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?

Ans:- D) None of these

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

Ans:- A) max_depth

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

Ans:- A) SMOTE

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.

Ans: 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 centre

Ans:- A) 3-1-2

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?

Ans:- D) Logistic Regression

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

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 regularisation if you take a large value of regularisation constant(lambda), which of the following things may occur?

Ans:- A) Ridge will lead to some of the coefficients to be very close to 0

D) Lasso will cause some of the coefficients to become 0

9. Which of the following methods can be used when there are correlated features in the dataset?

Ans:- C) Use ridge regularisation

D) use Lasso regularisation

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?

Ans:- A) Overfitting

C) Underfitting

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 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.

Also Where For categorical variables where ordinal relationship exists, the one hot encoding is not enough. We have to use Label Encoder for ordinal data.

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

Ans:- An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed.

Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the

assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class.

Two approaches to make a balanced dataset out of an imbalanced one are under-sampling and over-sampling.

1) Under-sampling

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.

2) Over-sampling

On the contrary, oversampling is used when the quantity of data is insufficient. It tries to balance dataset by increasing the size of rare samples.

Rather than getting rid of abundant samples, new rare samples are generated by using e.g. repetition, bootstrapping or SMOTE (Synthetic Minority Over-Sampling Technique).

3) Cluster-Based Over Sampling

In this case, the K-means clustering algorithm is independently applied to minority and majority class instances. This is to identify clusters in the dataset.

Subsequently, each cluster is oversampled such that all clusters of the same class have an equal number of instances and all classes have the same size.

4) Modified synthetic minority oversampling technique (MSMOTE) for imbalanced data

It is a modified version of SMOTE. SMOTE does not consider the underlying distribution of the minority class and latent noises in the dataset. To improve the

performance of SMOTE a modified method MSMOTE is used.

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

Ans:- SMOTE: Synthetic Minority Over sampling Technique (SMOTE) algorithm applies KNN approach where it selects K nearest neighbors,

joins them and creates the synthetic samples in the space. The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors.

The difference is multiplied by random number between (0, 1) and it is added back to feature. SMOTE algorithm is a pioneer algorithm and many other algorithms are derived from SMOTE.

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. 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: Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model.

There are libraries that have been implemented, such as GridSearchCV of the sklearn library, in order to automate this process.

Grid Search can be thought of as an exhaustive search for selecting a model. In Grid Search, the data scientist sets up a grid of hyperparameter values

and for each combination, trains a model and scores on the testing data. In this approach, every combination of hyperparameter values is tried

and when running it on larger dataset can be very inefficient.

For example, searching 20 different parameter values for each of 4 parameters will require 160,000 trials of cross-validation.

This equates to 1,600,000 model fits and 1,600,000 predictions if 10-fold cross validation is used.

While Scikit Learn offers the GridSearchCV function to simplify the process, it would be an extremely costly execution both in computing power and time.

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

Ans: There are three main errors (metrics) used to evaluate models, Mean absolute error, Mean Squared error and R2 score.

Mean Absolute Error (MAE): Lets take an example where we have some points. We have a line that fits those points.When we do a summation of the absolute value distance from

the points to the line, we get Mean absolute error.The problem with this metric is that it is not differentiable.

Mean Squared Error (MSE): Mean Squared Error solves differentiability problem of the MAE. Consider the same diagram above. We have a line that fits those points. When we do a summation of

the square of distances from the points to the line, we get Mean squared error.

R2 Score: R2 score answers the question that if this simple model has a larger error than the linear regression model. However, it terms of metrics the answer we need is how much larger.

The R2 score answers this question. R2 score is 1 — (Error from Linear Regression Model/Simple average model).

Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y,

disregarding the input features, would get a R^2 score of 0.0.