COMP4432 Machine Learning

Tutorial Questions on Clustering (with answers)

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.

   Ans.

	P1	P2	P3	P4	P5	P6	P7	P8
P1	0	$\sqrt{16}$	$\sqrt{72}$	$\sqrt{13}$	$\sqrt{61}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
P2		0	$\sqrt{40}$	$\sqrt{13}$	$\sqrt{29}$	$\sqrt{20}$	$\sqrt{17}$	$\sqrt{13}$
P3			0	$\sqrt{25}$	$\sqrt{1}$	$\sqrt{4}$	$\sqrt{53}$	$\sqrt{41}$
P4				0	$\sqrt{20}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
P5					0	1	$\sqrt{40}$	$\sqrt{34}$
P6						0	$\sqrt{29}$	$\sqrt{29}$
P7							0	$\sqrt{58}$
P8								0

   (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

   Ans.

   i) *d(a,b)* denotes the Eucledian distance between *a* and *b*. It is obtained directly from the distance matrix or calculated as follows: $d(a,b)=sqrt((x_b-x_a)^2+(y_b-y_a)^2))$

   seed1=P1=(2,10), seed2=P4=(5,8), seed3=P7=(1,2)

   Iteration (Epoch) 1– start:
   P1:
   d(P1, seed1)=0 as P1 is seed1
   d(P1, seed2)= $\sqrt{13}$ >0
   d(P1, seed3)= $\sqrt{65}$ >0
   ➔ P1 ∈ cluster1

   P2:
   d(P2,seed1)= 4
   d(P2, seed2)= $\sqrt{13}$ smaller
   d(P2, seed3)= $\sqrt{17}$
   ➔ P2 ∈ cluster 2

   P3:
   d(P3,seed1)= $\sqrt{72}$
   d(P3, seed2)= $\sqrt{25}$ smaller

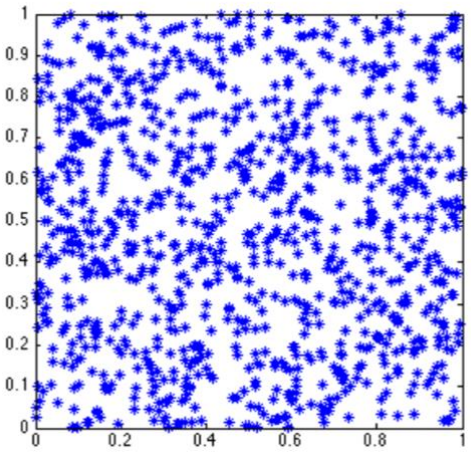
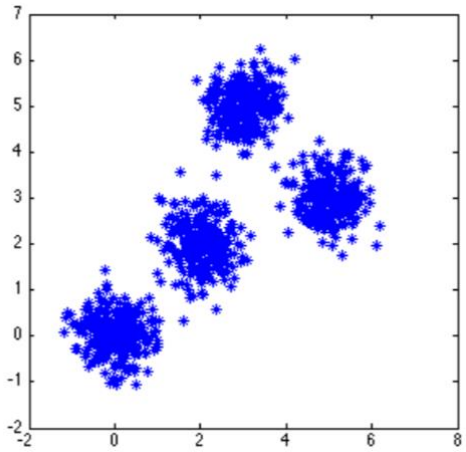
d(P3, seed3)= $\sqrt{53}$
➔ P3 ∈ cluster 2

Similarly,
P4∈cluster 2, P5∈cluster 2, P6∈cluster 2, P7∈cluster 3, P8∈cluster 2
Iteration 1- end

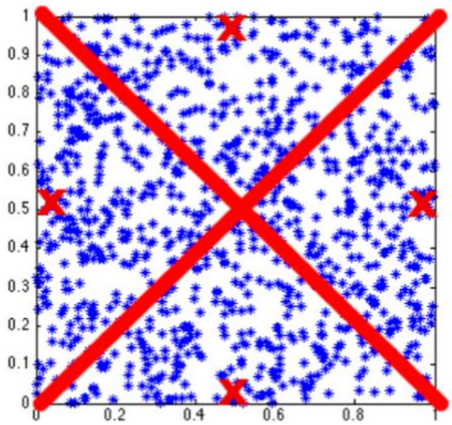
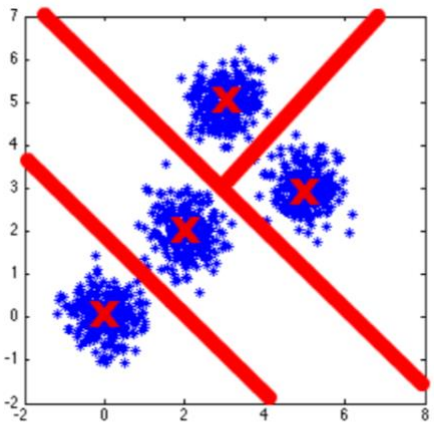
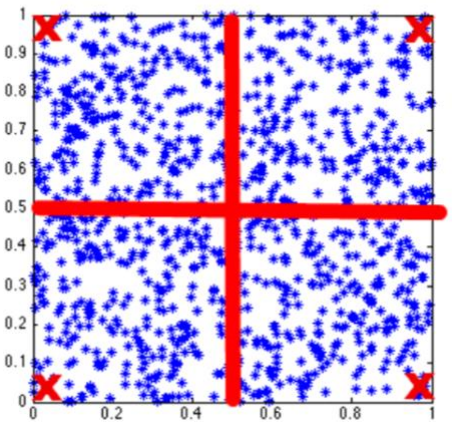
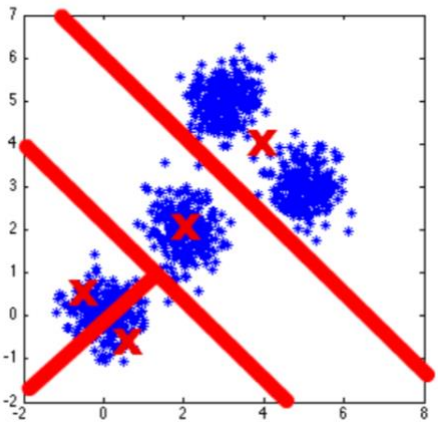
Thus, the new clusters are:
Cluster 1: {P1}; Cluster 2: {P2, P3, P4, P5, P6, P8}; Cluster 3: {P7}

ii) Centers of the new clusters:
C1 = (2, 10)
C2 = ((2+8+5+7+6+4)/6, (6+4+8+4+4+9)/6) = (5.33, 5.83)
C3 = (1, 2)

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.

Answer:

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.

	0001	0150	0553	1011	3997
0001	0				
0150	___	0			
0553	___	___	0		
1011	___	___	___	0	
3997	___	___	___	___	0

Ans.
By using the dissimilarity function:

$$Dissim(T_1, T_2) = 1 - \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

We have the following dissimilarity matrix:

	0001	0150	0553	1011	3997
0001	0	–	–	–	–
0150	3/5	0	–	–	–
0553	6/7	1	0	–	–
1011	1	1	1/2	0	–
3997	5/6	1	3/5	1	0

Other dissimilarity functions can be used. For example, by considering Drama, Mystery, Romance, Action, Thriller, Horror, Sci-Fiction, Fiction and Crime as binary attributes, one may use simple matching coefficient.

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.

Ans.
According to the dissimilarity matrix above, we need to
i) merge movies 0553 and 1011, then
ii) merge movies 3997 with new cluster for (0553 & 1011), then
iii) merge movies 0001 with 0150, and finally
iv) merge the new cluster in iii) and that in ii).
Hence, the dendrogram obtained will have

- movies 0553 and 1011 merged at 0.5 to form a new cluster C6
- movies 3997 and C6 merged at 0.6 to form a new cluster C7
- movie 0001 and 0150 merged at 0.6 to form a new cluster C8
- clusters C7 and C8 merged at 5/6 to form a single cluster for the five movies

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.

Ans.
One may make use of the *k*-means like algorithm, which is computationally efficient, to carry out the partition-based clustering first, say for k=100 (clusters). Based on the obtained 100 cluster centroids (i.e. n=100), the single linkage agglomerative clustering can then be executed efficiently. In other words, a hybridization of *k*-means and single linkage agglomerative clustering is proposed.