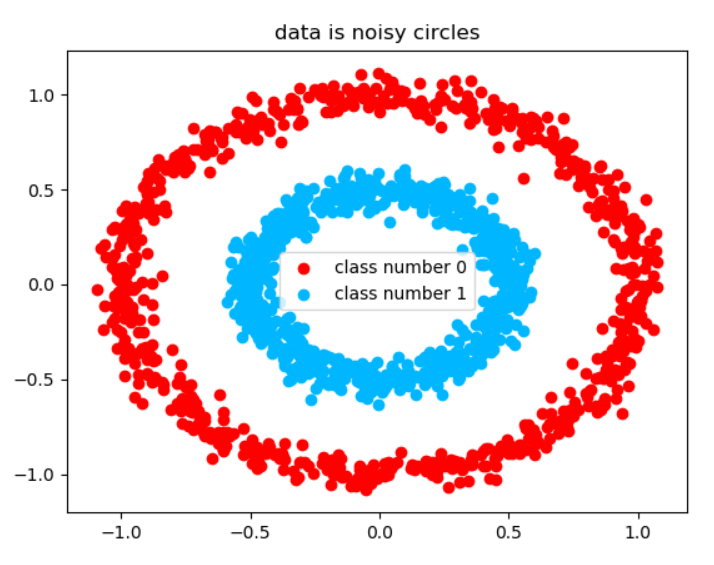
I used Birch and DBSCAN on three toy dataset and two real datasets. The toy data includes, noisy circles, noisy moons, and Blobs. The real data includes, iris and handwritten digits.

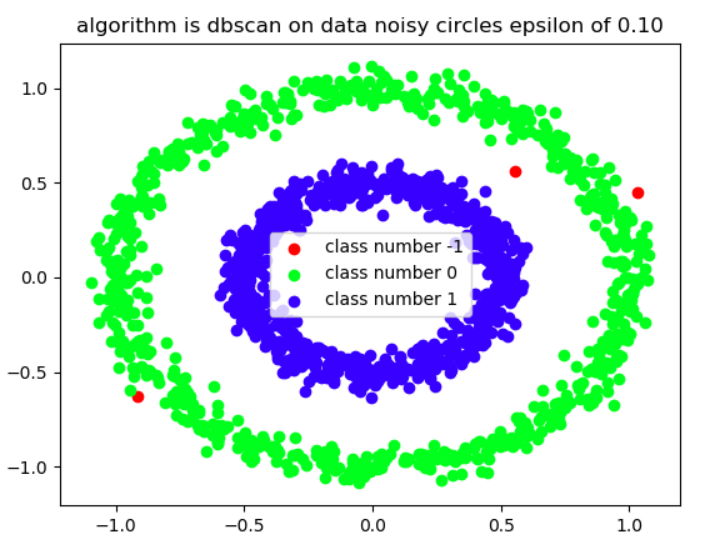
For DBSCAN I used different epsilon to get fairly nice results. And for Birch I used number of real clusters as we want to find the best possible answer which is a complete classifier, however there are some downside to this approach that it can lose homogeneity.

DATA:

Noisy Circle is consists of two inner and outer circle:



Which dbscan makes it:



-1 is outliers and the measurements are

algorithm: dbscan

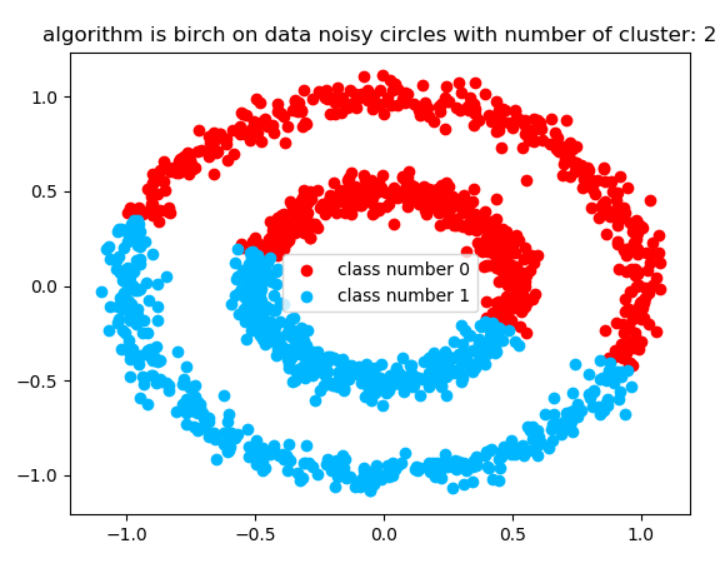
homogeneity score: 1.00

completeness score: 0.98

jaccard score: 1.00

normalized mutual information score: 0.99

silhouette score: 0.09



For Birch with metrics of:

homogeneity score: 0.00

completeness score: 0.00

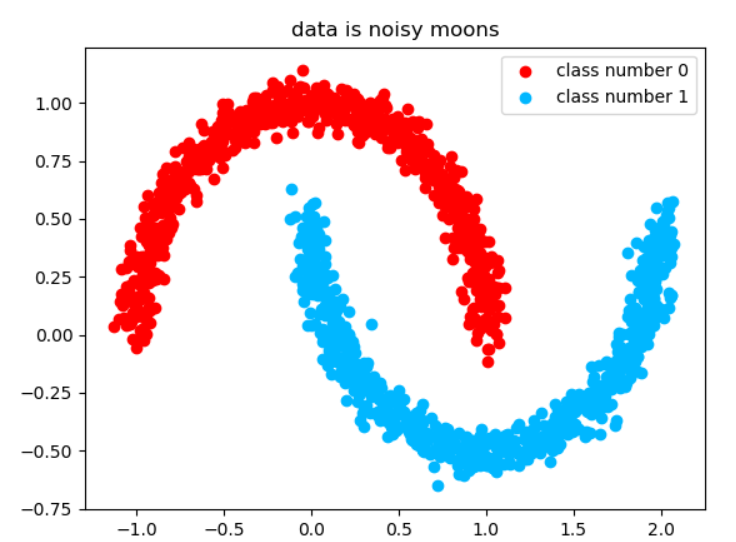
jaccard score: 0.50

normalized mutual information score: 0.00

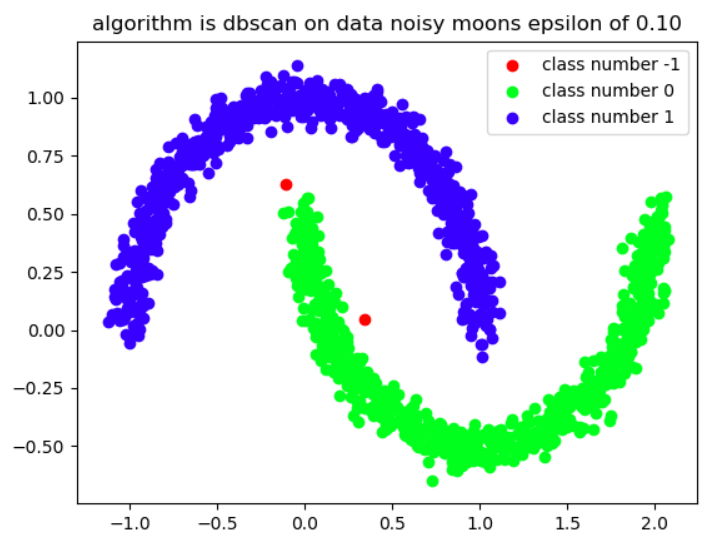
silhouette score: 0.35

The homogeneity is 0 because there are class 1 and 2 real data in our first class and completeness is 0 because our first class is consists of two different classes.

Noisy moon is two half circle which are separated from each other:



DBSCAN says:



algorithm: dbscan

homogeneity score: 1.00

completeness score: 0.99

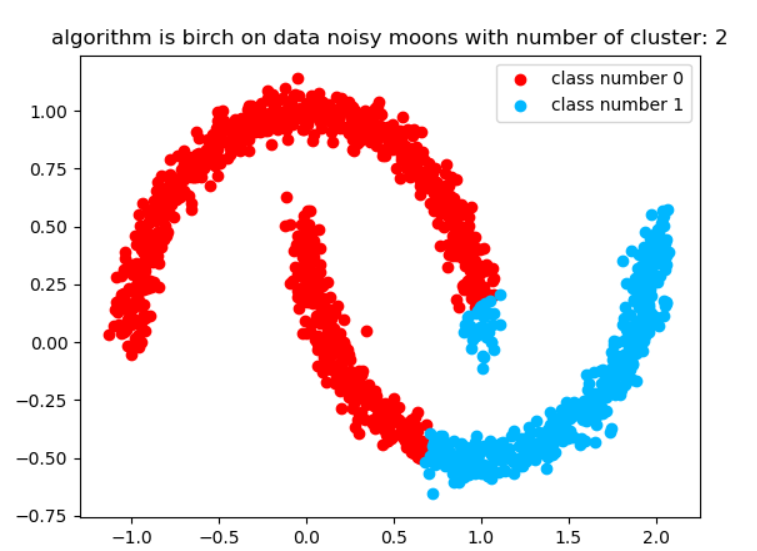
jaccard score: 0.00

normalized mutual information score: 0.99

silhouette score: 0.01

Has Jaccard 0 because although it is complete and homogeneous unfortunately our first class is real second class (the label mismatch) so that’s why we need homogeneity and completeness as our measurements.

For the birch:



homogeneity score: 0.27

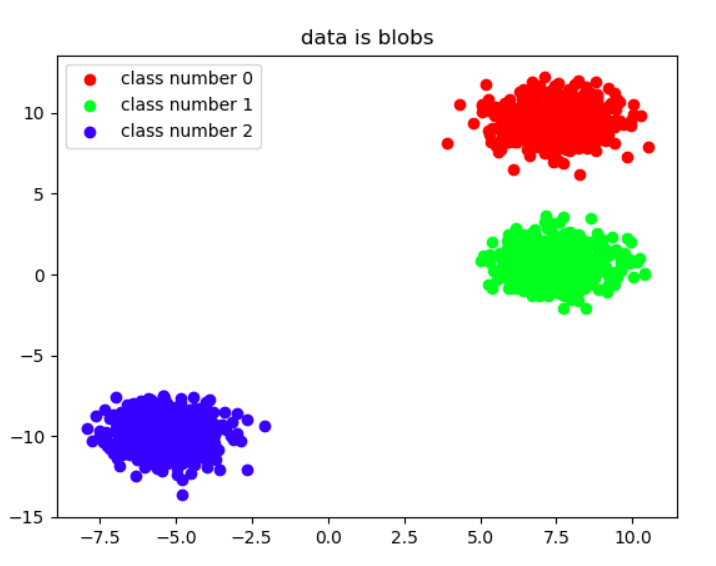
completeness score: 0.30

jaccard score: 0.77

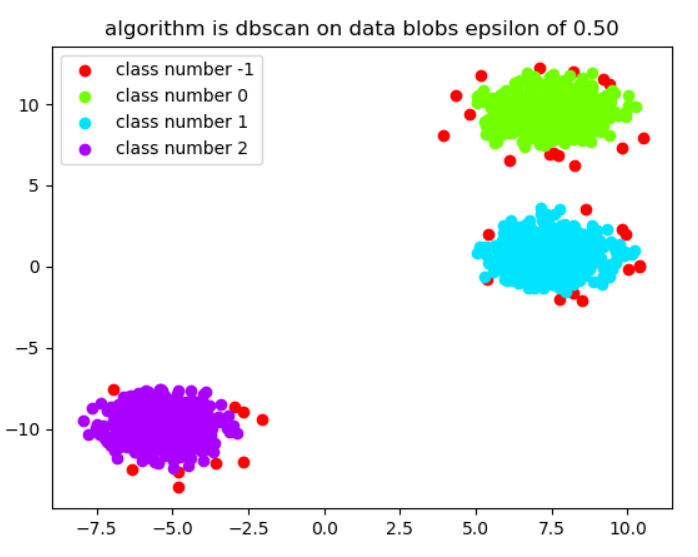
normalized mutual information score: 0.28

silhouette score: 0.46

Next data is Blob which is three different Gaussian mixture data and since they are separated it’s the easiest possible classification:



For DBSCAN:



homogeneity score: 0.98

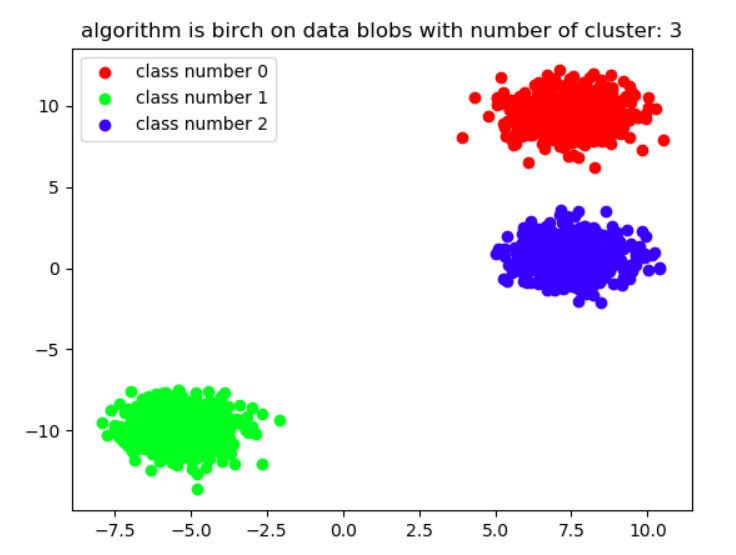
completeness score: 0.91

jaccard score: 0.98

normalized mutual information score: 0.94

silhouette score: 0.80

And for Birch:



homogeneity score: 1.00

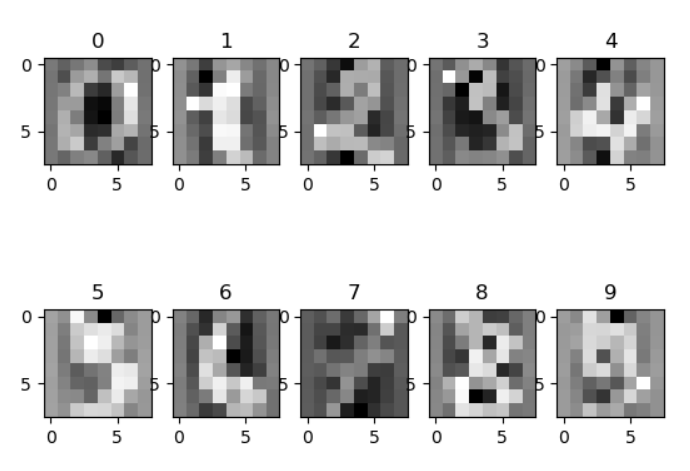
completeness score: 1.00

jaccard score: 0.33

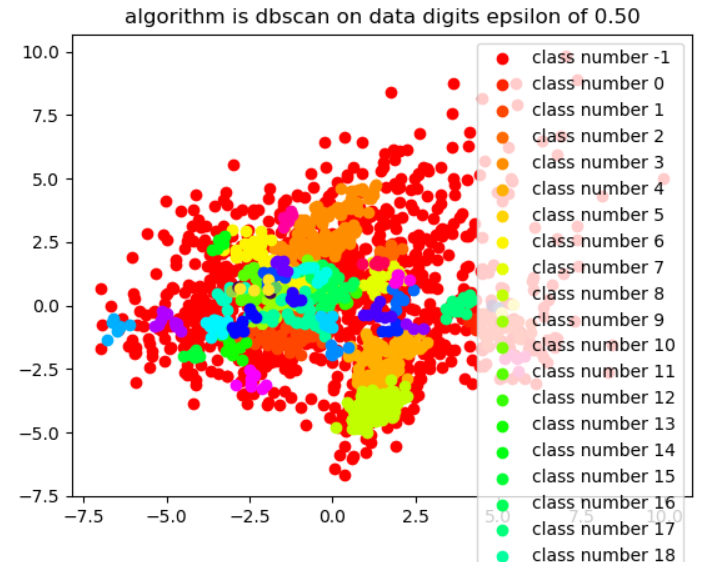
normalized mutual information score: 1.00

silhouette score: 0.83

Next data which is real life data is Handwritten digits:



For showing this data since we cannot use our whole feature data to show in two dimension we use PCA to map data in two dimension to show it there but use whole reduced feature data using PCA to number of classes to cluster them:



As we can see density base approach or DBSCAN finds much denses classes that are and can consist of same class each

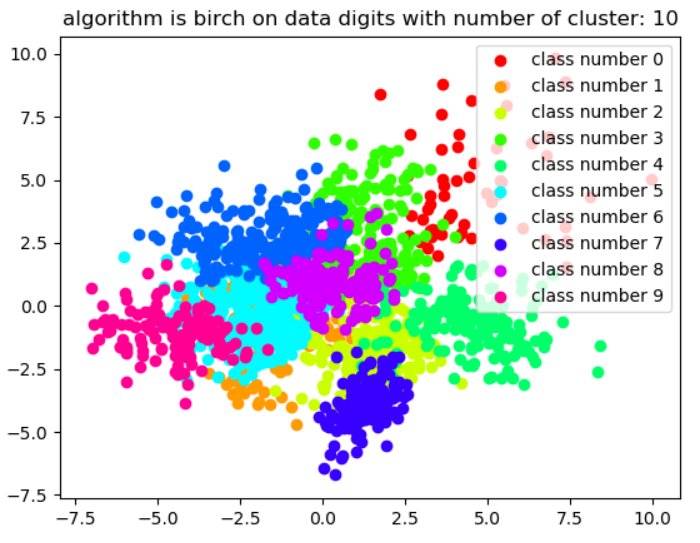
homogeneity score: 0.34

completeness score: 0.37

jaccard score: 0.02

normalized mutual information score: 0.36

But Birch:



homogeneity score: 0.53

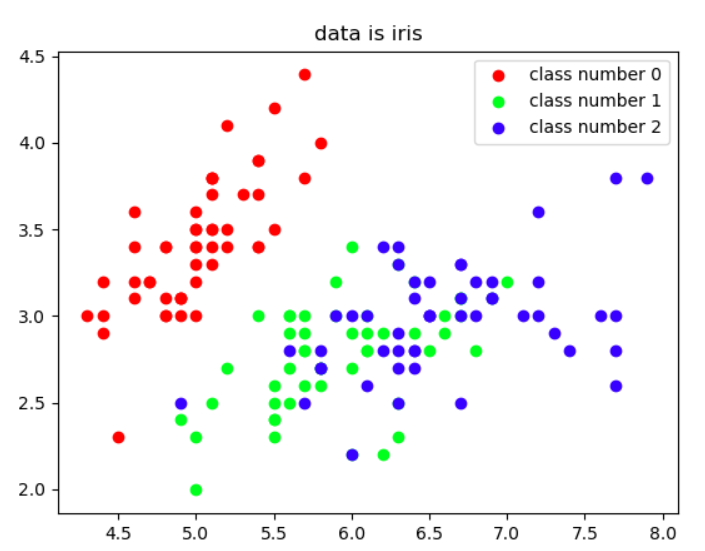
completeness score: 0.55

jaccard score: 0.15

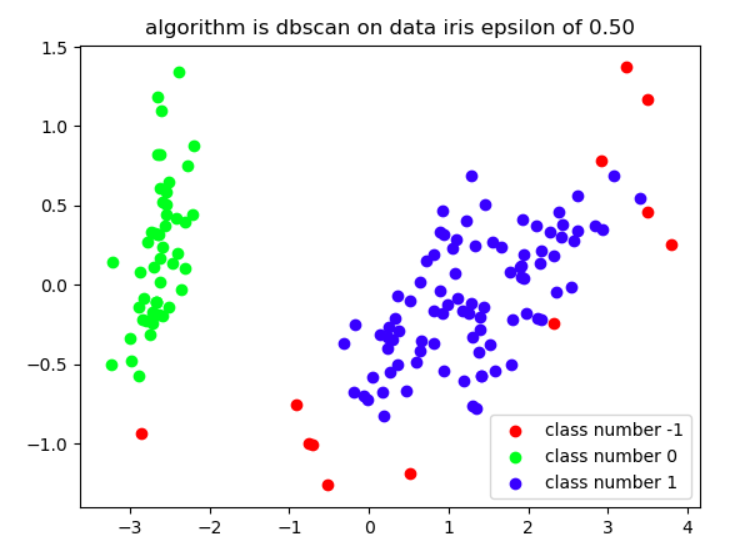
normalized mutual information score: 0.54

silhouette score: 0.30

find fairly good results and lastly the data is IRIS which is about flowers petal width and length and two more feature containing 3 classes:



Which DBSCAN:



homogeneity score: 0.56

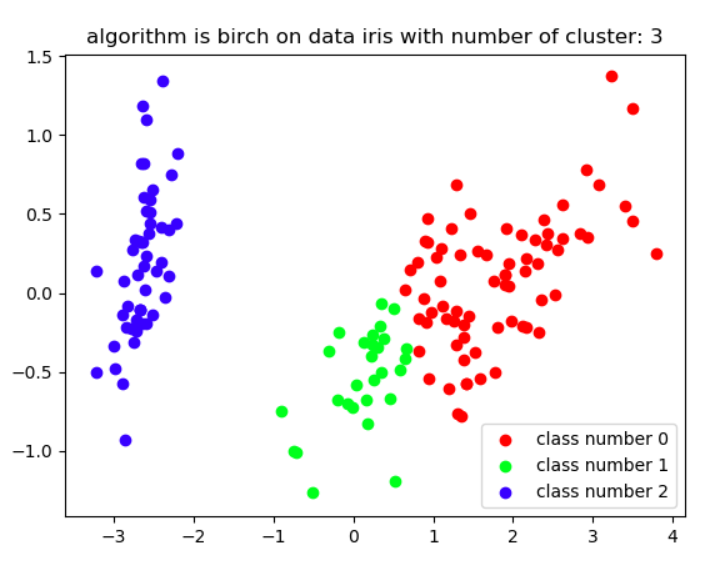
completeness score: 0.70

jaccard score: 0.63

normalized mutual information score: 0.63

silhouette score: 0.54

And Birch says that:



homogeneity score: 0.70

completeness score: 0.75

jaccard score: 0.18

normalized mutual information score: 0.72

silhouette score: 0.52

We can further improve on each results if we start tuning the hyper parameters but we decided to use same hyper for all of them.

You can see the results of fine tuning hyper parameters in the first two toy data because higher epsilon than what we issued to algorithm will give us poor results.