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) 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** B) n_estimators
   C) min_samples_leaf D) min_samples_splits

Ans:- (A) max_depth

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

Ans:- **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

Ans:- **B) 1 only**

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) 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

Ans:- **B) Support Vector Machines**

MACHINE LEARNING

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 createmultiway trees (more than two children for a node)
   D) None of the above

Ans:- (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?
   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:- (C) Ridge 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

Ans:- B)remove only one of the features

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:- C) Underfitting

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 insuch a case?

Ans- 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. Dummy technique should be used in such case because this categorical data encoding method transforms the categorical variable into a set of binary variables (also known as dummy variables). In the case of one-hot encoding, for N categories in a variable, it uses N binary variables. The dummy encoding is a small improvement over one-hot-encoding.

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

Ans :- Two approaches to make a balanced dataset out of an imbalanced are Resampling means one are under-sampling and over-sampling. This technique is used to up sample or down sample 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. Similarly, we can randomly delete rows from the majority class to match them with the minority class which is called under sampling. After sampling the data we can get a balanced dataset for both majority and minority classes. So, when both classes have a similar number of records present in the dataset, we can assume that the classifier will give equal importance to both classes.

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

Ans:- SMOTE: Synthetic Minority Over sampling Technique (SMOTE) algorithm applies KNN approach where it selects K nearest neighbors, joins them and creates the synthetic samples in the space. The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors. The difference is multiplied by random number between (0, 1) and it is added back to feature. SMOTE algorithm is a pioneer algorithm and many other algorithms are derived from SMOTE.

ADAptive SYNthetic (ADASYN) is based on the idea of adaptively generating minority data samples according to their distributions using K nearest neighbor. The algorithm adaptively updates the distribution and there are no assumptions made for the underlying distribution of the data. The algorithm uses Euclidean distance for KNN Algorithm. 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 distributions. The latter generates the same number of synthetic samples for each original minority sample.

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

Ans:- GridSearchCV is a technique to search through the best parameter values from the given set of the grid of parameters. It is basically a cross-validation method. the model and the parameters are required to be fed in. Best parameter values are extracted and then the predictions are made. 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. Thus Grid Search CV technique is not recommended for large-size datasets or param grids with a large number of components. For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible.

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

Ans:- 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 Squad Error:-

Mean Squared Error, or MSE for short, is a popular error metric for regression problems.
It is also an important loss function for algorithms fit or optimized using the least squares framing of a regression problem. Here “least squares” refers to minimizing the mean squared error between predictions and expected values.
The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.
MSE = 1 / N * sum for i to N (y_i – yhat_i)^2

Root Mean Squared Error

The Root Mean Squared Error, or RMSE, is an extension of the mean squared error.
Importantly, the square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted.
For example, if your target variable has the units “dollars,” then the RMSE error score will also have the unit “dollars” and not “squared dollars” like the MSE.
As such, it may be common to use MSE loss to train a regression predictive model, and to use RMSE to evaluate and report its performance.

The RMSE can be calculated as follows:
RMSE = sqrt(1 / N * sum for i to N (y_i – yhat_i)^2)
Where y_i is the i’th expected value in the dataset, yhat_i is the i’th predicted value, and sqrt() is the square root function.

We can restate the RMSE in terms of the MSE as:
RMSE = sqrt(MSE)

MACHINE LEARNING

Mean Absolute Error

Mean Absolute Error, or MAE, is a popular metric because, like RMSE, the units of the error score match the units of the target value that is being predicted.

Unlike the RMSE, the changes in MAE are linear and therefore intuitive.

That is, MSE and RMSE punish larger errors more than smaller errors, inflating or magnifying the mean error score. This is due to the square of the error value. The MAE does not give more or less weight to different types of errors and instead the scores increase linearly with increases in error.

As its name suggests, the 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.

The MAE can be calculated as follows:

MAE = 1 / N * sum for i to N abs(y_i – yhat_i)
Where y_i is the i'th expected value in the dataset, yhat_i is the i'th predicted value and abs() is the absolute function.

We can create a plot to get a feeling for how the change in prediction error impacts the MAE.

The example below gives a small contrived dataset of all 1.0 values and predictions that range from perfect (1.0) to wrong (0.0) by 0.1 increments. The absolute error between each prediction and expected value is calculated and plotted to show the linear increase in error.