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?

ANS. B) In hierarchical clustering you don't need to assign number of clusters in beginning

2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

ANS. 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?

ANS. A) SMOTE

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.

ANS. B) 1 only

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

ANS. 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?

ANS. C) K-Nearest Neighbors

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?

ANS. A) CART is used for classification, and CHAID is used for regression.

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?

ANS. 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

9. Which of the following methods can be used to treat two multi-collinear features?

ANS. 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?

ANS. A) Overfitting

B) Multicollinearity

C) Underfitting

Q11 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?

ANS. There are some situations where one-hot encoding might not be the best choice

1. High Cardinality Categorical Variables: If a categorical variable has a high number of unique values, one-hot encoding will create a large number of columns, making it difficult to process the data. In such cases, other encoding techniques such as frequency encoding or target encoding can be used.

2. Rare Categories: If a categorical variable contains rare categories with very few observations, one-hot encoding will create a sparse matrix with a large number of zeros, which can affect the performance of machine learning models. In such cases, other encoding techniques such as frequency encoding or target encoding can be used.

3. Sequential Data: If the categorical variable has an inherent order or sequence, one-hot encoding will not be able to capture this information. In such cases, other encoding techniques such as label encoding or ordinal encoding can be used.

4. Natural Language Text: If the categorical variable is text-based and contains a large number of unique values, one-hot encoding will not be a good choice due to the high dimensionality of the

resulting sparse matrix. In such cases, techniques such as word embedding or bag-of-words representation can be used.

one-hot encoding may not be the best choice in situations where the categorical variable has high cardinality, rare categories, sequential data, or natural language text. In such cases, other encoding techniques such as frequency encoding, target encoding label encoding, ordinal encoding, word embedding, or bag-of-words representation can be used depending on the specific requirements of the problem

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

ANS. Data imbalance is a common problem in dassification tasks where the number of instances in one class is significantly higher than the other classes. This can lead to biased models that perform poorly on the minority class. To address this problem, several techniques can be used to balance the dataset, such as:

1. Random Undersampling

2. Random Oversampling

3. Synthetic Minority Oversampling Technique (SMOT

4. Adaptive Synthetic Sampling (ADASYN)

5. cost-sensitive learning

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

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

SMOTE works by generating synthetic examples of the minority class by interpolating between the existing minority class examples. It selects an example from the minority class at random and then selects its k nearest neighbors. New synthetic examples are then created by taking linear combinations of the features of the original example and its neighbors.

ADASYN, on the other hand, is an extension of SMOTE that focuses on the more difficult examples of the minority class. ADASYN generates more synthetic examples in the vicinity of the existing minority class examples that are harder to learn. It does this by first calculating the density distribution of the minority class examples and then generating synthetic examples proportional to the density distribution.

In summary, the main difference between SMOTE and ADASYN is that SMOTE generates synthetic examples of the minority class by interpolating between existing examples, while ADASYN generates more synthetic examples in the vicinity of the harder-to-learn examples based on the density distribution. ADASYN can be more effective than SMOTE when the minority class is highly imbalanced and contains very difficult examples to learn. However, both techniques can be useful for improving the performance of machine learning models on imbalanced datasets

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 for tuning hyperparameters of a machine learning model. The purpose of using GridSearchCV is to systematically search for the best combination of hyperparameters that yields the best performance of the model. It does this by exhaustively searching over a grid of hyperparameter values and evaluating the performance of the model for each combination of hyperparameters.

GridSearchCV is particularly useful when working with complex models that have many hyperparameters to tune. It saves time and effort compared to manually tuning the hyperparameters by evaluating each combination separately, GridSearchCV can also be used to compare the performance of different machine learning algorithms, by evaluating the performance of each algorithm with different hyperparameters.

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

ANS. 1. Mean Squared Error (MSE): This is a popular metric for evaluating regression models. It measures the average of squared differences between predicted and actual values. The lower the MSE value, the better the model. However, it has a drawback of heavily penalizing large errors, making it sensitive to outliers.

2. Root Mean Squared Error (RMSE): This metric is similar to MSE but takes the square root of the average squared error. It's preferred over MSE because it is easier to interpret since it's in the same units as the target variable.

3. Mean Absolute Error (MAE) This metric measures the average absolute differences between predicted and actual values. It's less sensitive to outliers than MSE but also has a drawback of not penalizing large errors.

4. R-squared (R) This metric measures the proportion of variance in the target variable that can be explained by the model. It ranges between 0 and 1, where 1 indicates that the model perfectly fits the data.

5. Adjusted R-squared: This metric is similar to R but adjusts for the number of predictors in the model. It's useful for comparing models with different numbers of predictors.

6. Mean Absolute Percentage Error (MAPE): This metric measures the average percentage difference between predicted and actual values. It's commonly used in finance and economics but has a drawback of not being able to handle zero values in the actual data.

7. Mean Percentage Error (MPE): This metric measures the average percentage difference between predicted and actual values. It's similar to MAPE but can handle zero values in the actual data.