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 thedata?
   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 andtime, 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 (ChiSquare 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 createbinary 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

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), whichof 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. Whatare 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 insuch a case?
   Answer: One-hot encoding should be avoided when dealing with high cardinality categorical features, where the number of unique categories is very large. In such cases, it can result in a high number of new features, leading to the curse of dimensionality and may cause performance issues for some models. In such cases, entity embedding techniques can be used for encoding categorical feature

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

Answer: In case of data imbalance problem in classification, some techniques that can be used to balance the dataset are:

- **Random Undersampling:** In this technique, we randomly select some observations from the majority class and remove them from the dataset, so that the number of observations in the majority class is reduced and becomes closer to the number of observations in the minority class.

- **Random Oversampling:** In this technique, we randomly replicate some observations from the minority class to increase the number of observations in the minority class, so that it becomes closer to the number of observations in the majority class.

- **Synthetic Minority Oversampling Technique** or **SMOTE** is another technique to oversample the minority class. Simply adding duplicate records of minority class often don't add any new information to the model. In SMOTE new instances are synthesized from the existing data. If we explain it in simple words, SMOTE looks into minority class instances and use *k* nearest neighbour to select a random nearest neighbor, and a synthetic instance is created randomly in feature space.

- **BalancedBaggingClassifier :** When we try to use a usual classifier to classify an imbalanced dataset, the model favors the majority class due to its larger volume presence. A BalancedBaggingClassifier is the same as a sklearn classifier but with additional balancing. It includes an additional step to balance the training set at the time of fit for a given sampler. This classifier takes two special parameters "sampling_strategy" and "replacement". The **sampling_strategy** decides the type of resampling required (e.g. 'majority' – resample only the majority class, 'all' – resample all classes, etc) and **replacement** decides whether it is going to be a sample with replacement or not.

MACHINE LEARNING

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

Answer:

SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are two popular oversampling techniques used to address the class imbalance problem in machine learning.

The main difference between SMOTE and ADASYN is that SMOTE generates synthetic samples by interpolating between existing minority class samples, while ADASYN creates synthetic samples by focusing on the samples that are difficult to classify correctly. Specifically, ADASYN places more emphasis on the samples that are closer to the decision boundary of the classifier and generates synthetic samples in those regions.

Another difference between the two techniques is that ADASYN tends to generate more samples than SMOTE, as it uses a density distribution function to determine the number of synthetic samples to be generated for each minority class sample. This makes ADASYN more suitable for datasets with severe class imbalance.

Overall, while both SMOTE and ADASYN are effective in addressing class imbalance, ADASYN may perform better in certain scenarios, especially when the class imbalance is severe and the decision boundary is complex.

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

Answer: GridSearchCV is a technique used in machine learning to tune hyperparameters of a model by exhaustively searching over a specified parameter grid. It performs an exhaustive search over all possible combinations of hyperparameters to find the optimal set of hyperparameters that results in the best performance of the model on a given evaluation metric

The main purpose of using GridSearchCV is to find the best hyperparameters for a given model that maximize the performance of the model on the given data. This is important because different combinations of hyperparameters can lead to significantly different performance of the model, and finding the best hyperparameters can improve the accuracy and generalizability of the model.

Whether or not to use GridSearchCV for large datasets depends on the specific situation. On one hand, GridSearchCV can be computationally expensive, as it performs an exhaustive search over a parameter grid, which can be time-consuming for large datasets or models with many hyperparameters. In such cases, a randomized search or a Bayesian optimization technique might be a better option, as they can explore the hyperparameter space more efficiently.

On the other hand, in some cases, the performance of a model on a large dataset can be sensitive to the choice of hyperparameters, and finding the optimal hyperparameters can have a significant impact on the performance of the model. In such cases, using GridSearchCV to exhaustively search the hyperparameter space might be necessary, even if it is computationally expensive.

In summary, the decision to use GridSearchCV for large datasets depends on the specific situation, and other hyperparameter tuning techniques may be more suitable depending on the computational resources available and the sensitivity of the model to hyperparameters

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

Regression models are used to predict continuous numerical values, and they are evaluated based on their ability to accurately predict these values. Here are some of the commonly used evaluation metrics for regression models:

Mean Squared Error (MSE): The MSE measures the average squared difference between the predicted values and the actual values. It is calculated by taking the average of the squared differences between each predicted and actual value. The lower the MSE, the better the model performs.

MACHINE LEARNING

Root Mean Squared Error (RMSE): The RMSE is the square root of the MSE and it measures the average distance between the predicted values and the actual values. Like MSE, a lower RMSE indicates a better performing model.

Mean Absolute Error (MAE): The MAE is the average absolute difference between the predicted and actual values. It is calculated by taking the average of the absolute differences between each predicted and actual value. MAE is less sensitive to outliers than MSE and RMSE, but it does not penalize large errors as heavily.

R-squared (R2): The R-squared value measures the proportion of variance in the dependent variable that is explained by the independent variable(s). It ranges from 0 to 1, with a value of 1 indicating that the model explains all of the variance in the dependent variable.

Mean Absolute Percentage Error (MAPE): The MAPE measures the average percentage difference between the predicted and actual values. It is calculated by taking the average of the absolute percentage differences between each predicted and actual value. MAPE is useful for comparing models with different units and scales, but it is sensitive to outliers and can be undefined when the actual value is zero.

Coefficient of Determination (COD): It measures the goodness of fit of the model. It ranges between 0 to 1. The higher the COD value, the better the model is at predicting the target variable.

These metrics are used to evaluate the performance of a regression model and to compare it to other models to determine which one performs the best.