MACHINE LEARNING ASSIGNMENT - 8

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.

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?

Answer:- One-hot encoding may not be the best choice when dealing with high cardinality categorical features, i.e. categorical features with a large number of unique values. This is because one-hot encoding can result in a very high-dimensional and sparse feature space, which can be computationally expensive and may lead to overfitting.

In such cases, an alternative encoding technique that can be used is target encoding, also known as mean encoding. Target encoding involves replacing each categorical value with the mean of the target variable for that value. This can be an effective way to represent categorical variables as numerical features, without creating a high-dimensional and sparse feature space.

Target encoding works particularly well when there is a strong relationship between the categorical variable and the target variable. However, it is important to be careful when using target encoding, as it can result in overfitting if the relationship between the categorical variable and the target variable is not strong enough or if there are too few samples for some categories. In such cases, cross-validation can be used to estimate the performance of the model and to detect overfitting.

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

Answer:- Data imbalance occurs when the classes in a classification problem have significantly different number of samples. In such cases, the classifier may be biased towards the majority class and perform poorly on the minority class. There are several techniques that can be used to balance the dataset:

1: Undersampling: In this technique, some samples from the majority class are randomly selected and removed from the dataset until the class distribution becomes balanced. However, undersampling can result in the loss of important information from the majority class.

2: Oversampling: In this technique, new synthetic samples are generated for the minority class to match the number of samples in the majority class. The most commonly used oversampling technique is Synthetic Minority Over-sampling Technique (SMOTE), which generates new synthetic samples by interpolating between neighboring minority class samples. Oversampling can be effective in balancing the dataset, but it can also result in overfitting.

3: Cost-sensitive learning: In this technique, the misclassification cost of the minority class is given more weight during training to make the classifier more sensitive to the minority class. This can be achieved by modifying the loss function or by assigning weights to each sample based on its class.

4: Ensemble methods: In this technique, multiple classifiers are trained on different subsets of the dataset and their predictions are combined to make the final prediction. Ensemble methods such as Bagging and Boosting can be used to balance the dataset by training the classifiers on different subsets of the dataset.

5: Anomaly detection: In this technique, the minority class is treated as an anomaly and a separate model is trained to detect the anomalies. This can be effective if the minority class represents a rare event or a critical outcome.

It is important to choose the appropriate technique based on the problem at hand and to evaluate the performance of the classifier using appropriate metrics such as precision, recall, F1-score and AUC-ROC. It is also important to be careful when using data balancing techniques, as they can result in overfitting, underfitting or bias if not used properly. Cross-validation and hyperparameter tuning can be used to optimize the performance of the classifier.

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 in machine learning to handle class imbalance. Both methods create synthetic examples of the minority class to balance the class distribution.

The main difference between SMOTE and ADASYN is how they generate synthetic examples. SMOTE generates synthetic examples by interpolating between existing examples of the minority class. Specifically, SMOTE selects two examples of the minority class and creates a new example along the line connecting the two examples. The degree of interpolation is controlled by a parameter called the sampling ratio.

ADASYN, on the other hand, uses a density distribution to generate synthetic examples. The density distribution is estimated based on the ratio of the number of examples of the minority class to the number of examples of the majority class. The synthetic examples are generated with higher density in regions of the feature space where the density of the minority class is lower.

In summary, while both SMOTE and ADASYN are oversampling techniques used to handle class imbalance, SMOTE generates synthetic examples by interpolating between existing examples, while ADASYN generates synthetic examples based on the density distribution of the minority class in the feature space.

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 a technique used in machine learning to tune the hyperparameters of a model. It performs an exhaustive search over a specified

parameter grid and returns the best set of hyperparameters that optimize a chosen performance metric.

The purpose of using GridSearchCV is to automate the hyperparameter tuning process and save time and effort. Instead of manually trying different combinations of hyperparameters, GridSearchCV allows us to define a range of values for each hyperparameter and automatically search over all possible combinations.

GridSearchCV can be useful for large datasets because it saves time and effort in the hyperparameter tuning process. However, it may not always be preferable for large datasets because it can be computationally expensive. GridSearchCV performs an exhaustive search over all possible combinations of hyperparameters, which can be time-consuming and resource-intensive for large datasets.

In cases where the dataset is too large for GridSearchCV, other techniques such as random search or Bayesian optimization can be used instead. These methods are more efficient and require fewer computations compared to GridSearchCV. However, they may not always find the optimal set of hyperparameters. Therefore, it is important to carefully choose the appropriate hyperparameter tuning method based on the size of the dataset, the complexity of the model, and the available computational resources.

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

Answer:- Mean Squared Error (MSE): It is one of the most commonly used evaluation metrics for regression models. It calculates the average squared difference between the predicted and actual values. A lower MSE value indicates better model performance.

Root Mean Squared Error (RMSE): This is the square root of the MSE. It is also a widely used metric in regression analysis. The RMSE is used to measure the difference between the predicted and actual values in the same units as the target variable. A lower RMSE indicates better model performance.

Mean Absolute Error (MAE): MAE calculates the average absolute difference between the predicted and actual values. Unlike MSE, it is not sensitive to outliers. A lower MAE indicates better model performance.

R-Squared (R2): R2 is a metric that measures the proportion of the variance in the target variable that is explained by the regression model. It takes values between 0 and 1, with higher values indicating better model performance.

Adjusted R-Squared (Adjusted R2): Adjusted R2 is similar to R2 but takes into account the number of predictor variables in the model. It is adjusted for the degree of freedom in the model. A higher Adjusted R2 value indicates better model performance.

Mean Absolute Percentage Error (MAPE): This metric measures the average percentage difference between the predicted and actual values. It is particularly useful for evaluating models that are used for forecasting or prediction. A lower MAPE indicates better model performance.

Mean Percentage Error (MPE): MPE is similar to MAPE but measures the average percentage difference between the predicted and actual values, without considering the absolute values. It is useful for understanding the direction of the error. A lower MPE indicates better model performance.

Mean Absolute Scaled Error (MASE): MASE is a metric that measures the accuracy of the forecast of a time series model. It compares the forecast error of the model to the forecast error of a naive model that uses the previous value as the forecast. A lower MASE indicates better model performance.