**MACHINE LEARNING [ASSIGNMENT NO -1]**

**In Q1 to Q7, only one option is correct, Choose the correct option:**

**Q.** 1. What is the advantage of hierarchical clustering over K-means clustering?

**ANS. D)None of these**

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

**ANS. A) max\_depth**

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

**ANS. A) SMOTE**

Q. 4. Which of the following statements is/are true about “Type-1” and “Type-2” errors?

**Ans. D) 2 and 3**

Q. 5. Arrange the steps of k-means algorithm in the order in which they occur:

**ANS. A) 3-1-2**

Q. 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**

Q. 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:**

**Q.** 8. In Ridge and Lasso regularization if you take a large value of regularization 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.**

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

**ANS. C) Use ridge regularization D) use Lasso regularization**

Q. 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**

**Q11 to Q15 are subjective answer type questions, Answer them briefly.**

Q. 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.**

For categorical variables where no such ordinal relationship exists, the integer encoding is not enough.In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).

In this case, a one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.

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.

* That categorical data is defined as variables with a finite set of label values.
* That most machine learning algorithms require numerical input and output variables.
* That an integer and one hot encoding is used to convert categorical data to integer data.

Q. 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.

Four 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.

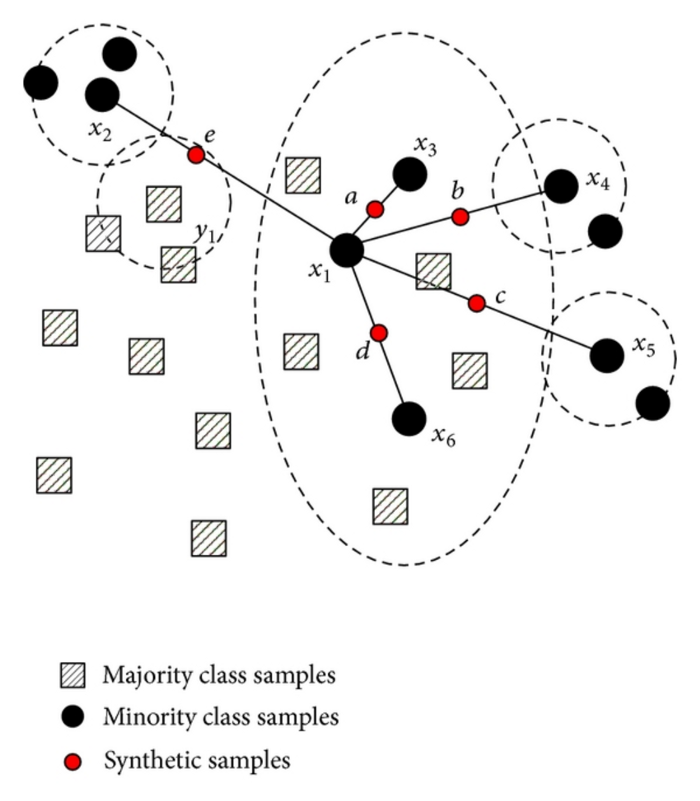
4. Modified synthetic minority oversampling technique (MSMOTE) for imbalanced data

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.

Q.NO.13. What is the difference between SMOTE and ADASYN sampling techniques?

ANS.NO.13.  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.

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.



See the above image so it finds the 5 nearest neighbors to the sample points. then draws a line to each of them. Then create samples on the lines with class == minority class.

ADASYN

Its a 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.

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.

Q.NO.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.

GridSearch Cv will calculate the average of out of fold recall for each combination of parameters, the set of parameters with best score, will be chosen by Grid search CV. It is fine to use the entire dataset, as you are using Cv method, which will check the score on out of fold set, hence you are not evaluating performance on Training data (on which model is trained) for parameter selection. for example: we want to tune max depth of tree, let's say maximum depth parameter we want to test are 5,10,15. Grid Search Cv will calculate recall score on out of fold set for all three value. The max depth value corresponding to best score on out of fold set will be chose

After spending hours on cleaning the data to fit your model and tuning the parameters using GridSearchCV, you may come to find that all that hypertuning didn’t improve your model performance by very much. Was it worth it? Maybe. It all depends on how much time you spent on tuning your model. If you took a whole day to test out parameters and only improved your model accuracy by 0.5%, perhaps that wasn’t the best use of your time.

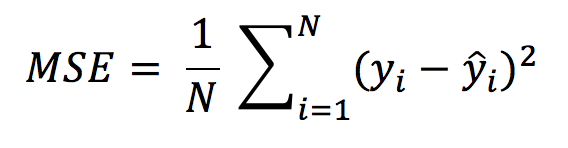
Q.NO.15.) List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief .

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

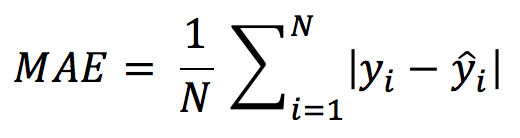
* Mean Squared Error (MSE).
* Root Mean Squared Error (RMSE).
* Mean Absolute Error (MAE)

MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret many insights from one single result but it gives you a real number to compare against other model results and help you select the best regression model.

**Root Mean Square Error(RMSE)** is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.



**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 the absolute value of error.



Mean Absolute Error formula

Compare to MSE or RMSE, MAE is a more direct representation of sum of error terms. MSE gives larger penalization to big prediction error by square it while MAE treats all errors the same.

R Square/Adjusted R Square is better used to explain the model to other people because you can explain the number as a percentage of the output variability. MSE, RMSE, or MAE are better be used to compare performance between different regression models.