Assignment 4 Report

Course: MSBD 5002 Data Mining Course

Name: YANG Rongfeng

Data: 2020-05-07

Q1 Fuzzy Clustering using EM

Assume o_i represents the i-th object and c_j represents the j-th center. The sum of the squared error is computed by

$$SSE(C) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^2 \cdot dist(o_i, c_j)^2$$

We set two initial centroids

$$C_1 = (1, 1), C_2 = (2, 2)$$

In the first three iterations, the updated centers and SSE

Iteration	Updated Centers	SSE
0	C1 = (0.539,0.552), C2 = (0.474,0.610)	2736.853
1	C1 = (0.601,0.466),C2 = (0.432,0.635)	843.983
2	C1 = (0.726,0.329),C2 = (0.305,0.773)	853.136

Table 1. The updated centers and SSE during clutering iterations

After about 10 iterations, we can get the final converged centers

$$C_1 = (1.04, -0.0289), C_2 = (-0.0339, 1.133)$$

According to the figures we draw, we can clearly see that predicted clustering has a more distinct boundary than the clutering divided by original labels. I have drawed the clutering boundary below:

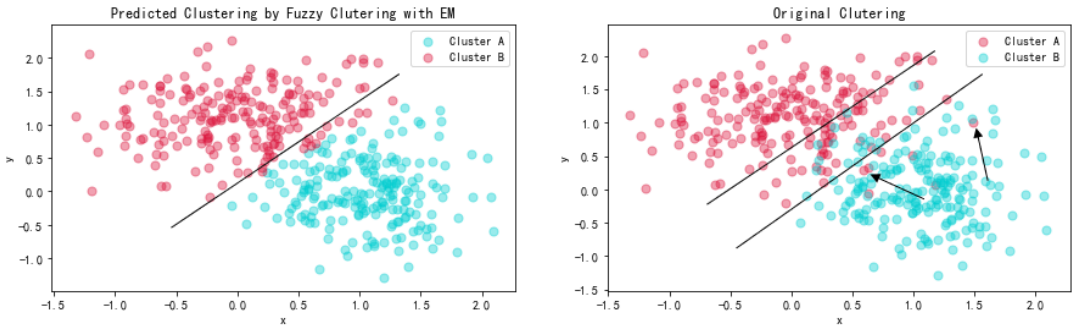

Figure 1. Comparison of predicted clustering and original clustering

We can easily draw a line to distinguish two cluters in the left figure. However, in the figure on the right, though we widen the boundary, there are still some outliers existing. This is because in reality, the noise exists which would make some points fluctuate near the classification boundary. However, The algorithm just consider the ideal situation and cannot consider noise.

Q2 DBSCAN

The DBSCAN clutering results of different combination of parameters are shown in Figure 2:

From the figures above, we could see that the outliers are mainly distributed in the edge area, which shown in dark BLUE. When using parameters `eps = 10 and min_points = 5`, the clusters are clear and distinct with less number of outliers. If the parameters are set as `eps = 5 and min_points = 10`, the number of outliers are up to 238! Not very well.

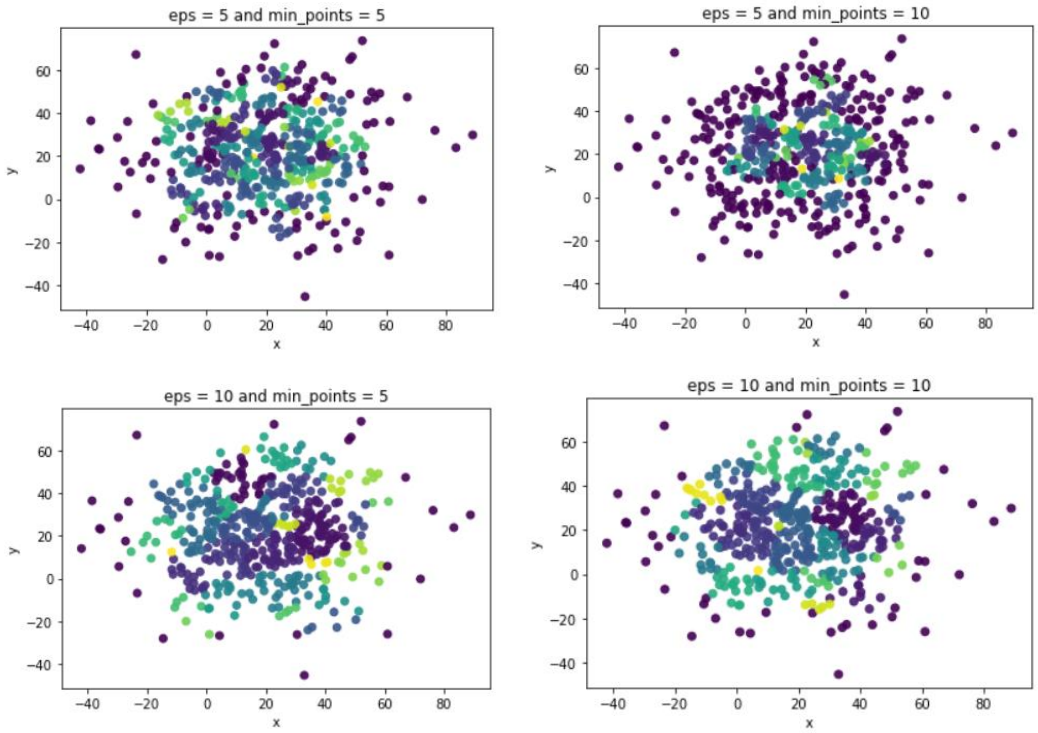

Figure 2. The DBSCAN clustering results through different combination of parameters

The detailed results are shown in the table:

eps	min_points	n_clusters	n_outliers
5	5	72	91
5	10	34	238
10	5	37	25
10	10	29	43

Table 2. The number of clusters and outliers through different combination of parameters