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?
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

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

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.

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

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?

B) Support Vector Machines

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

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.

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

C) Use ridge regularization
D) Use Lasso regularization

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

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 creates d-dimensional vectors for each instance where d is the unique number of feature values in the dataset. **For a feature having a large number of unique feature values or categories**, one-hot encoding is not a great choice. To fight the curse of dimensionality, **binary encoding** might be a good alternative to one-hot encoding because it creates fewer columns when encoding categorical variables. Ordinal encoding is a good choice if the order of the categorical variables matters.

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

ANS- There are7 Techniques to Handle Imbalanced Data

1. **Use the right evaluation metrics**

Applying inappropriate evaluation metrics for model generated using imbalanced data can be dangerous. Imagine our training data is the one illustrated in graph above.

2. **Resample the training set**

Apart from using different evaluation criteria, one can also work on getting different dataset. Two approaches to make a balanced dataset out of an imbalanced one are under-sampling and over-sampling

3. **Use K-fold Cross-Validation in the Right Way.**

It is noteworthy that cross-validation should be applied properly while using over-sampling method to address imbalance problems.

4. **Ensemble Different Resampled Datasets.**

The easiest way to successfully generalize a model is by using more data. The problem is that out-of-the-box classifiers like logistic regression or random forest tend to generalize by discarding the rare class. One easy best practice is building n models that use all the samples of the rare class and n-differing samples of the abundant class.

5. **Resample with Different Ratios.**

The previous approach can be fine-tuned by playing with the ratio between the rare and the abundant class. The best ratio heavily depends on the data and the models that are used.

6. **Cluster the abundant class.**

An elegant approach was proposed by Sergey on Quora [2]. Instead of relying on random samples to cover the variety of the training samples, he suggests clustering the abundant class in r groups, with r being the number of cases in r.

7. **Design Your Models.**

All the previous methods focus on the data and keep the models as a fixed component. But in fact, there is no need to resample the data if the model is suited for imbalanced data.

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

ANS-

1-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.

2- 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-

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.

Grid Search CV technique is not recommended for large-size datasets or param grids with a large number of components.

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

ANS-

There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are:

1) **Mean Squared Error (MSE).** MAE is a very simple metric which calculates the absolute difference between actual and predicted values.

2) **Root Mean Squared Error (RMSE).** As RMSE is clear by the name itself, that it is a simple square root of mean squared error.

3**) Mean Absolute Error (MAE)** MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.