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=C) Both are equally proficient

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

Ans=B) n\_estimators

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

Ans=B) RandomOverSampler

1. Which of the following statements is/are true about “Type-1” and “Type-2” errors?

Ans=C) 1 and 3

1. Arrange the steps of k-means algorithm in the order in which they occur:

Ans=B) 2-1-3

1. 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=B) Support Vector Machines

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

Ans=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).

In Q8 to Q10, more than one options are correct, Choose all the correct options:

1. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?

Ans=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

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

Ans=C) Use ridge regularization D) use Lasso regularization

1. 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=C) Underfitting D) Outliers

Q10 to Q15 are subjective answer type questions, Answer them briefly

1. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?

Ans=One-Hot-Encoding has the advantage that the result is binary rather than ordinal and that everything sits in an orthogonal vector space. The disadvantage is that for high cardinality, the feature space can really blow up quickly and you start fighting with the curse of dimensionality.

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

A one hot encoding allows the representation of categorical data to be more expressive. Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers. This is required for both input and output variables that are categorical.

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

Ans=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.

Dealing with imbalanced datasets entails strategies such as improving classification algorithms or balancing classes in the training data (data preprocessing) before providing the data as input to the machine learning algorithm. The later technique is preferred as it has wider application.

1. 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.

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

Ans=GridSearchCV tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method. Hence after using this function we get accuracy/loss for every combination of hyperparameters and we can choose the one with the best performance.

GridSearchCV is a library function that is a member of sklearn's model\_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyperparameters.

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

Ans=1. R Square/Adjusted R Square

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)

3. Mean Absolute Error(MAE)

R Square/Adjusted R Square

R Square measures how much of variability in dependent variable can be explained by the model. It is square of Correlation Coefficient(R) and that is why it is called R Square.

Mean Square Error(MSE)/Root Mean Square Error(RMSE)

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

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.