Machine Learning Assignment - 8

1. What is the advantage of hierarchical clustering over K-means clustering?
Ans : 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?
Ans: max_depth

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

4. Which of the following statements is/are true about "Type-1" and "Type-2" errors?
Ans : 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 center
Ans : 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?
Ans : Support Vector Machines

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?
Ans : 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?
Ans : Lasso will lead to some of the coefficients to be very close to 0

9. Which of the following methods can be used to treat two multi-collinear features?
Ans : remove both features from the dataset

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 : Overfitting

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. There are various other techniques to encode the categorical (ordinal or nominal) features.
Time-based features such as day of month, day of week, day of year, etc have a cyclic nature and have many feature values. One-hot encoding day of monthfeature results in 30 dimensionality vector, day of year results in 366 dimension vector. It's not a great choice to one-hot encode these features, as it may lead to a curse of dimensionality.
The elegant solution to encode these cyclic features can be using mathematical formulation and trigonometry.
day of week the feature has 7 unique feature values. Taking the sin and cosine of the feature values will create 2 dimensionality features.Now, instead of creating a 7-dimensionality feature vector using One-hot encoding, a 2-dimensional transformed feature vector will serve the purpose to represent the entire feature. Now, let's visualize the new 2-dimensional transformed feature vector with a scatterplot.
The scatterplot clearly shows the cyclic nature of the day of week feature. The 7-feature values (from 0 to 6) are now encoded into a 2-dimensional vector.

12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.
Ans : Imbalanced datasets are a special case for classification problem where the class distribution is not uniform among the classes. Typically, they are composed by two classes: The majority (negative) class and the minority (positive) class
Below are the techniques by which we can handle it:
1. Random Undersampling and Oversampling
A widely adopted and perhaps the most straightforward method for dealing with highly imbalanced datasets is called resampling. It consists of removing samples from the majority class (under-sampling) and/or adding more examples from the minority class (over-sampling).
2. Undersampling and Oversampling using imbalanced-learn
imbalanced-learn(imblearn) is a Python Package to tackle the curse of imbalanced datasets.

It provides a variety of methods to undersample and oversample.

a. Undersampling using Tomek Links:
One of such methods it provides is called Tomek Links. Tomek links are pairs of examples of opposite classes in close vicinity.

In this algorithm, we end up removing the majority element from the Tomek link, which provides a better decision boundary for a classifier.
b. Oversampling using SMOTE:
In SMOTE (Synthetic Minority Oversampling Technique) we synthesize elements for the minority class, in the vicinity of already existing elements.

3. Class weights in the models

Most of the machine learning models provide a parameter called class_weights. For example, in a random forest classifier using, class_weights we can specify a higher weight for the minority class using a dictionary.
4. Change your Evaluation Metric
Choosing the right evaluation metric is pretty essential whenever we work with imbalanced datasets. Generally, in such cases, the F1 Score is what I want as my evaluation metric.

The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall.

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.

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 the process of performing hyperparameter tuning in order to determine the optimal values for a given model. As mentioned above, the performance of a model significantly depends on the value of hyperparameters. Note that there is no way to know in advance the best values for hyperparameters so ideally, we need to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus we use GridSearchCV to automate the tuning of hyperparameters.
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.

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