CS722/822 MACHINE LEARNING

Homework #4

Due: Nov. 30th, end of the day

This is a programming homework assignment. You need to upload your code and a report to Blackboard by the specified due date. You can use either Python (preferred), or Matlab, or any other programming language in your choice.

1. Implement your own K-means algorithm. Apply the algorithm to the dataset stored in file "A.txt" coming with this document. This dataset includes the coordinates of 174 2-D data points as plotted in the following figure.

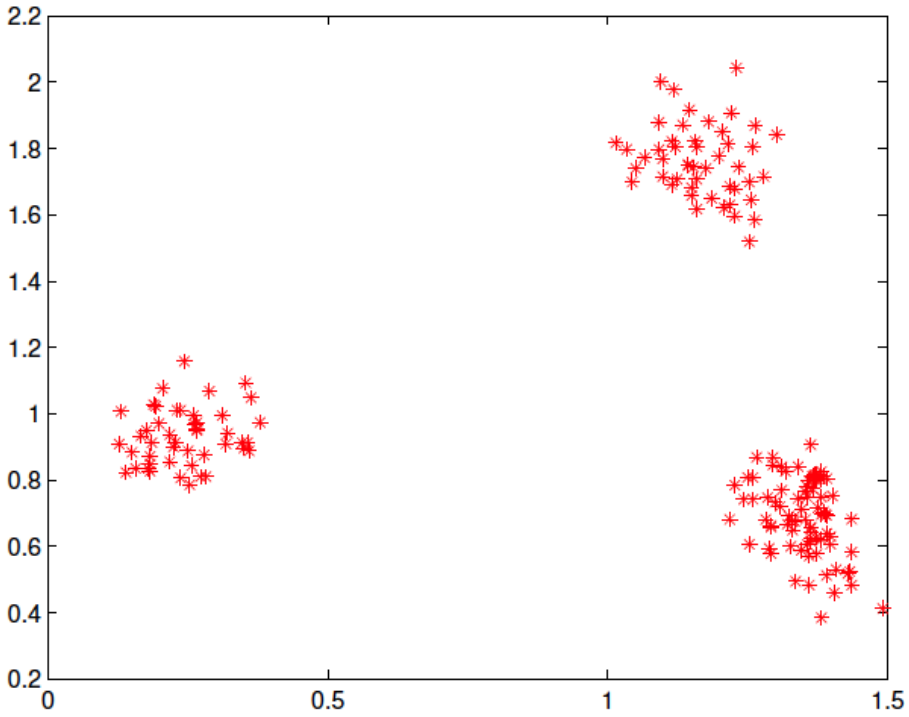

Figure 1: Input for K-means algorithm

(1) Run your algorithm with different choices of K's (ranging from 2 to 10 with step size 1). Calculate the Sum of Squared Errors (SSE) for clustering resulted from each run, and plot SSE against K in a figure as shown on 7th slide from lecture 16. Note for each K, multiple runs of the algorithm with different initial centroids may be needed, in order to have a curve that is monotonically non-increasing.

(2) Run your algorithm with K=3. Show the clustering results in a figure with varying colors or shapes of data points to represent different cluster assignment (as in those figures on

6th slide from lecture 16). Note multiple runs may be needed in order to obtain a satisfactory clustering result.

2. Implement your own agglomerative hierarchical clustering algorithm with Euclidean distance to measure distance between any pair of data points and four different ways to measure inter-cluster similarity: MIN, MAX, Group Average, and Distance between Centroids. Hierarchical clustering can work directly with distance matrix. However, if you prefer to work on similarity (proximity) matrix as in the examples shown in class, you can use the following formula to covert distance to similarity, $s_{ij} = 1/(1 + d_{ij})$, where s_{ij} represents the similarity between any two points i and j, and d_{ij} is the distance between the two.

   Apply your algorithm to the dataset stored in "B.txt" coming with this document to obtain **two** clusters. This dataset includes the coordinates of 218 2-D data points as plotted in Figure 2, with data points at the center from one cluster and those surrounding from the other. Show the clustering solutions resulted from all four ways of defining inter-cluster similarity in plots, with different signs (color or shape) representing different cluster assignment. Which inter-cluster similarity measure gives you the desired solution? Why?

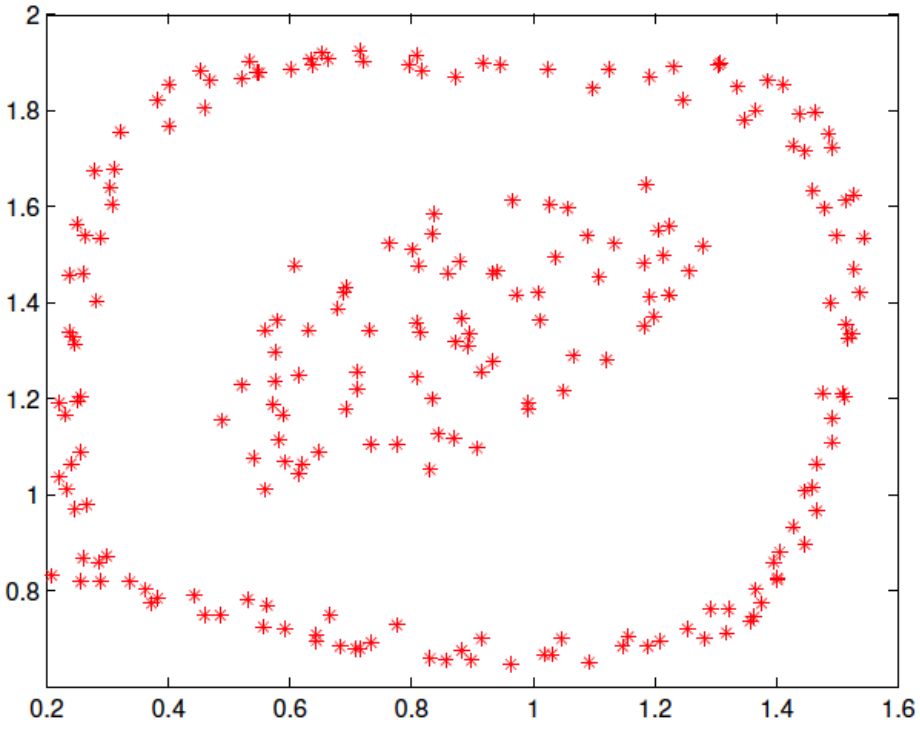

Figure 2: Input for Hierarchical Clustering Algorithm