Name of the Assignment: Machine Learning (Worksheet-8)

Submitted by : Shalini Joshi

Designation : Data Science Intern

Date of Submission : 7th Mar,2023

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

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?

A) max_depth B) n_estimators C) min_samples_leaf D) min_samples_splits

Ans: A) max_depth

3. Which of the following is the least preferable resampling method in handling imbalance datasets?

A) SMOTE B) RandomOverSampler C) RandomUnderSampler D) ADASYN

Ans: 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

Ans: C) 1 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

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?

A) Decision Trees B) Support Vector Machines C) K-Nearest Neighbors D) Logistic Regression

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?

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

Ans: 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)

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

Ans: C) &D)

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

Ans: B),C) &D)

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

Ans: A) Overfitting, B) Multicollinearity & D) Outliers

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: We should not use the One Hot Encoding method when: When the categorical features present in the dataset are ordinal i.e for the data being like Junior, Senior, Executive, Owner.

Label Encoding can be used in such a case. Label Encoding in Python can be achieved using Sklearn Library. Sklearn provides a very efficient tool for encoding the levels of categorical features into numeric values. LabelEncoder encode labels with a value between 0 and n_classes-1 where n is the number of distinct labels.

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

Ans: . In case of data imbalance problem in classification, the following techniques can be used to balance the dataset:

- **Resampling (Oversampling and Undersampling)** : This technique is used to upsample or downsample the minority or majority class. When we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, we can randomly delete rows from the majority class to match them with the minority class which is called undersampling.
- **Synthetic Minority Oversampling Technique** or **SMOTE** is another technique to oversample the minority class. Simply adding duplicate records of minority class often don't add any new information to the model. In SMOTE new instances are synthesized from the existing data.
- **BalancedBaggingClassifier**: A BalancedBaggingClassifier is the same as a sklearn classifier but with additional balancing. It includes an additional step to balance the training set at the time of fit for a given sampler. This classifier takes two special parameters "sampling_strategy" and "replacement".
- **Threshold moving:** In the case of our classifiers, many times classifiers actually predict the probability of class membership. We assign those prediction's probabilities to a certain class based on a threshold which is usually 0.5, i.e. if the probabilities < 0.5 it belongs to a certain class, and if not it belongs to the other class.

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.

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 technique for finding the optimal parameter values from a given set of parameters in a grid. It's essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made.

For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible.

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

Ans: Evaluation metrics to evaluate a regression model:

- Mean Squared Error: It calculates the average of the square of the errors between the actual and the predicted values. Lower the value, better the regression model.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Here yi denotes the true score for the ith data point, and ŷi indicates the predicted value and n is the number of data points.

- RMSE is the most popular metric to measure the error of a regression model.This metric is calculated as the square root of the average squared distance between the actual and the predicted values.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

- Mean Absolute Error: It is calculated as the mean of the absolute difference between the actual and the predicted values.

$$MAE = \frac{1}{N}\sum_{i=0}^{N}|y_i - \hat{y}_i|$$

Where ŷi is the predicted value of the ith sample, and yi is the corresponding actual value, and N is the number of samples.

- R-Square: It measures the proportion of variance of the dependent variable explained by the independent variable.

***********************The End*********************************