Abhishek Singhal

Homework-3

Clustering Algorithms:

The code trains 3 models of clustering:

1) KMeans
2) DBSCAN
3) Birch

I have used scatter plots to see the clustering visualization.

We can select any two attribute from dataset to plot scatter plot. This is done through widgets.

Some help is taken from code demo, mainly the selection of attributes part.

Feature Selection:

I select features for training models by :

```
digits_feature = digits[[#attribute_list[0],
                         attribute_list[1],
                        attribute_list[2],
                         #attribute_list[3],
                         # attribute_list[4],
                         # attribute_list[5]
                        ]]
```

Here we can use one or multiple features to train the models.

Case1: If we use feature that are being shown on X-axis and Y-axis then the regions are clearly defined by all three algos.

DBSCAN also catches noise (shown by Orange circles)

For ex: I use attribute[1],attribute[2] for features (Grocery and Milk);

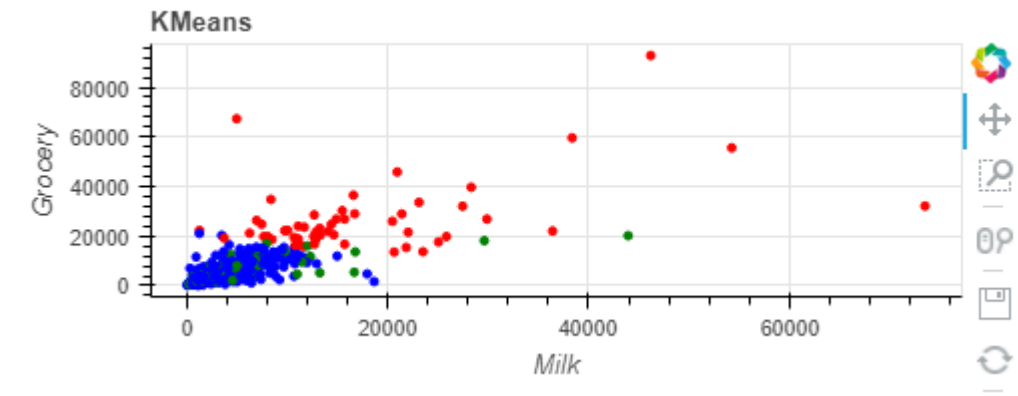
However,

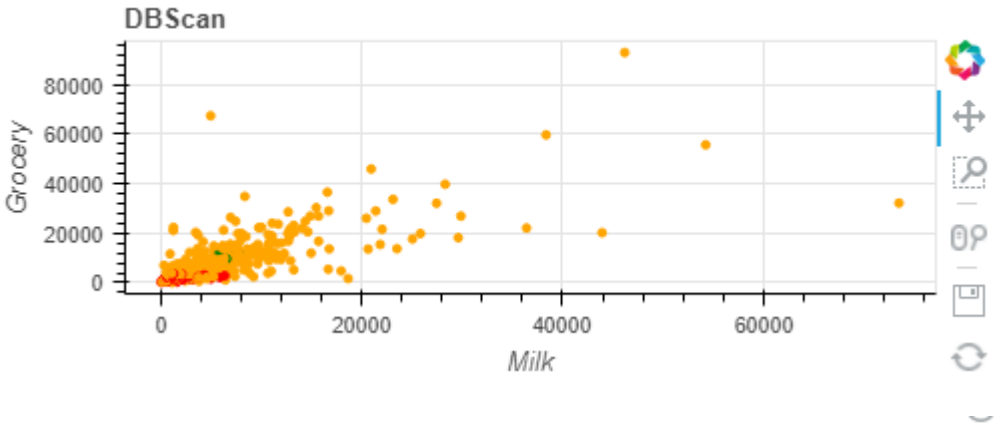
Case2: The clustering results are quite different if all **the features are used.**

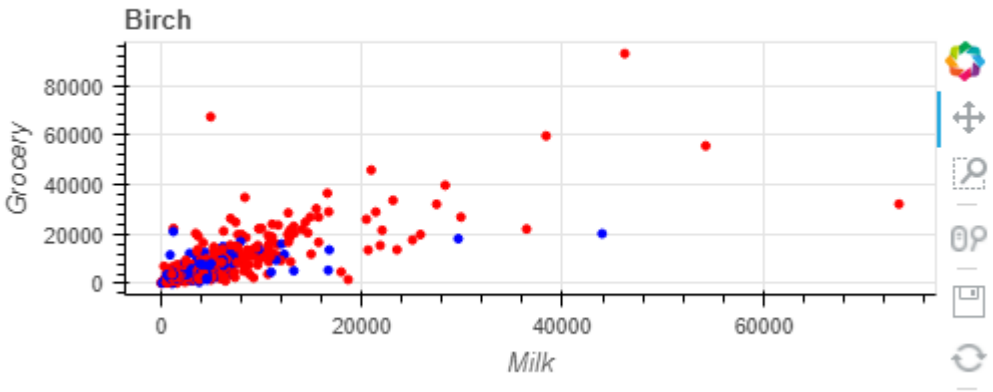
One thing to notice is that DBSCAN rejects almost complete dataset as noise.

Birch algorithm also fails to a certain extent.

However, KNN performs fairly well.







In the above visualizations,

- n_clusters=3 for KNN
- eps=2000 , min_samples =4 for DBSCAN
- threshold=5, branching factor =20, n_clusters =2 for Birch
- DBSCAN is sensitive to number of minimum points in a cluster.
- eps<=1000 treats every data point as noise.

KMeans:

As we increase n_clusters from 2 to 3 to 4, the original clusters are partitioned. It sometimes fails to catch outliers or noise like DBSCAN does.