

## Assignment 4

This assignment is due at 10, April, 2006. Early hand-in is encouraged.

- 4.1 The leader algorithm represents each cluster by a point, known as a leader, and assign each point to the cluster corresponding to the closest leader, unless this distance is above a user-specified threshold. In that case, the point becomes the leader of a new cluster. Compare the advantages and disadvantages of the leader algorithm to K-means and propose ways to improve the leader algorithm.
- 4.2 Program the k-means and EM algorithms, respectively in C or matlab to cluster a given set of points in two dimension into clusters. The input of your program includes the point set  $S$ , the initial psarameters used in the algorithm and  $k$ , number of clusters. The output of your program should give the final clusters and their centers. You should hand in hard copy of your code and printout of your example questions. You are also required to give examples to show the influence of different starting centers and the number  $k$  of clusters on the final outcome. You can send TA the source code of your program by email for testing.
- 4.3 Give few ways to represent documents by vectors in a certain space and discuss the advantages and disadvantages of text mining algorithms based on vector representation.