Home assignment 1

Jan Viilma

March 20, 2018

Contents

1 Distance functions

When evaluating different distance functions, I noticed that the most common, euclidian distance, is the most successful.

2 K-means algorithm

K-means algorithm was noticably more precise when the clusters were shaped as circles. Also when graphing the data, you have to be able to tell the quantity of the clusters. For my datasets, around 10 epochs was enough for the best results. The results didn't really get better with more epochs.

Compared to the built in k-means algorithm, mine didn't performed at least as well. When using the algorithms on the same dataset, the built in was pretty accurate. My algorithm somewhat depended on the random positions of the initial centroids. Because of that the results sometimes varied from better to the same as the built in algorithm.

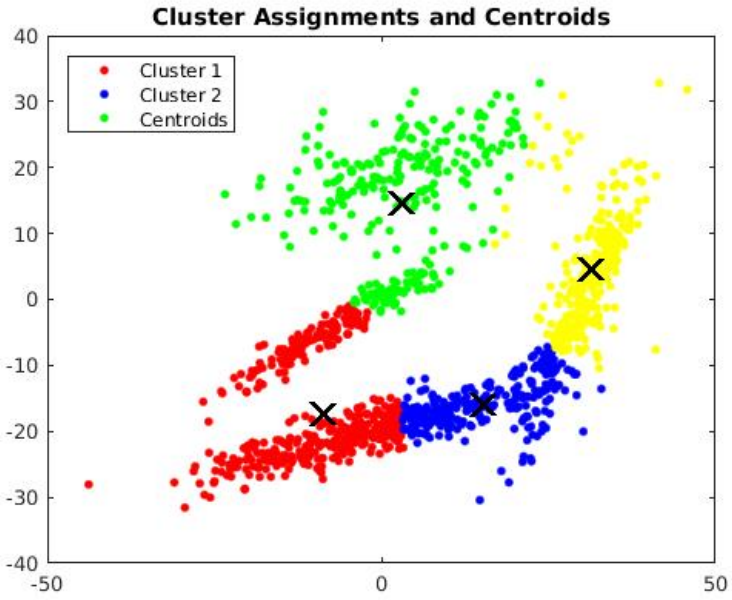

Figure 1: Matlab built in k-means clustering algorithm.

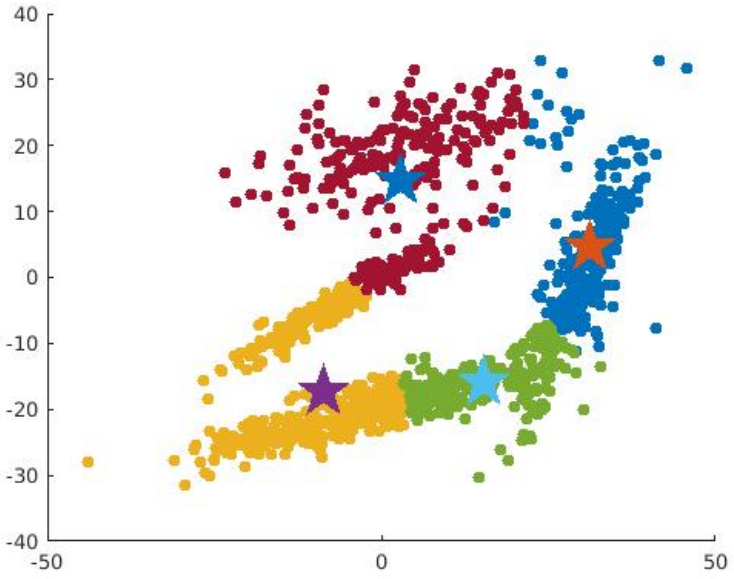

Figure 2: My k-means clustering algorithm.

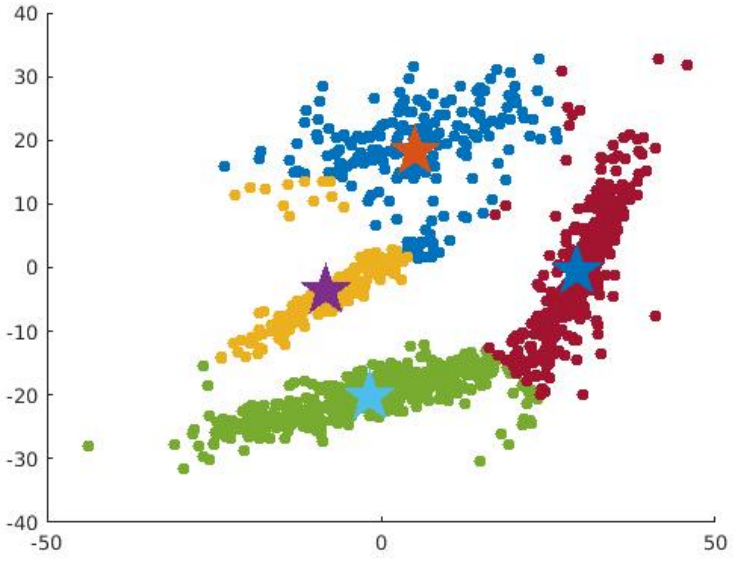

Figure 3: My k-means clustering algorithm (better results).

3 DBSCAN

The most successful algorithm ofcourse, was DBSCAN. Only thing that you had to know, was the scale of the dataset to assign a logical epsilon (distance between core points). The second parameter, you have to assign is the minimum points a core point needs in its epsilon neighbourhood. This algorithm was a lot better compared to k-means, because the clusters didn't have to be shaped circular. When comparing the results with different distance functions, no mentionable differences were noticed, so using the euclidean distance, seems like the choice to go with.

Unfortunately I didn't find any built in dbscan algorithm in my matlab.

The results compared to the k-means algorithm can be seen on the picture below. It didn't classify the two clusters below as well as it could have. By trying different parameters the results can definately be improved. Compared to the k-means algorithm, dbscan also takes a lot more time to get the results. The blue points are classified as noise.

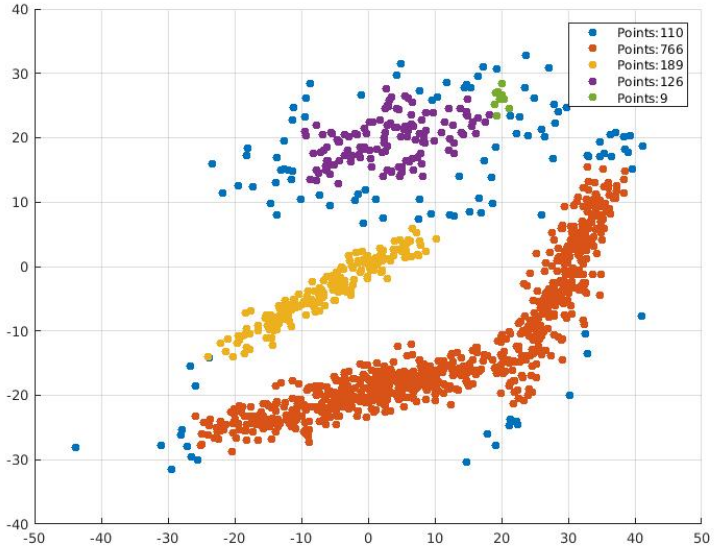

Figure 4: DBSCAN clutering algorithm.