**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

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

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

A) SMOTE B) RandomOverSampler

C) RandomUnderSampler 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. A) 1 and 2 B) 1 only

C) 1 and 3 D) 2 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

A) 3-1-2 B) 2-1-3

C) 3-2-1 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?

A) Decision Trees B) Support Vector Machines

C) K-Nearest Neighbors D) Logistic Regression

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

**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.

**MACHINE LEARNING**

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

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

**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?

If the cardinality (the # of categories) of the categorical features is low (relative to the amount of data) one-hot encoding will work best. You can use it as input into any model.But if the cardinality is large and your dataset is small, one-hot encoding may not be feasible, and you model may not be able to efficiently learn. An alternative is to try an approach like numeric or binary encoding [1], and hope that your model can learn something useful from this, at the risk of overfitting. Or you can use a model like LightGBM, which can directly handle categorical variables, by using an alternative method for identifying variable splits.

12. In case of data imbalance problem in classification, what techniques can be used to balance the

dataset? Explain them briefly.

* Oversampling Techniques

Oversampling methods duplicate examples in the minority class or synthesize new examples from the examples in the minority class.Some of the more widely used and implemented oversampling methods include:

* Random Oversampling
* Synthetic Minority Oversampling Technique (SMOTE)
* Borderline-SMOTE
* Borderline Oversampling with SVM
* Adaptive Synthetic Sampling (ADASYN)
* Undersampling Techniques

Undersampling methods delete or select a subset of examples from the majority class. Some of the more widely used and implemented undersampling methods include:

* Random Undersampling
* Condensed Nearest Neighbor Rule (CNN)
* Near Miss Undersampling
* Tomek Links Undersampling
* Edited Nearest Neighbors Rule (ENN)
* One-Sided Selection (OSS)
* Neighborhood Cleaning Rule (NCR)
* Combinations of Techniques

Although an oversampling or undersampling method when used alone on a training dataset can be effective, experiments have shown that applying both types of techniques together can often result in better overall performance of a model fit on the resulting transformed dataset. Some of the more widely used and implemented combinations of data sampling methods include:

* SMOTE and Random Undersampling
* SMOTE and Tomek Links
* SMOTE and Edited Nearest Neighbors Rule

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

* SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample as a point along that line.
* Adaptive Synthetic Sampling (ADASYN) is another extension to SMOTE that generates synthetic samples inversely proportional to the density of the examples in the minority class. It is designed to create synthetic examples in regions of the feature space where the density of minority examples is low, and fewer or none where the density is high.

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

GridSearchCV 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. GridSearchCV will go through all the intermediate combinations of hyperparameters which makes grid search computationally very expensive.

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

Regression metrics help us to measure the goodness of the fit of our model.We have various regression evaluation metrics to measure how well our model fits the data.Below we will see some of the most commonly used metrics to asses the regression model.

* + MEAN SQUARED ERROR: The first metric we are going to see is the mean squared error.It calculates the average of the square of the errors between the actual and the predicted values.Lower the value, better the regression model.
  + ROOT MEAN SQUARED ERROR: RMSE is the most popular metric to measure the error of a regression model. This metric is calculated as the square root of the average squared distance between the actual and the predicted values. Taking the square root of the mean squared error will give you RMSE. Since we are, taking the square root it reverts the unit of measurement to its original scale.
  + MEAN ABSOLUTE PERCENTAGE ERROR(MAPE): MAPE measures the error in percentage terms. MAPE is calculated as the absolute difference between the actual and predicted values divide over every observation. It is multiplied by 100 to make it a percentage error. Where n is the size of the sample, ŷt is the value predicted by the model, and yt is the actual value. The problem with MSE is that since the values are squared the unit of measurement is changed. To fill this deficiency, we looked at another metric called RMSE, which reverts the value to its original unit of measurement by taking a square root. MAPE can be used to compare two models of different scales.