MACHINE LEARNING WORKSHEET – 1

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

Answer : **B**

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

Answer : **A**

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

A) SMOTE B) RandomOverSampler

C) RandomUnderSampler D) ADASYN

Answer : **C**

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

Answer : **D**

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

Answer : **D**

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

Answer : **B**

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

Answer : **C**

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.

Answer : **A,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

Answer : **B,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

Answer : **A,B**

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?**

Answer :

When the categorical features present in the dataset are ordinal. When the number of categories or the independent features in the dataset is quite large. One Hot Encoding should be avoided in this case as it can lead to high memory consumption.

A large number of levels are present in data. If there are multiple categories in a feature variable in such a case we need a similar number of dummy variables to encode the data. For example, a column with 30 different values will require 30 new variables for coding.

If we have multiple categorical features in the dataset similar situation will occur and again we will end to have several binary features each representing the categorical feature and their multiple categories e.g a dataset having 10 or more categorical columns.

We apply Label Encoding when:

1. The categorical feature is **ordinal** (like Jr. kg, Sr. kg, Primary school, high school)
2. The number of categories is quite large as one-hot encoding can lead to high memory consumption

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

Answer :

In case of Data Imbalance there are many Techniques where we can Balance the Dataset.

Resampling (Oversampling and Undersampling):

Under sampling is the process where you randomly delete some of the observations from the majority class in order to match the numbers with the minority class.

The second resampling technique is called, **Oversampling**. This process is a little more complicated than under sampling. It is the process of generating synthetic data that tries to randomly generate a sample of the attributes from observations in the minority class. There are a number of methods used to oversample a dataset for a typical classification problem. The most common technique is called **SMOTE (**Synthetic Minority Over-sampling Technique). In simple terms, it looks at the feature space for the minority class data points and considers its *k* nearest neighbours.

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

Answer:

What smote does is simple. First it finds the n-nearest neighbors in the minority class for each of the samples in the class . Then it draws a line between the the neighbors an generates random points on the lines. hen create samples on the lines with class == minority class.

ADASYN is an improved version of Smote. What it does is same as SMOTE just with a minor improvement. After creating those sample it adds a random small values to the points thus making it more realistic. In other words instead of all the sample being linearly correlated to the parent they have a little more variance in them i.e they are bit scattered.

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

Answer :

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.

When working with large datasets, using GridSearchCV for hyperparameter tuning may not always be the most efficient approach. GridSearchCV works by exhaustively searching through all possible combinations of hyperparameters specified in a grid, which can be computationally expensive and time-consuming for large datasets.

Overall, the choice of hyperparameter tuning method depends on the specific characteristics of the dataset and the computational resources available. GridSearchCV can still be a good option for smaller datasets, while randomized search or Bayesian optimization may be more suitable for larger datasets.

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

Answer :

There are several evaluation metrics that can be used to evaluate a regression model's performance. These are some common ones used:

1. Mean Squared Error (MSE): MSE measures the average of the squared differences between predicted and actual values. It is calculated by taking the sum of the squared errors and dividing it by the number of data points. The lower the MSE, the better the model is performing.
2. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and provides a measure of the average magnitude of the errors in the predicted values. RMSE is more interpretable than MSE because it is in the same units as the response variable.
3. Mean Absolute Error (MAE): MAE is similar to MSE, but instead of squaring the errors, it takes the absolute value of the differences between predicted and actual values. MAE measures the average magnitude of the errors in the predicted values and is less sensitive to outliers than MSE.
4. R-squared (R2): R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. R2 ranges from 0 to 1, with a value of 1 indicating a perfect fit. However, R2 can be misleading when used alone and should be used in conjunction with other metrics.