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?

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. This might sound complicated.

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

This is most famous and important topics in data science. If you have spent some time in the area, you would have definitely come across imbalanced class distribution. This is a scenario where the number of observations belonging to one class is significantly lower than those belong to the other class.It comes problem when we try to predict lower ratio class. In general, Machine learning algorithms are designed to improve accuracy by reducing the error. However, When we build model to predict lower number of class, It will go fail and doesn't predict that class. That assumes it as noise and ignore it . Even it comes worse when we check accuracy of model , it shows more than good like 96.2 %. So , what will we have strategy to handle like these imbalanced data .Maybe there are too many strategy , but here i only try to say that how to handle simple way and efficiently these data without going all of theories . So, There are some points below

-Challenges of Imbalanced Classification

-Confusion Matrix, precision, recall, and F1

-Approach to handling Imbalanced Datasets.

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

The major difference between SMOTE and ADASYN is the difference in the generation of synthetic sample points for minority data points. In ADASYN, we consider a density distribution r_x which thereby decides the number of synthetic samples to be generated for a particular point, whereas in SMOTE, there is a uniform weight for all minority points.

MACHINE LEARNING

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

Cross-validation is the process of splitting the same dataset in K-partitions, and for each split, we search the whole grid of hyperparameters to an algorithm, in a brute force manner of trying every combination.

Note that I'm referring to K-Fold cross-validation (CV), even though there are other methods of doing CV.

In an iterative manner, we switch up the testing and training dataset in different subsets from the full dataset. We usually split the full dataset so that each testing fold has of the full dataset.

rom this image of cross-validation, what we do for the grid search is the following; for each iteration, test all the possible combinations of hyperparameters, by fitting and scoring each combination separately.

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

Confusion Matrix
F1 Score
Gain and Lift Charts
Kolmogorov Smirnov Chart
AUC – ROC
Log Loss
Gini Coefficient
Concordant – Discordant Ratio
Root Mean Squared Error
Cross Validation