Machine Learning

Lab-2: Data Preprocessing

16BCE1259
Shushil Kumar Ravishankar

Why Preprocessing?

Data is generated at a large scale nowadays. But the generated data is not readily usable. There may be error in the value generated, values may not have been standardized, different units of measurements, null values may be present and many more. To make the data usable for us, it needs to be preprocessed.

About the dataset:

The dataset taken for preprocessing is 'Big Mart Sales' data. This dataset is a collection of the sales data of each item at different store type, stores of different establishment years and details of items like weight, fat content, etc. This dataset has been taken from kaggle.com and is one of the basic dataset to start working on.

Preprocessing:

The preprocessing is done with the help of python libraries pandas and seaborn.

After pair-plotting the dataset using seaborn with hue=fat-content, we notice that there are 5 labels which represent only 2 values because the values were not standardized while creating the dataset.

We can see that there are null values present in the dataset. On executing the isnull() on the dataset, we found **1463 values of Item-Weight and 2410 values of Outlet-size are missing**(null). We use the mean value to fill the null data values of Item-weight and fill the outlet-size as "Medium" (mean of the sizes of the outlets).

Once the null values have been taken care of, we plot histograms, distribution plot, violin plots to understand the distribution of the dataset and get a different view of the dataset. For example, we found that the outlet sales are more for low fat content items when compared to items having regular fat content. But in contrast, no of items with low fat content are much less when compared to no of items with regular fat content.

Conclusion:

After preforming the above steps, the dataset is now ready to be used further in programs and algorithms.