COMP4432 Machine Learning

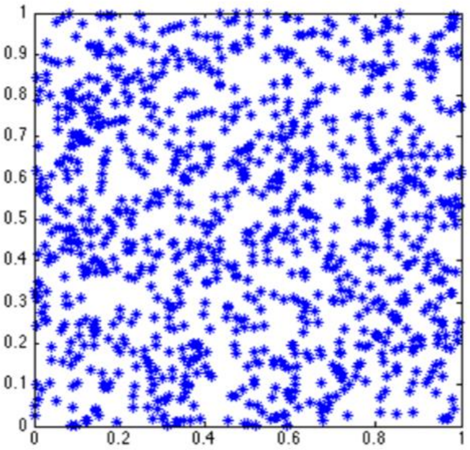
Tutorial Questions on Clustering

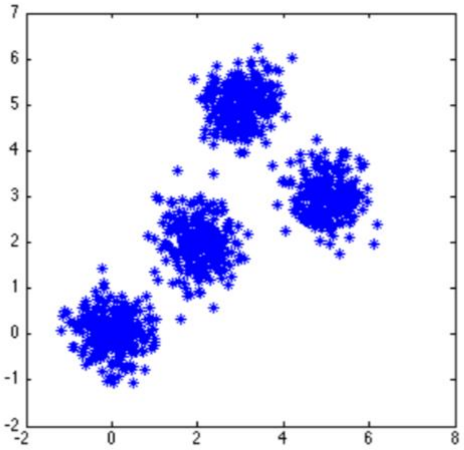
1. You are asked to use the k-means algorithm to cluster the following 8 examples into 3 clusters:
   P1=(2,10), P2=(2,6), P3=(8,4), P4=(5,8), P5=(7,4), P6=(6,4), P7=(1,2), P8=(4,9).

   (a) Compute the distance matrix based on the Euclidean distance.

   (b) Suppose that the initial seeds (centroids of each cluster) are P1, P4 and P7. Run the k-means algorithm for 1 iteration ONLY and then write down:
   i) The new clusters (i.e. the examples belonging to each cluster)
   ii) The centroids of the new clusters

2. Given the following two artificial datasets with 1000 2-D points each. We want to find 4 clusters in each of them by using k-mean clustering. Given two examples for each of them to illustrate their sensitivity to initialization.





3. Suppose you are asked to provide data mining consulting services to an Internet DVD shop. After interviewing the shop's manager and the database administrator, the following movie database is collected.

Movie Database

Movie ID	Movie Name	Types
0001	The Hobbits	Story, Sci-Fiction, Drama, Mystery
0150	The Intern	Drama, Story, Fiction
0553	Avengers II	Action, Sci-Fiction, Thriller, Horror
1011	Poltergeist	Horror, Thriller
3997	Batman Vs Superman	Action, Crime, Sci-Fiction

   a) If you are asked to cluster the movies, propose an appropriate dissimilarity measure for it and prepare the following dissimilarity matrix for the five movies in the database above.

$$\begin{array}{cc} & \begin{array}{ccccc} 0001 & 0150 & 0553 & 1011 & 3997 \end{array} \\ \begin{array}{c} 0001 \\ 0150 \\ 0553 \\ 1011 \\ 3997 \end{array} & \begin{bmatrix} 0 & & & & \\ \_\_\_ & 0 & & & \\ \_\_\_ & \_\_\_ & 0 & & \\ \_\_\_ & \_\_\_ & \_\_\_ & 0 & \\ \_\_\_ & \_\_\_ & \_\_\_ & \_\_\_ & 0 \end{bmatrix} \end{array}$$

b) Based on your dissimilarity matrix obtained in part (a), apply the single linkage agglomerative clustering algorithm to cluster the five movies. Draw the dendrogram obtained.

c) The single linkage agglomerative clustering has been suffering from the weakness of low scalability (high time complexity). Other than the traditional sampling approach, propose a NEW way to speed up its computation.