MSBD 5002
Assignment-4

Gardas Nuthan

20643274

1. **Fuzzy Clustering using EM algorithm:**

Here in this algorithm, I've computed the final clusters by taking the first two points as the initial clusters- C1, C2;

According to this, the SSE was calculated for the first two iterations and the updated C1,C2 for every iteration also:

1. For iteration-1: Updated clusters are: C1(0.44528628 0.31233445) and C2(0.42661802 0.95099126), the SSE of this iteration is: 721.4956974886536

2. For iteration-2: Updated clusters are C1(0.71141242 0.12192137), and C2(0.30463132 1.0173773), the SSE is : 395.2495171277726

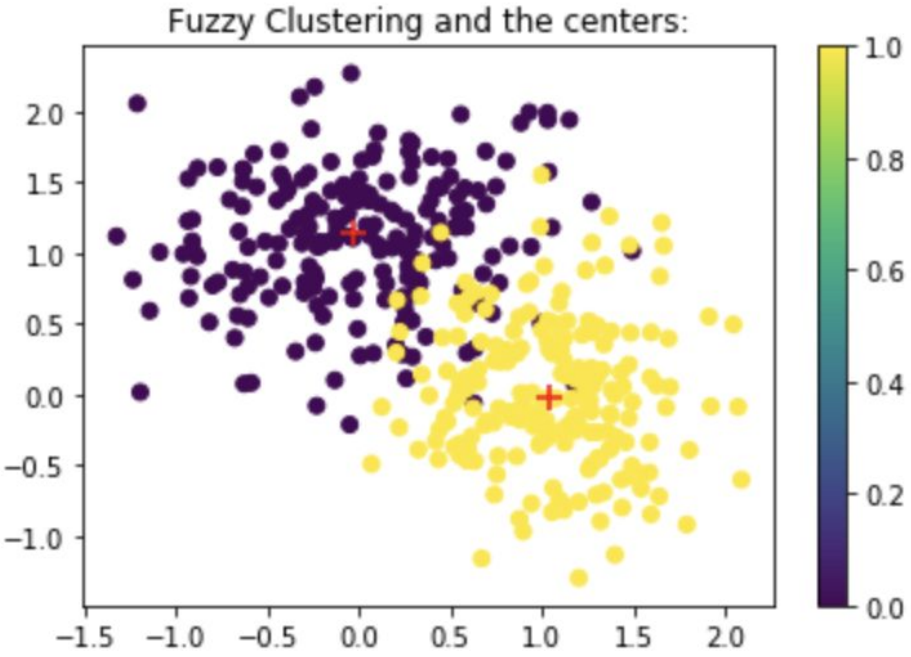
The iterations to converge is 10 here for these initial clusters and the updated centers are:

C1=(1.04436187 -0.02897887)
C2=(-0.03375124 1.13281557)

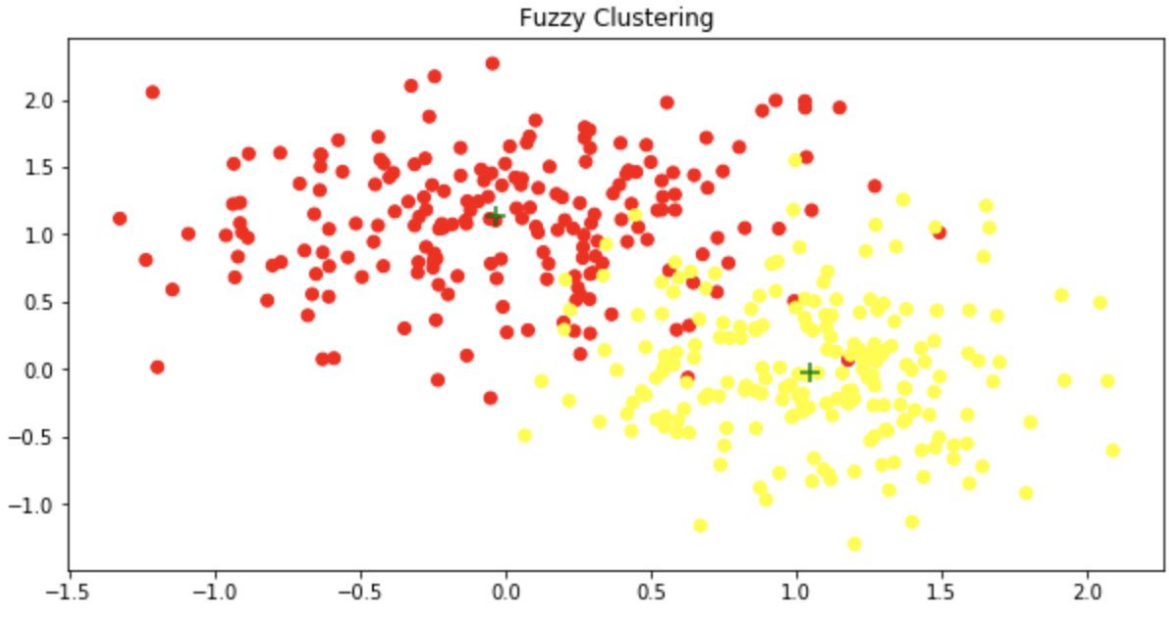
And I've also done the computation by taking the initial clusters as C1(0,0) and C2(1,1) instead of just taking the first initial two points from the data given.

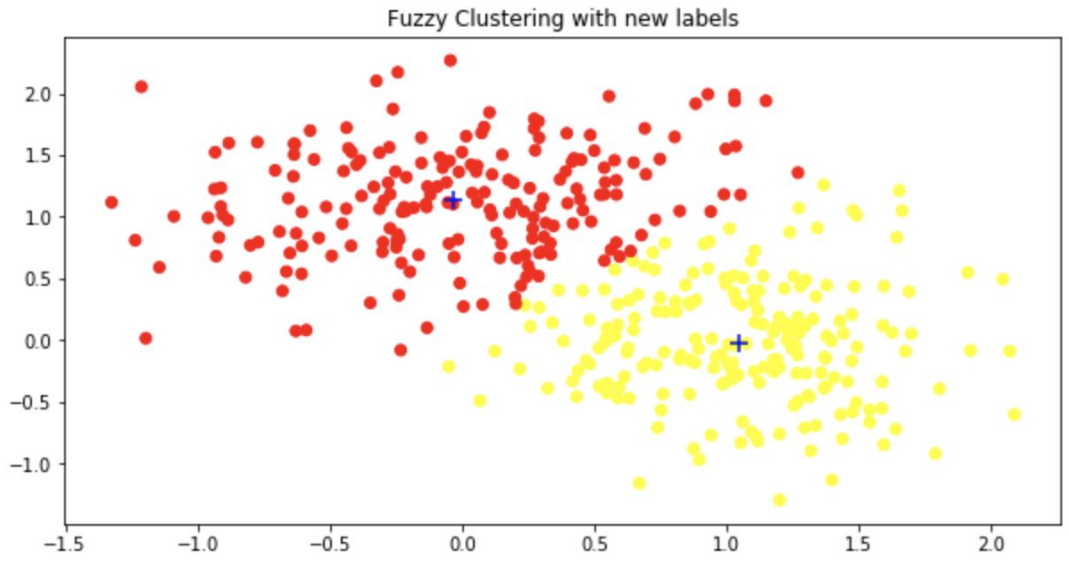
The total iterations is 12, 2 iterations more than the first case above, and the final updated centers are closely similar to the final updated centers above with just some minor changes in the numbers.

Either way, 10 iterations is better than 12, in this case, the first case is better to compute the clustering centers.



Centers are marked with red colours in the first figure.Here from the diagram, you can see that the centers are somewhat in resemblance to the labels, but in the lower part, some of the other labels are in the yellow part of the figure, let's expand the figure a bit and check:





After finding the converged centers, I've labelled them with the points given in order to cluster them more accurately, you can see that in the first figure before labelling, some of the other cluster points overlap with the yellow labels(C1), but after labelling again, they have been accurately clustered into two separate clusters. Yellow being cluster C1, red being cluster C2.

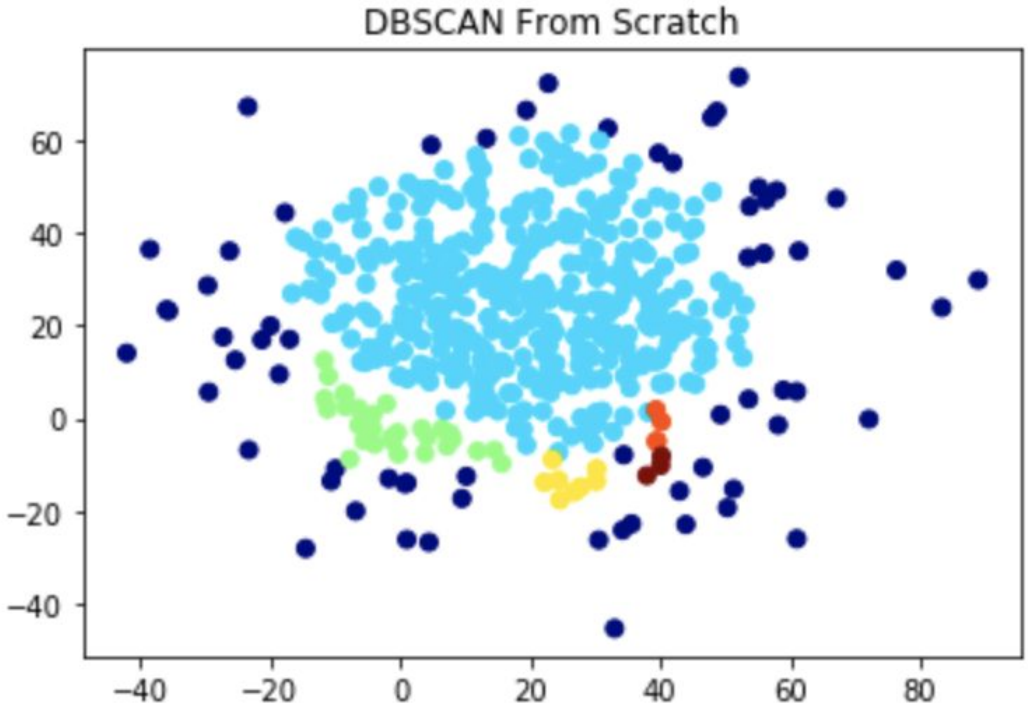
So there are 3-4 labels in the other yellow cluster (first figure), meaning some of them are misclassified.But the performance of the EM algorithm helps us in identifying the centers easily and classify the labels correctly.

2. **DBSCAN:**

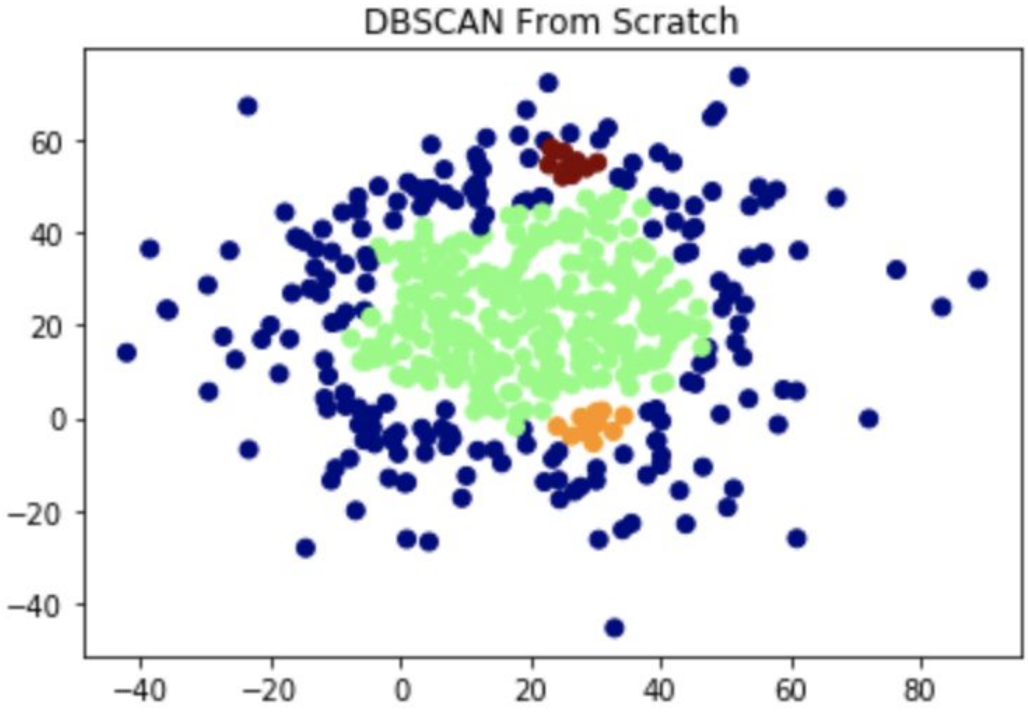
Density based algorithm in which there are two input parameters:
eps=neighbourhood radius of the cluster (Density of the cluster)
min_pts=number of points in order to be classified as a core cluster.

So we plotted for 4 different parameters and labelled the outliers as 'blue' or 'dark blue' and initiated a table in order to compare the performance of these given parameters.
Here are the figure of different parameter (eps,min_pts):

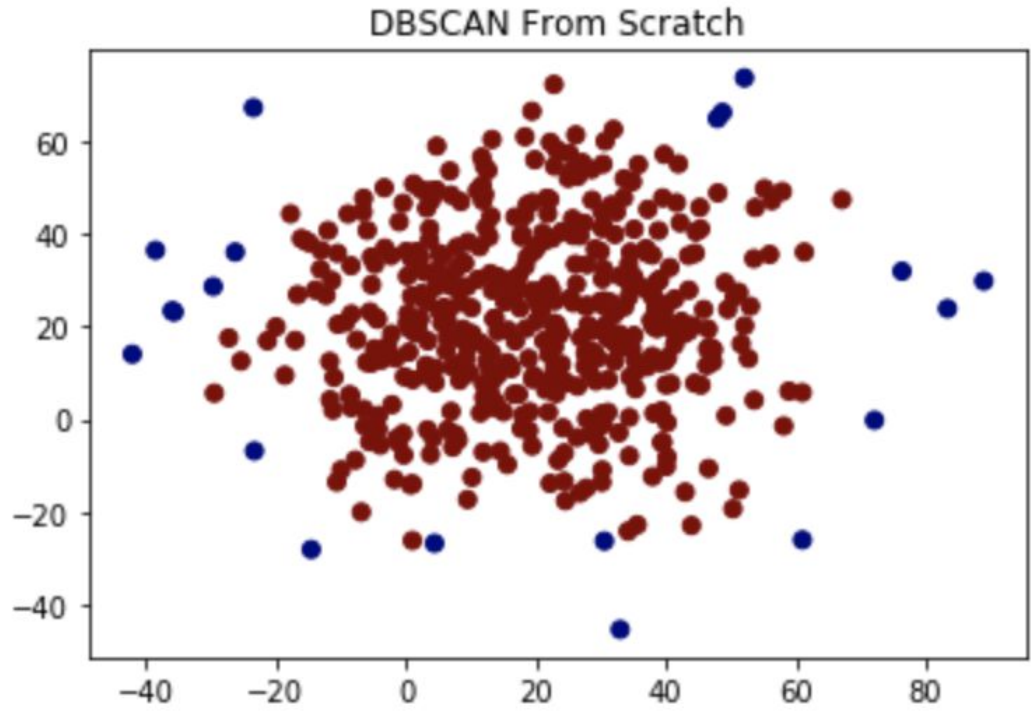
1. eps=5,min_pts=5



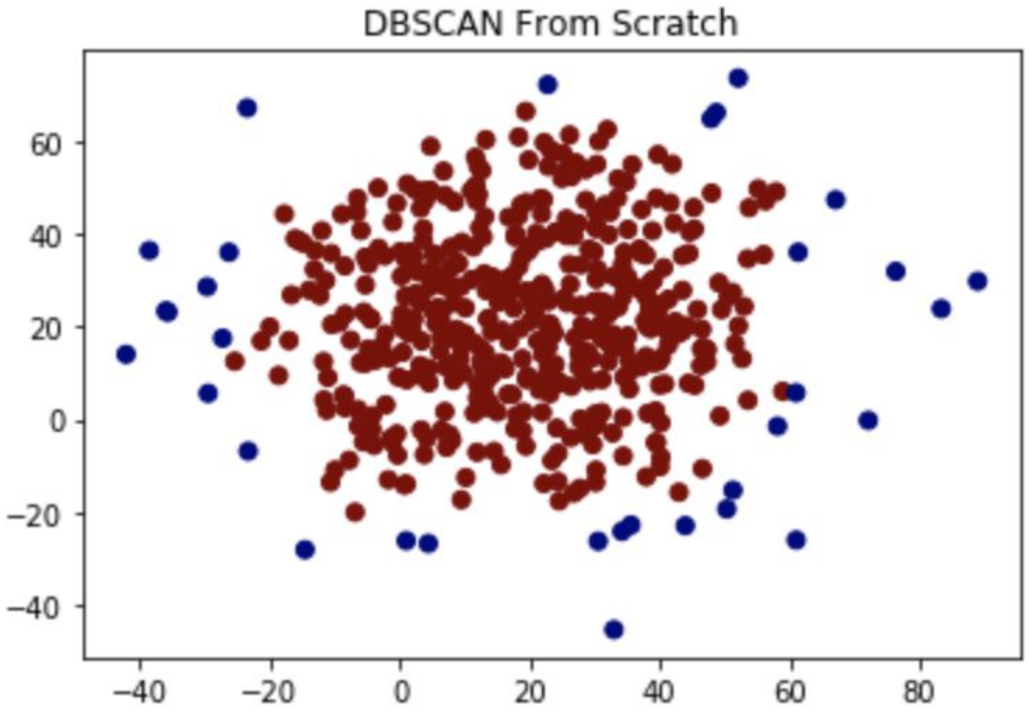
2. eps=5,min_pts=10



3. eps=10,min_pts=5



4. eps=10,min_pts=10



The table below gives us a more information about the no.of clusters and outliers:

	(eps,min_pts)	No.of Clusters	Outliers	Performance
1	(5,5)	6	65	Good
2	(5,10)	4	194	Bad
3	(10,5)	2	20	Bad
4	(10,10)	2	33	Bad

So according to me, the parameters of 5,5 is much more suitable to the data, as taking the min_pts parameter higher will contradict with the smaller eps radius, thereby misclassifying the points as outliers(194) here. When you take the radius higher than or equal to the min_pts, given the size of the data, it doesn't justify the clustering perfectly and there by clustering them into two giant clusters. Ultimately it comes down to the size of the data where the parameters are justified.

Therefore **(5,5)** is the better suited one when compared to others.