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. 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?
12. In case of data imbalance problem in classification, what techniques can be used to balance thedataset? Explain them briefly.
13. What is the difference between SMOTE and ADASYN sampling techniques?
14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why orwhy not?
15. List down some of the evaluation metric used to evaluate a regression model. Explain each of themin brief.

11.

Insituation One-hot encoding must be avoided? Which encoding technique can be used in such a case?

One-hot encoding is a popular technique used in machine learning for transforming categorical data into numerical data. However, there are situations where one-hot encoding might not be the best option, such as:

High Cardinality Features: One-hot encoding can be problematic when dealing with high cardinality categorical features. This is because one-hot encoding creates a new binary column for each unique category, which can lead to a high-dimensional sparse feature matrix and cause memory and computational issues. In such cases, feature hashing or entity embedding techniques can be used as an alternative to one-hot encoding.

Ordinal Data: One-hot encoding treats all categories as independent and equally important, which is not always true for ordinal data. For example, if a feature has categories like "low," "medium," and "high," one-hot encoding does not take into account the order and relationship between these categories. In such cases, label encoding can be used to encode the categories with numerical values based on their order.

Imbalanced Data: One-hot encoding can also be problematic when dealing with imbalanced data, where some categories have a much higher frequency than others. In such cases, encoding the categories based on their frequency or using target encoding techniques can be more useful for improving model performance.

In summary, while one-hot encoding is a powerful encoding technique for categorical data, it might not always be the best option. It's essential to consider the nature of the data and the problem at

MACHINE LEARNING

hand when choosing an appropriate encoding technique.

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

Data imbalance is a common problem in classification tasks where one or more classes have significantly fewer samples than the others. This can lead to biased models that perform poorly on the minority classes. Here are some techniques to balance the dataset:

Under sampling: This technique involves removing samples from the majority class to balance the class distribution. Random undersampling removes random samples from the majority class, while intelligent undersampling techniques use heuristics or algorithms to select samples to remove.

Over sampling: This technique involves generating synthetic samples for the minority class to balance the class distribution. Random oversampling duplicates samples from the minority class, while intelligent oversampling techniques generate new samples based on the existing minority class samples.

Synthetic Minority Over-sampling Technique (SMOTE): SMOTE is a popular oversampling technique that generates synthetic samples by interpolating between minority class samples. It selects a minority sample and finds its k-nearest neighbors. It then generates synthetic samples by interpolating between the minority sample and its neighbors.

Ensemble Techniques: Ensemble techniques such as bagging, boosting, and stacking can be used to balance the class distribution. Ensemble techniques involve combining multiple models to improve performance. For example, you can use bagging to train multiple models on random subsets of the majority class samples and combine their predictions.

Cost-sensitive learning: Cost-sensitive learning involves adjusting the cost function to give more weight to the minority class samples. This can be done by assigning a higher misclassification cost to the minority class samples.

In summary, there are various techniques to balance the dataset in classification tasks. Choosing the appropriate technique depends on the nature of the data and the problem at hand. A combination of multiple techniques can also be used to achieve better results.

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

Both **SMOTE (Synthetic Minority Over-sampling Technique)** and **ADASYN (Adaptive Synthetic Sampling)** are oversampling techniques used to handle class imbalance problems in classification tasks. However, there are some differences between them:

SMOTE generates synthetic samples by interpolating between minority class samples and their nearest neighbors. It selects a minority sample and finds its k-nearest neighbors in the feature space. It then generates synthetic samples by interpolating between the minority sample and its neighbors. The

MACHINE LEARNING

number of synthetic samples generated is equal to a user-defined ratio of the difference between the number of samples in the majority and minority classes.

ADASYN, on the other hand, generates synthetic samples by using a density distribution-based approach. It focuses on generating synthetic samples in the regions that are difficult to learn by the classifier. It calculates the density distribution of minority samples in each feature space and generates synthetic samples proportionally to the density distribution. This means that it generates more synthetic samples for the minority class samples that are more difficult to learn.

Another difference between **SMOTE and ADASYN** is that ADASYN introduces a random perturbation into each synthetic sample to make it more diverse than SMOTE. This random perturbation helps the classifier to avoid overfitting to the synthetic samples and makes it more robust to noise.

In summary, while both SMOTE and ADASYN are oversampling techniques used to handle class imbalance problems, they differ in their approach to generating synthetic samples. SMOTE generates synthetic samples by interpolating between minority class samples and their nearest neighbors, while ADASYN generates synthetic samples based on the density distribution of the minority class samples in each feature space. Additionally, ADASYN introduces a random perturbation to make the synthetic samples more diverse.

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

GridSearchCV is a function in scikit-learn library that is used to perform hyperparameter tuning of a machine learning model. It exhaustively searches all possible combinations of hyperparameters specified in a grid and evaluates the model performance using cross-validation. The purpose of using GridSearchCV is to find the best combination of hyperparameters that results in the highest performance of the model.

GridSearchCV is a widely used technique for hyperparameter tuning, but it can be computationally expensive, especially for large datasets. This is because it requires fitting the model multiple times for each combination of hyperparameters and cross-validation folds, which can be time-consuming and resource-intensive. For very large datasets, it may not be practical to use GridSearchCV because of the computational cost.

However, **GridSearchCV** can still be useful in some cases, even for large datasets. For example, if the model has a relatively small number of hyperparameters, GridSearchCV may still be feasible to use, especially if the number of possible values for each hyperparameter is also small. In addition, GridSearchCV can help identify the most promising hyperparameters, which can then be further optimized using other techniques such as randomized search or Bayesian optimization.

Overall, the use of GridSearchCV depends on the specific problem and dataset at hand. While it is a powerful technique for hyperparameter tuning, its feasibility and usefulness depend on the computational resources available and the complexity of the model and dataset.

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

MACHINE LEARNING

There are several evaluation metrics used to assess the performance of regression models. Here are some commonly used metrics and a brief explanation of each:

Mean Squared Error (MSE): MSE is the most widely used metric for evaluating regression models. It measures the average squared difference between the predicted and actual values. It is calculated as the average of the squared differences between the predicted and actual values of the target variable.

Root Mean Squared Error (RMSE): RMSE is similar to MSE, but it takes the square root of the MSE to make the units of the error the same as the target variable. It is calculated as the square root of the average of the squared differences between the predicted and actual values of the target variable.

Mean Absolute Error (MAE): MAE is another commonly used metric for evaluating regression models. It measures the average absolute difference between the predicted and actual values. It is calculated as the average of the absolute differences between the predicted and actual values of the target variable.

R-squared (R^2): R-squared is a metric that measures the proportion of variance in the target variable that is explained by the model. It ranges from 0 to 1, with a higher value indicating a better fit. It is calculated as the ratio of the explained variance to the total variance.

Mean Absolute Percentage Error (MAPE): MAPE is a metric that measures the average percentage difference between the predicted and actual values. It is calculated as the average of the absolute differences between the predicted and actual values of the target variable, divided by the actual value.

Coefficient of Determination (COD): COD is a metric that measures the proportion of total variation in the dependent variable that is explained by the independent variables in the regression model. It ranges from 0 to 1, with a higher value indicating a better fit. It is calculated as the ratio of the explained variance to the total variance.

Explained Variance Score (EVS): EVS is a metric that measures the proportion of variance in the target variable that is explained by the model. It ranges from 0 to 1, with a higher value indicating a better fit. It is calculated as the ratio of the explained variance to the total variance of the target variable.

In summary, there are several evaluation metrics that can be used to assess the performance of regression models. The choice of the metric depends on the specific problem and the nature of the dataset. MSE, RMSE, and MAE are commonly used error metrics, while R^2, MAPE, COD, and EVS are commonly used goodness-of-fit metrics.