Data Mining Assignment II
Clustering using K-Means vs. DBSCAN

Submit your solutions to:
Per.Gustafsson@it.uu.se

Due date: 28/11/2004, 12:00 (**NOTE TIME!**)

1 Background

This assignment focusses on two clustering techniques: K-means and DBSCAN. K-means is a partitional algorithm, is one of the most commonly used clustering methods as it is quite easy to understand and implement. DBSCAN [1] is a density-based clustering method. (The paper is available via the course's homepage.) This assignment asks you to implement both algorithms and examine their characteristics on two different 4-dimensional data sets.

2 Assignment & Tasks

The data sets can be found in the file `assignment2.mat` which is accessible through the course's assignments page. When you load this file in MATLAB, you will find two matrices:

1. *Patterns* a 150*4 matrix where each row contains one pattern,
2. *Patterns2* a 200*4 matrix where each row contains one pattern,

These tasks should be performed on both data sets; see Sect. 3 on how you should report on them.

Use principal components analysis (PCA) to project the 4-dimensional pattern-vectors on the two principal components.[1] This makes it possible to view the patterns in two dimensions. This will give you some hints as to how many clusters there are, which in turn will be useful to decide which parameters to use for the different algorithms.

Implement the k-means clustering algorithm and find a suitable value for **k**. Remember that the clustering provided by the k-means algorithm depends on the initial placements of the clusters so it might be wise to make several runs for each **k** and choose the clustering that gives the lowest mean distance to cluster center.

Then implement the DBSCAN algorithm. Set the **MinPts** value to 5. Create a graph of the 5-dist value of the patterns (as described in [1, Sect. 4.2]) and use this to estimate the amount of noise in the data set. Then make a choice of **Eps** that gives you the correct amount of noise.

[1] You can assume that no normalisation is needed. The data is given in the correct units. Use the MATLAB command *princomp* to find the principal components.

3 Report

As in all assignments of this course, you are allowed and encouraged to work in pairs. Your report should contain information about:

- Your names and e-mails.
- Pointer to MATLAB files and running directions. Please make sure that your code runs on MATLAB 6.0. (The version of MATLAB that is started with the `matlab` command in the shell.) Needless to mention, you are not allowed to update these files after the deadline.
- Pointer to a `.mat` file that contains (only) the variables:

  K1 containing your choice of K for the first data set.

  K2 containing your choice of K for the second data set.

  Eps1 containing your choice of Eps for the first data set.

  Eps2 containing your choice of Eps for the second data set.

  ClusterK1 A vector containing a number for each pattern which shows which cluster it belongs to in the clustering performed with **K1**.

  ClusterK2 A vector containing a number for each pattern which shows which cluster it belongs to in the clustering performed with **K2**.

  ClusterEps1 A vector containing the cluster number when DBSCAN was used with **Eps1** as parameter. If the pattern was not assigned to a cluster it should contain a zero.

  ClusterEps1 A vector containing the cluster number when DBSCAN was used with **Eps2** as parameter. If the pattern was not assigned to a cluster it should contain a zero.

- Figures that show the two datasets in the principal component plane after classification. Use different colors or different markers to show what cluster each data point belongs to for all four clusterings that you submit. Use *Export* from the menu in the figure window and choose the *color eps*-format (can be viewed in ghostview).
- A discussion on the differences of the two datasets based on both the two dimensional PCA representation and the results of the clustering.
- The reasoning behind your choice of parameters; in particular how you've choosen the values **K1, K2** for k-means.
- A discussion of pros and cons of the two algorithms that you have examined.
- A suggestion how a k-dist graph can be used to remove noise when the k-means algorithm is used.

Good luck!

References

[1] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226-231, 1996.