Uppsala University
Department of Information Technology
Kjell Orsborn

DATA MINING I - 1DL360

Results form for Assignment 2

Date: 25/4/2015

Name: Al Hinai, Abdullah
John Chiramel, George
Noorani Subramanian, Varun

KMEANS:

	clusterdata1.nt	clusterdata2.nt
number of clusters (k parameter)	4	5
SSE value for original data	42.905	143.311

DBSCAN:

	clusterdata1.nt	clusterdata2.nt
eps parameter	0.3	0.3
minpts parameter	5	5
number of core points generated	47	99
number of clusters generated	6	6
SSE value for original data	8.426	20.3811

Questions:

1. In what way are the two datasets different?

In data set 1, density and shape are similar for all the cluters whereas data set 2, density is same and outliers are present .

2. Can we compare SSE values between KMEANS and DBSCAN? Why?

No, that beacuse for KMEANS we give pre-specified number of cluster, k, then we just apply the algorithem which minimize the SSE for k numbers centroids. In contrast, DBSCAN is density based depend on eps & minPts values, and not converges to smalles SSE

3. How did you choose the parameters? In particular, how did you choose number of clusters for KMEANS?

Clusters are chosen by inspection. Points which lie close enough are included in a cluster.

4. What are the advantages and disadvantages of the two algorithms?

K means algorithm is advantageous when the shape of clusters is globular and the size of clusters are similar. It works well with small sized data sets. K means has the disadvantages of not able to work with clusters of differing sizes and differing densities. DBscan

5. How can a k-dist graph be used to remove noise when the k-means algorithm is used?

Lets take an example where there are 100 points and 96 points form clusters and k = 5. The remaining 4 points cannot find the 5th smallest point and they end up as noise points

Please give brief answers. In order to get a pass grade,

- your script file must work
- your answers to the questions above must be correct
- your answers at the oral examination must be satisfactory

Date Grade/Comment Signature+Stamp