Feature Selections Technique for Machine Learning.

11 July 2023 19:29

Here in this notebook I will give the notes to a student who is struggling with Feature Selection Techniques.
I will assume the student has a brief knowledge of python and, is compatible with sicket learn library

Also, As Feature Selection requires a fair understanding of features in general. I will assume the student has a basic idea over statistics.

Feature Selection is basically a technique where you use some statistical methods to filter out best features who have a great impact with you target column.

It mostly used in high dimensional datasets, It helps us in reducing the dimensionality of the data and to improve the interpretability of our model.

There are basically 3 type of feature Selection:

- Filter based Techniques
- Wrapper Method
- Embedded Method
- Hybrid Method

TODAY WE WILL SEE SOME FILTER BASED LEARNING

BEFORE DISCUSSING THE TECHNIQUES,

1st point is to remove all the duplicate columns present in your data, sometimes there might be conditions where there are lot of features and you might find same features that are same so we should remove those features.

VARIANCE THRESHOLD : -

This is a method where we remove those features those variance are constant or quasi constant, meaning if we have some features whose variance is 0 or very nearer to 0 such as 0.00442 we will remove those features.

- We will set a threshold of a certain variance depends on the dataset but (0.01- 0.1) can varies depending on the situation.
- Our data should be normalized to apply this method because if the data is not normalized it can vary a lot depending on the range of the data in a particular column, so we should normalize the data.
- We will calculate the variance of all the columns and decide according to our threshold.
- Then we will keep the features which have a variance greater than our required threshold.

So, the concept behind this strategy is that, understand we have to focus on the features which have a high variance and assuming that the features having a high variance will give us meaningful insights to our model.

```
From sklearn.feature_selectionimportVarianceThreshold

sel=VarianceThreshold(threshold=0.05)
```

There are some disadvantages of using the variance Threshold methods which we have to consider for the stages:

- 1st condition is this method doesn't see the relationship with the feature with the target variable, So we have to be careful about that.
- It ignores the feature interaction with other features, for example : - If there is 2 col's such as longitude and latitude. If we get that longitude has a very low variance and we remove it is a blunder in our dataset. So we have to keep in mind the interaction among the features
- Before implementing this method we need to ensure that the datapoints are properly scaled.

CORRELATION : -

Correlation is a very basic approach to deal with features selection. As we understand what is correlation, the relationship between different features in the dataset. We will find all the corelation co-efficient among the features and we will then seeing the co-efficient decides which features are needed and which features to be discarded.

corr_matrix = X_train.corr() -------> (Simple corr() method)

Here we tackle a very major problem of Machine learning that is Multicollinearity problem. So the concept is we have to see for e.g.

F1 - - - -F2
F1 - - - - -F3

F1, F2 and F3 are 3 features having a high collinearity, So we will keep F1 and will simply remove f2 and f3. as they have a strong correlation with F1.

To do this we don't need to do all the maths pandas has there function to do this for us. We just need to understand the correlation between different features and remove those that falls below the threshold number set by us.

THINGS TO CONSIDER BEFORE APPLYING THIS TECHNIQUE:

This method works great with our linear bounded ml models like logistic and linear regression But, it doesn't tell us anything about the non- linear relation between the features.

This also has an ambiguity of over the threshold how to choose best threshold to gain the best features is subject to the dataset.

As we are working over linear model this is also subject to outlier sensitivity.