used to combat the issue of unbalanced classes in a dataset. We sometimes do this in order to avoid overfitting the data with a majority class at the expense of other classes whether it's one or multiple.

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

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

GridSearchCV is a technique for **finding the optimal parameter values from a given set of parameters in a grid**. It's essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made.

GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. As mentioned above, the performance of a model significantly depends on the value of hyperparameters.

- Advantages: exhaustive search, will find the absolute best way to tune the hyperparameters based on the training set.
- Disadvantages: time-consuming, danger of overfitting.

Methods for Hyperparameters Tuningused in acse of large dataset.

For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible. One can shift to Random Search CV where the algorithm will randomly choose the combination of parameters

For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible. One can shift to Random Search CV where the algorithm will randomly choose the combination of parameters

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; they are:

Mean Squared Error (MSE).

Root Mean Squared Error (RMSE).

Mean Absolute Error (MAE)

Mean Squared Error (MSE) :

The Mean Squared Error **measures how close a regression line is to a set of data points**. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.

The Mean Squared Error measures how close a refression line is to set a data points. It is a risk function corresponding to the expected value of the squared error loss.

Mean square Error is calculated by taking the average ,specifically the mean, of errors squared from data as it relates to a function .

Root Mean Squared Error (RMSE).

What is Root Mean Square Error (RMSE)? Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of **the differences between values (sample or population values) predicted by a model or an estimator and the values observed**.

Mean Absolute Error (MAE) :

The Mean absolute error is calculated by **adding up all the absolute errors and dividing them by the number of errors**.

Absolute Error is the amount of error in your measurements. It is the difference between the measured value and " true" value.

Mean Squared Error (MSE) :

MSE, take the observed value, subtract the predicted value, and square that difference. Repeat that for all observations. Then, sum all of those squared values and divide by the number of observations.

The Mean Squared Error **measures how close a regression line is to a set of data points**