MACHINE LEARNING

ASSIGNMENT – Worksheet 8 Answers – Vivek Kumar Sahu – Internship 35)

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. 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?
Ans: One-hot encoding may not be the best option in situations where the number of categories or features is very large. When the number of categories is large, the resulting one-hot encoded feature set can be very sparse and high-dimensional, which can lead to memory and computational issues.

In such situations, an alternative encoding technique like Hashing Trick or Binary Encoding can be used. The Hashing Trick is a method that maps a categorical feature to a fixed-length vector, regardless of the number of categories. Binary Encoding, on the other hand, creates binary codes for each category that are shorter than the one-hot encoding, resulting in a more compact representation.

12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.
a) Random Under-Sampling: Random Under- sampling aims to balance class distribution by randomly eliminating majority class examples. This is done until the majority and minority class instances are balanced out.

b) Random Over-Sampling increases the number of instances in the minority class by randomly replicating them to present a higher representation of the minority class in the sample.

c) Informed Over Sampling: Synthetic Minority Over-sampling Technique for imbalanced data (SMOTE): This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models.

d) ADASYN (Adaptive Synthetic) (ADASYN) is based on the idea of adaptively generating minority data samples according to their distributions using K nearest neighbor. The algorithm adaptively updates the distribution and there are no assumptions made for the underlying distribution of the data. The algorithm uses Euclidean distance for KNN Algorithm.

13. What is the difference between SMOTE and ADASYN sampling techniques?
Ans: The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed distributions. The latter generates the same number of synthetic samples for each original minority sample.

14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?
Ans: To tune hyperparameters for best possible prediction results. It depends on the usage – for large datasets it takes a lot of time, but if we require quality, and have time we can use it. However, if time is critical, we can use Randomized Search CV instead.

MACHINE LEARNING

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief.
Ans: There are 3 main metrics for model evaluation in regression:

- R Square/Adjusted R Square. R Square is calculated by the sum of squared of prediction error divided by the total sum of square which replace the calculated prediction with mean. R Square value is between 0 to 1 and bigger value indicates a better fit between prediction and actual value. To avoid overfitting, Adjusted R Square is introduced because it will penalise additional independent variables added to the model and adjust the metric to prevent overfitting issue.

- Mean Square Error(MSE): MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives an absolute number on how much our predicted results deviate from the actual number. We cannot interpret much insights from one single result but it gives a real number to compare against other model results and helps to select the best regression model.

MACHINE LEARNING

MACHINE LEARNING

- • Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and make it easier for interpretation.

- • Mean Absolute Error (MAE): Mean Absolute Error (MAE) is similar to Mean Square Error (MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of absolute value of error. Compared to MSE or RMSE, MAE is a more direct representation of sum of error terms. MSE gives larger penalisation to big prediction error by squaring it while MAE treats all errors the same.

11.