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?

   Answer: 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?

   Answer: A) max_depth

****
****

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

   Answer: C) Random Under Sampler

****
****

4) Which of the following statements is/are true about "Type-1" and "Type-2" errors?

   Answer: C) 1 and 3

****
****

5) Arrange the steps of k-means algorithm in the order in which they occur:

   Answer: 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?

Answer: B) Support Vector Machines

****
****

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

Answer: 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?

Answer: B) Lasso 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?

Answer: 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?

Answer: A) Overfitting

B) Multicollinearity

****
****

Q11 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?

Answer: We use this categorical data encoding technique when the features are nominal(do not have any order). In one hot encoding, for each level of a categorical feature, we create a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category. These newly created binary features are known as Dummy variables. The number of dummy variables depends on the levels present in the categorical variable.

****
****

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

Answer:

In Imbalanced dataset first we need to think what we require to oversampling or down sampling as per that we need to use the sampling technique .There are different sampling techniques to use, mostly people prefer SMOTE to balance the data.

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:

1. Random Oversampling
2. Synthetic Minority Oversampling Technique (SMOTE)
3. Borderline-SMOTE
4. Borderline Oversampling with SVM
5. 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:

1. Random Undersampling
2. Condensed Nearest Neighbor Rule (CNN)
3. Near Miss Undersampling
4. Tomek Links Undersampling
5. Edited Nearest Neighbors Rule (ENN)
6. One-Sided Selection (OSS)
7. Neighbourhood 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:

1. SMOTE and Random Undersampling
2. SMOTE and Tomek Links
3. SMOTE and Edited Nearest Neighbors Rule

****
****

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

Answer:

- **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.

The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed.

****
****

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

Answer: GridSearchCV is used to find the best parameters out of given parameters on the best score which is defined in its definition. So it is a method of tuning the parameters of given estimator.

Since GridSearch takes into account each and every parameter combination to check for the best result therefore in case of large datasets it will be very time taking task to get best parameters using Grid search instead RandomSearch will reduce the time complexity since it takes random combinations of the parameters.

****
****

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

Answer:

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 assess 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 it 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.

****
****