MACHINE LEARNING WORKSHEET 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

➔ C) Both are equally proficient

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

➔ 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
➔ 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
➔ 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
A) 3-1-2 B) 2-1-3
C) 3-2-1 D) 1-3-2
➔ 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

➔ B) Support Vector Machines

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
➔
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)

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.
➔A, B

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
➔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
➔A,C

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?
➔ Machine learning models require all input and output variables to be numeric.
his means that if your data contains categorical data, you must encode it to numbers before you can fit and evaluate a model.
The two most popular techniques are an Ordinal Encoding and a One-Hot Encoding.
In this tutorial, you will discover how to use encoding schemes for categorical machine learning data.
After completing this tutorial, you will know:
Encoding is a required pre-processing step when working with categorical data for machine learning algorithms.
How to use ordinal encoding for categorical variables that have a natural rank ordering.

How to use one-hot encoding for categorical variables that do not have a natural rank ordering.

12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.
➔ Under-sampling balances the dataset by reducing the size of the abundant class. This method is used when quantity of data is sufficient. By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved for further modelling.

13. What is the difference between SMOTE and ADASYN sampling techniques?
➔ This approach by itself is known as the SMOTE method (Synthetic Minority Oversampling TEchnique). ADASYN is an extension of SMOTE, creating more examples in the vicinity of the boundary between the two classes than in the interior of the minority class.

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

➔ This article was published as a part of the Data Science Blogathon

Data-Driven decision-making has large involvement of Machine Learning Algorithms. For a business problem, the professional never rely on one algorithm. One always applies multiple relevant algorithms based on the problem and selects the best model based on the best performance metrics shown by the models. But this is not the end. One can increase the model performance using hyperparameters. Thus, finding the optimal hyperparameters would help us achieve the best-performing model. In this article, we will learn about Hyperparameters, Grid Search, Cross-Validation, GridSearchCV, and the tuning of Hyperparameters in Python.

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief.
➔ There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are: Mean Squared Error (MSE). Root Mean Squared Error (RMSE). Mean Absolute Error (MAE)