1

DATA PREPROCESSING

The Foundation Of Machine Learning

As a Student, it becomes essential to understand what is being thrown at you and what do you interpret from it. It does matter that one's interpretation goes well with the contextual understanding of the subject but at the same time thinking of something different can work really well too. So this blog provides you with the interpretations made by a student and his experiences.

Just to put things in perspective, we all must know that everything in this small world is actually a data. So just think about this small world producing large amount of data, thanks to the advancements in technology, and now in parallel think about a elementary data analysing technique going over thousands of Exabyte of data. Now when I said small world and large data, I actually meant it, looking at the amount of data produced by the entire world some 15 years back, I found out it to be legit around 100 Exabyte mark. As I said earlier, thanks to the rise in computer technology, introduction of super computers and mobile computing, this amount of data is expected to shoot up to 41,000 Exabyte mark by the end of second decade of the 21st Century. The world already seems small?

Now, just think about processing this large amount of data with elementary methods used in data analysis. Possibilities are that some of the data might never get used. Here is where Machine Learning finds its roots. We all know that Machine Learning help us predict data by working on the given data and it has the potential to work on large amount of data.

The Machine Learning Model is completely dependent on the data that has been provided to it. To understand this, let us assume there's a dataset that has 4 columns, a machine learning model needs to know the independent and dependent variable. The independent variables of a dataset are used to predict the dependent variable. This is where data preprocessing comes into picture. Our data needs to be cleaned and made free from vulnerabilities before even applying a machine learning model.

Any vulnerability in our data will make our Machine Learning Model inconsistent. Several question arises, like, what are these vulnerabilities, how do they affect the model and how are they removed etc.

Vulnerabilities commonly found in our data sets are

1. Missing Data

   A Machine Learning Model can never predict the dependent variable if the data is missing in independent variables. Here, we need to understand that only a few rows in the entire column may have such scenario. The most common way to handle missing data is to replace it by mean/median/most frequent observation of the entire column.

2. Categorical Data

   A Machine Learning Model is purely based on mathematics; therefore we need to take care of data that is anything other than numbers. Strings are the prime example. This kind of data is called Categorical Data. To remove it from our dataset we encode it to numbers while making sure that our data doesn't loose on integrity.

Some other data preprocessing techniques include; Splitting Data into Training Set and Test Set, it means that we split our dataset into two subsets, training and test. As the name suggests Training set trains the dataset to understand the data and Test set examines whether or not the machine learning model has understood the correlation between independent and dependent variables. Another important technique is Feature Scaling, as most of the machine learning models are based on Euclidean distances ($\sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}$), it becomes important that one value doesn't dominate over the other, so it is necessary to scale down the values to a common range, say -1 to 1. This is achieved by either Standardisation or Normalisation of the entire column.

As we have seen that the data holds the key to success of any machine learning model, hence it becomes important that a dataset free from all the vulnerabilities is fed to the model to have accurate results.
Thus we can say that Data Preprocessing is the foundation of Machine Learning.