MACHINE LEARNING ASSIGNMENT – 8

**In Q-1 to Q-7, 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 Q-8 to Q-10, 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

Q-10 to Q-15 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?

Ans: One-hot encoding is a commonly used technique to convert categorical variables into a format that can be used for machine learning algorithms. However, there are situations where one-hot encoding may not be the best choice.

One such situation is when the categorical variable has a large number of categories. One-hot encoding can lead to a very high-dimensional and sparse feature space, which can be computationally expensive and may require a lot of memory. For example, if you have a categorical variable with 1000 categories, one-hot encoding would create 1000 new features.

In such situations, an alternative encoding technique that can be used is called "Hashing Trick" or "Hashing Encoding". This technique involves hashing the categorical variables into a fixed number of features instead of creating a new feature for each category. The hashing function maps the categories to a fixed number of buckets, and each bucket represents a feature. The number of buckets is typically much smaller than the number of categories, so this technique can lead to a more compact representation of the data.

However, it is important to note that this technique may result in collisions, where different categories are mapped to the same bucket. Collisions can lead to a loss of information and may affect the performance of the model. Therefore, it is important to choose the number of buckets carefully to minimize collisions.

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

Ans: Imbalanced datasets can be a common problem in classification tasks where one class has significantly fewer samples than the other(s). This can lead to poor performance of the classifier, as it tends to favor the majority class and ignores the minority class. To address this issue, several techniques can be used to balance the dataset. Here are some of the most commonly used techniques

1. Undersampling: In this technique, a random subset of the majority class samples is removed from the dataset to balance the class distribution. The main drawback of undersampling is that it may lead to a loss of information, especially if the dataset is already small.

2. Oversampling: In this technique, new samples are generated for the minority class to balance the class distribution. The most popular oversampling technique is Synthetic Minority Oversampling Technique (SMOTE), which creates synthetic samples by interpolating between minority class samples.

3Class weighting: This technique involves assigning higher weights to the minority class samples to make them more important during training. This ensures that the classifier pays more attention to the minority class, even though it is outnumbered.

1. Cost-sensitive learning: In this technique, a cost matrix is introduced to penalize misclassification of the minority class more heavily than the majority class. This can be useful in cases where the cost of misclassifying the minority class is much higher than the majority class.

2. Ensemble learning: This technique involves combining multiple models trained on different subsets of the dataset to improve classification performance. Ensemble methods such as Bagging, Boosting, and Stacking can be used to combine multiple classifiers and improve the overall performance of the model.

3. Undersampling: In this technique, a random subset of the majority class samples is

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

Ans: Both SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are oversampling techniques used to address the problem of imbalanced datasets in classification tasks. However, there are some differences between the two techniques:

1. SMOTE generates synthetic samples for the minority class by interpolating between existing samples, whereas ADASYN generates synthetic samples by adding noise to existing samples.

2. SMOTE creates synthetic samples by randomly selecting a minority class sample and its k-nearest neighbors, whereas ADASYN introduces a density distribution function to generate synthetic samples based on the density of minority class samples in their local neighborhood.

3. ADASYN is adaptive and focuses on generating more samples in areas where the class distribution is sparser, whereas SMOTE generates synthetic samples uniformly across the feature space.

4. ADASYN aims to address the problem of overfitting by generating fewer synthetic samples in regions where the density of minority class samples is high, whereas SMOTE does not consider this problem

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

Ans: GridSearchCV is a method used in machine learning to tune the hyperparameters of a model. The purpose of using GridSearchCV is to find the optimal combination of hyperparameters that maximizes the performance of the model on a validation set.

GridSearchCV works by searching through a predefined grid of hyperparameters and performing cross-validation on each combination of hyperparameters to evaluate the model's performance. The hyperparameters with the highest cross-validation score are then selected as the optimal hyperparameters for the model.

GridSearchCV is particularly useful when working with large datasets because it automates the process of hyperparameter tuning, which can be time-consuming and computationally expensive. It also ensures that the hyperparameters are tuned properly, leading to better performance and more accurate predictions.

However, GridSearchCV can also be computationally expensive, especially when the grid of hyperparameters is large. In some cases, it may be necessary to use other methods for hyperparameter tuning, such as RandomizedSearchCV, which samples from a distribution of hyperparameters instead of searching through a grid.

Overall, GridSearchCV is a powerful tool for hyperparameter tuning and can be very effective for optimizing model performance on large datasets, but it's important to be mindful of the computational cost and to explore alternative methods when necessary.

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

Ans: There are several evaluation metrics that can be used to evaluate the performance of a regression model. Some of the commonly used ones are:

1. Mean Squared Error (MSE): The MSE measures the average of the squared differences between the predicted and actual values. A lower MSE indicates a better fit between the predicted and actual values.

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

3. Mean Absolute Error (MAE): The MAE measures the absolute differences between the predicted and actual values. It provides a more intuitive measure of the average error and is less sensitive to outliers than MSE or RMSE.

4. R-squared (R2): The R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit between the predicted and actual values.

5. Adjusted R-squared: The adjusted R-squared adjusts for the number of independent variables in the model and provides a more accurate measure of the goodness of fit.

6. Mean Percentage Error (MPE): The MPE measures the average percentage difference between the predicted and actual values. It provides a measure of the average percentage error in the predictions.

7. Mean Absolute Percentage Error (MAPE): The MAPE measures the absolute percentage difference between the predicted and actual values. It is similar to MPE but provides a more intuitive measure of the average percentage error.

It's important to choose the appropriate evaluation metric based on the specific problem and the goals of the analysis. Some metrics may be more appropriate for certain types of data or models, and it's important to consider the strengths and limitations of each metric when evaluating the performance of a regression model.Top of Form