Junyi CHEN

CS5487

2017/11/04

CS5487 Programming Assignment 2: clustering

# PART 1 pOLYNOMINAL FUNCTION

## implementation

Implementation written in python is attached in the source code files.

## Run the algorithms on the three synthetic datasets

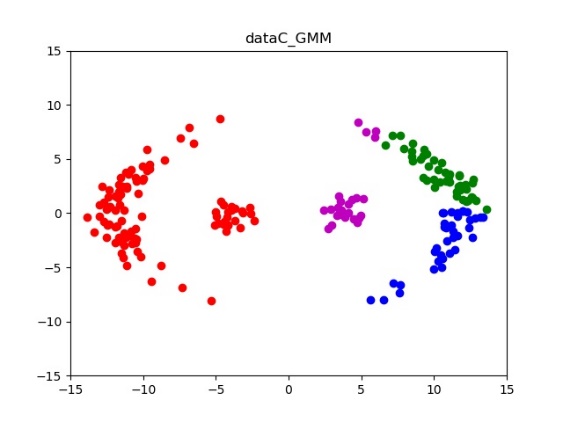
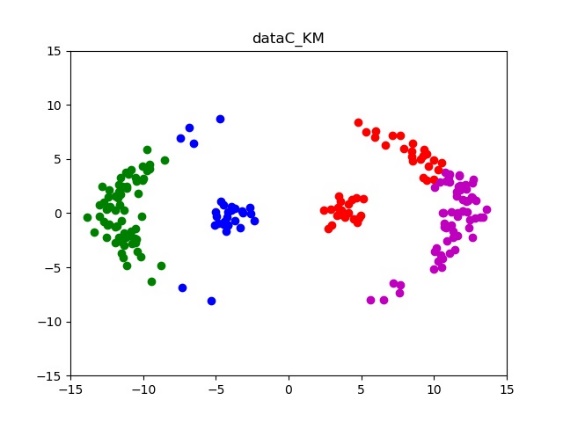
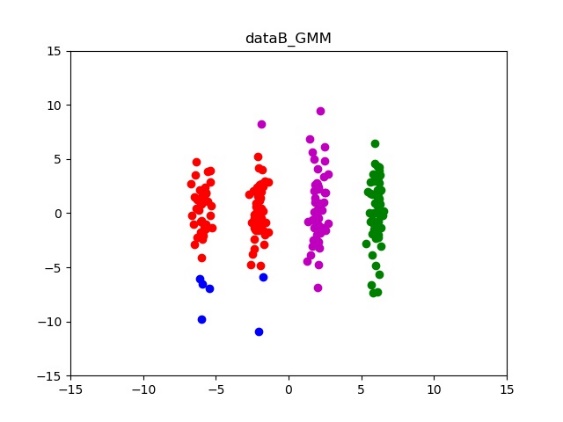
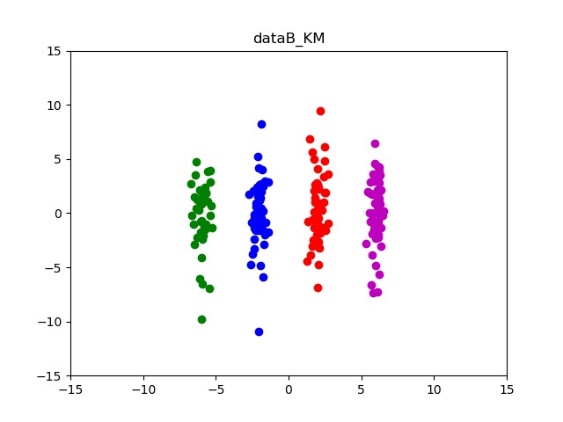
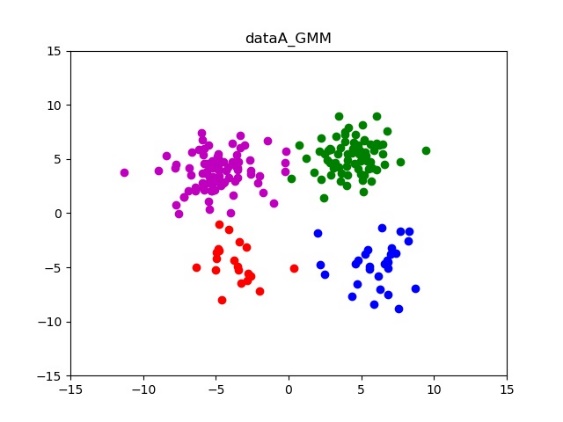
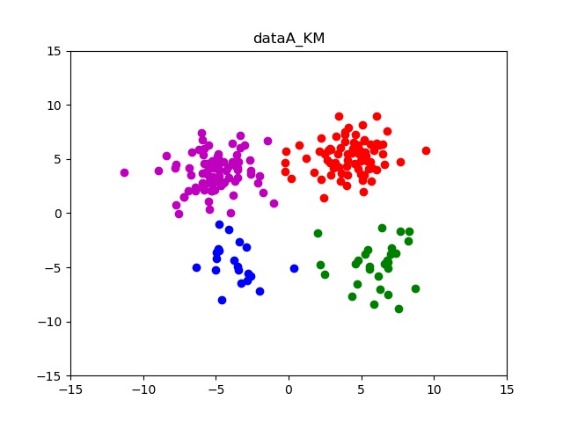


Figure 1: Predict of different clusters, the figures are K-means (left) and Gaussian mixture model (right)

In terms of K-means (KM) and Gaussian mixture model (GMM) the performance of data A is similar, both of them can correctly cluster four sets of data points. In terms of data B, K-means out puts better performance than Gaussian mixture model because GMM fail to discriminate the first and the second column wise data points. However, both KM and GMM do not work well in data C, KM cannot handle with points stay in the margin from the leftmost cluster. While GMM fail to separate the first and the second from left hand side.

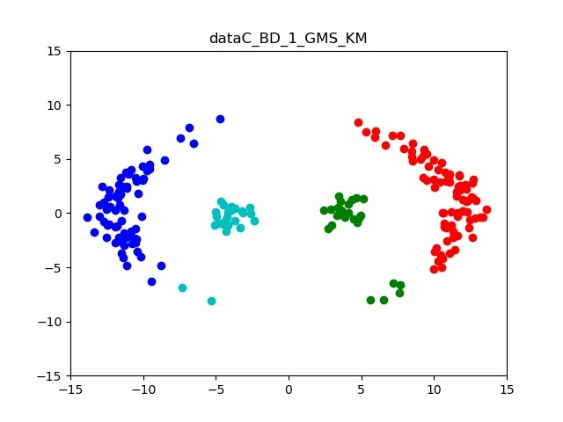
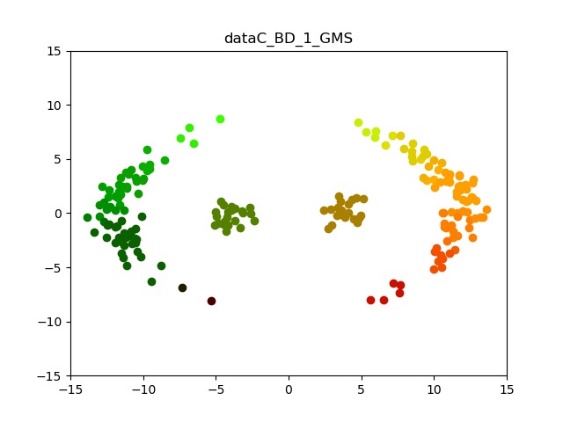
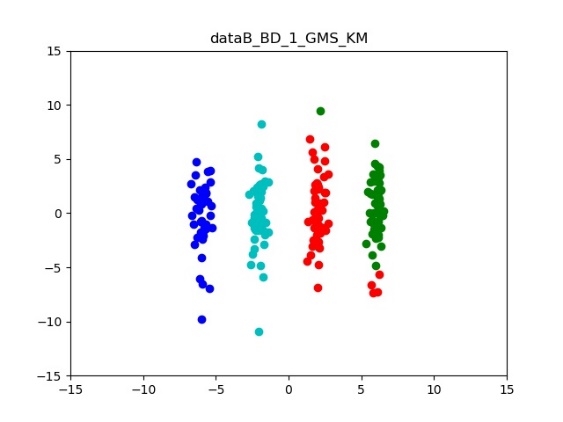
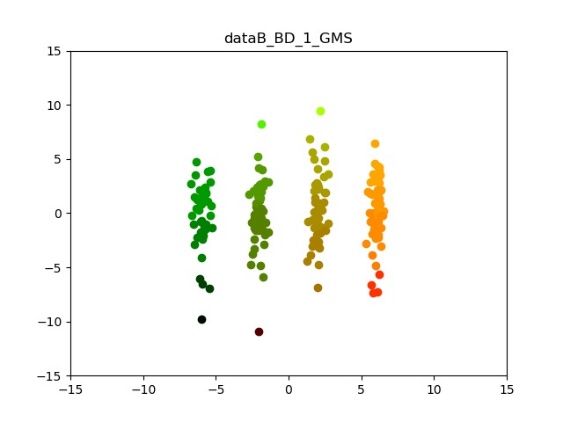
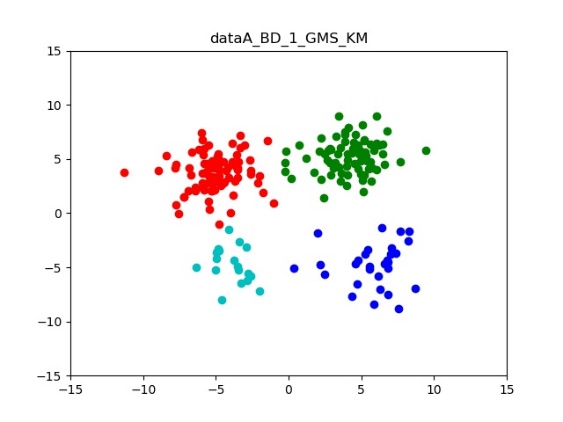
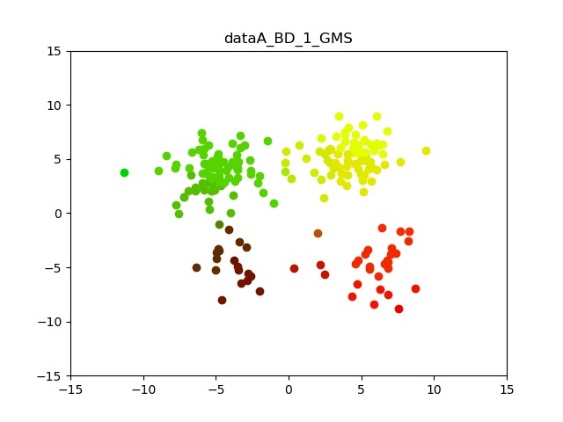


Figure 2 Predict of mean shift clustering with bandwidth equals to 1, color is displayed according to the value of local peak (left) and plots of local peak values clustered by K-means(right) to limit the cluster number of mean shift

Figure 2 are cluster result of mean shift clustering (MS). Because MS calculate the value of local peak, not represented by cluster center. To visualize the clustering result, the data point in figure 2 are colored by the value of local peak. Hence, points converge to the same peak will have similar color in figure 2. We can see that MS is working good on data A for the colors in 4 clusters are well separated. In terms of data B and C, colors of the 4 clusters can basically be discriminated. Anyway, from the plot of data B of MS, data points from the bottom seems to form their own cluster like GMM. Plots form right hand side is picture we apply KM on the output local peaks of MS. The reason why we do the second clustering is to limit the number of clusters. Qualitatively, MS is not as good as KM in data A and B but it works well in data C.

In summary, distance based KM works good in data A and B, and it is easy to implement. But the performance is not good using KM when the densities between clusters are not equal like data C. Also, cluster number is fixed and affiliation of data points is hard. One point belongs to one clusters only. GMM can output probability of cluster affiliation, but the cluster may not be as precise as KM and the initialization of estimated values potentially affect clustering result. Cluster number of GMM is also fixed. MS works well in our data A, B and C, and the representation is not by cluster center so it is flexible. But the limitation is that the stagey of cluster prototyping will affect the performance of cluster and the speed of GMM is much slower than KM and GMM.

## SENSITIVITY OF MEAN SHIFT BANDWIDTH

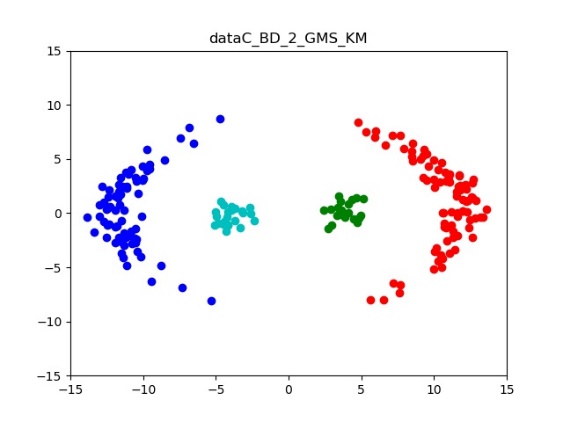
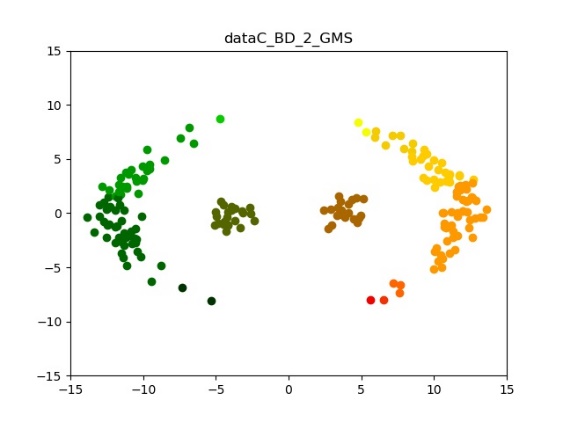
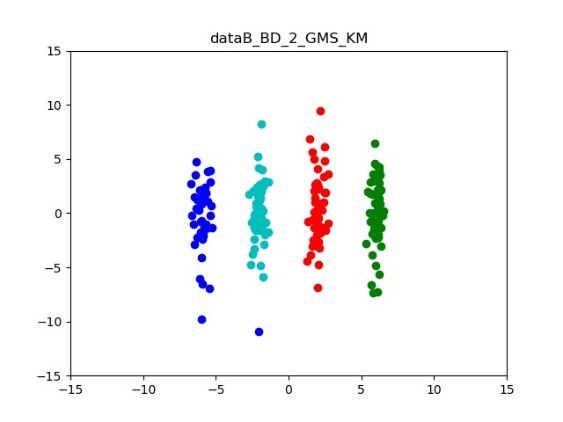
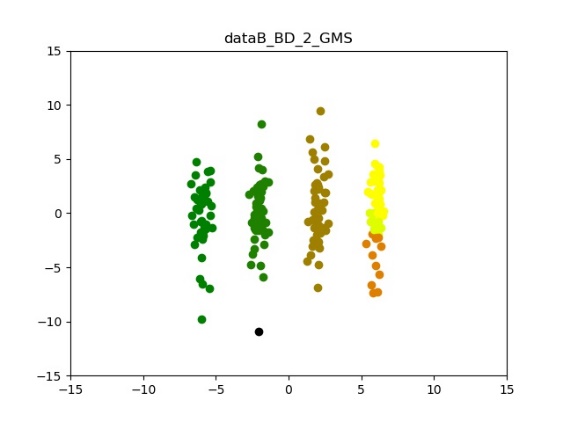
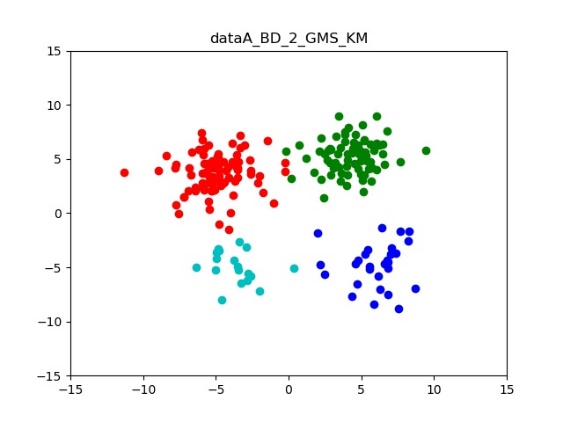
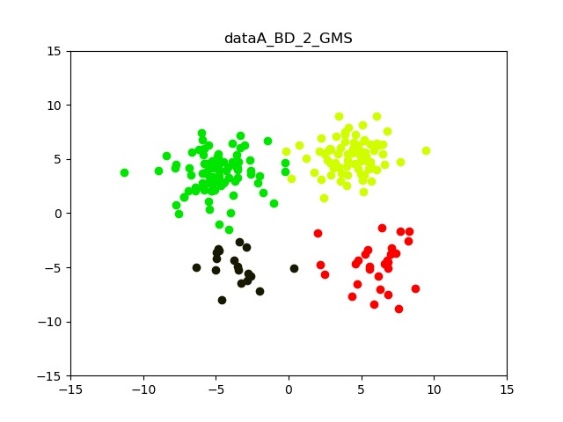


Figure 3 Predict of mean shift clustering with bandwidth equals to 2, color is displayed according to the value of local peak (left) and plots of local peak values clustered by K-means(right)

To analysis the sensitivity of MS bandwidth. We tested different bandwidth values on three datasets. By increasing the bandwidth to 2 in Figure 3, the clustering performance is better than 1. Using KM to limit the cluster value to 4, data C is clustered without fault. But when the bandwidth increases to 10 shown in Figure 4, data point on data A and data B cannot be distinguished while data C can be divided into 2 clusters. When the bandwidth increases to 20, all data set are clustered in one single cluster and failed to form valid clusters.

In summary, when the bandwidth increases, number of clusters will decrease, vice versa. In the presented problem, we have prior know ledge of cluster number so this prior knowledge can help us to generate good result using MS algorithm. And one optimum we tested is bandwidth equals to 2 and using KM to limit the number of local peaks.

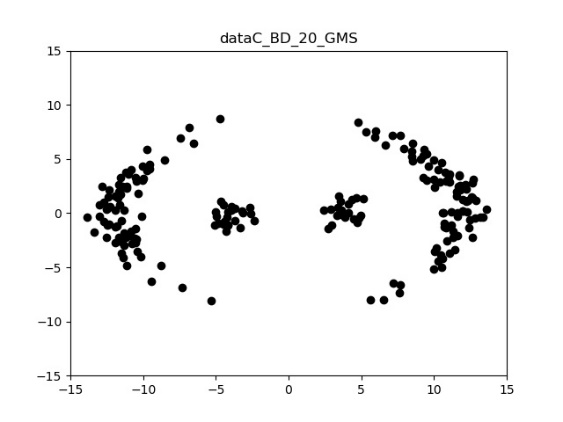
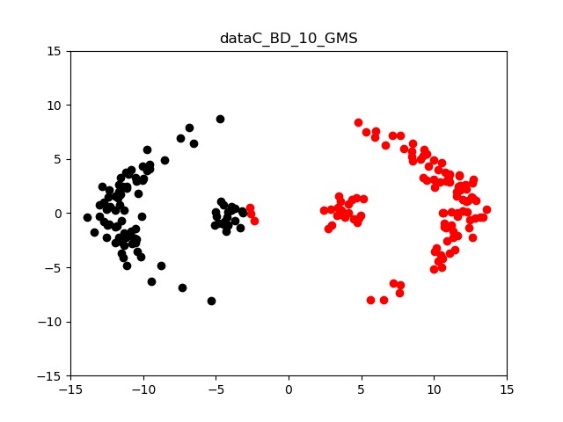
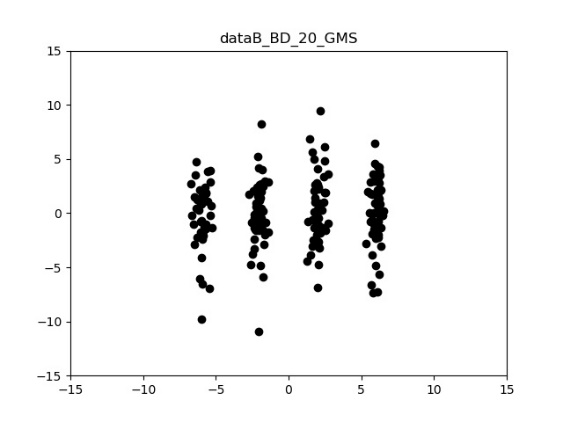
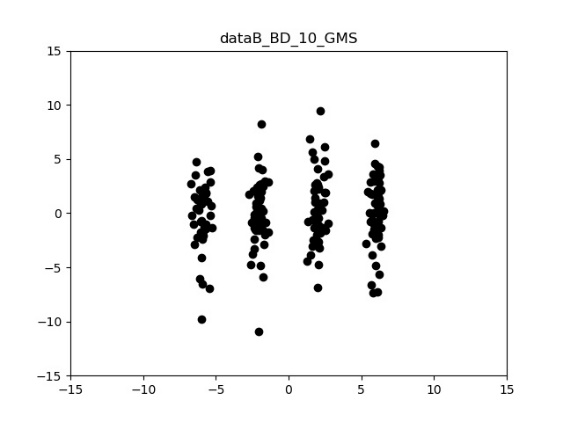
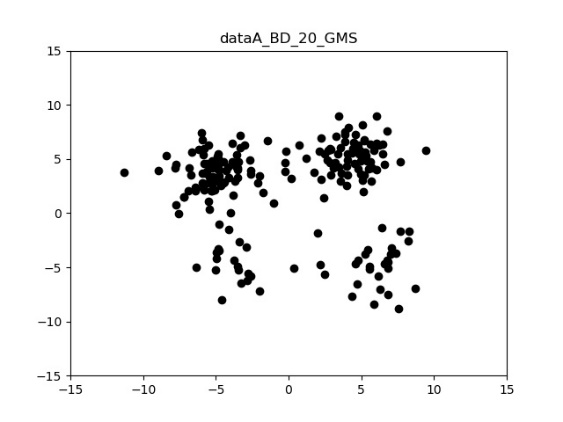
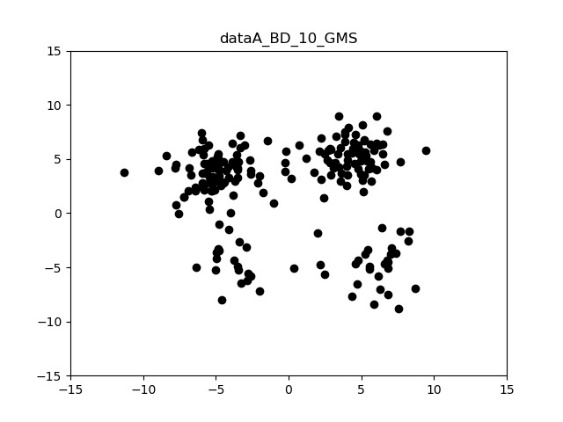


Figure 4 Predict of different bandwidths, the figures are bandwidth = 10 (left) and bandwidth = 20 (right)

# Part 2 A REAL-WORLD clustering problem – image segmentation

## result comparison, hpyer-parameters and properties

## Feature scaling