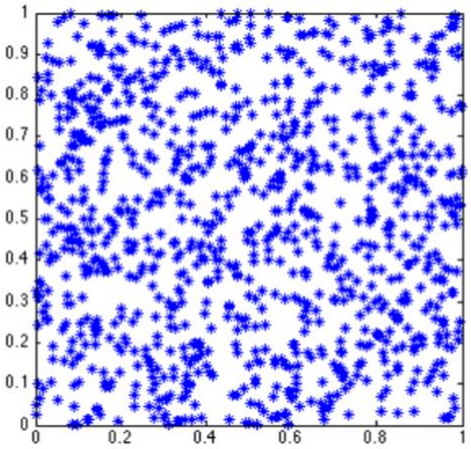
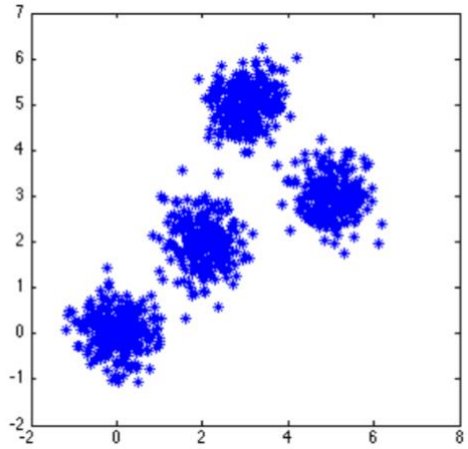
COMP4433 Data Mining and Data Warehousing

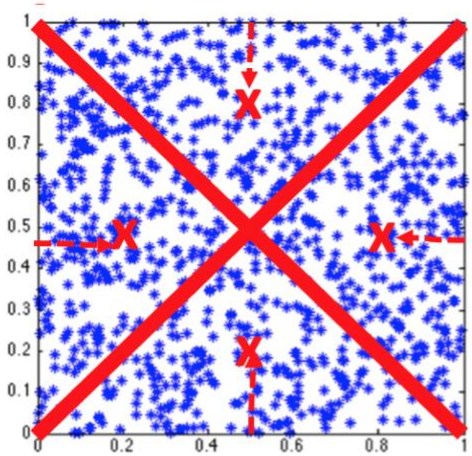
FAQ on Clustering I (with suggested answers)

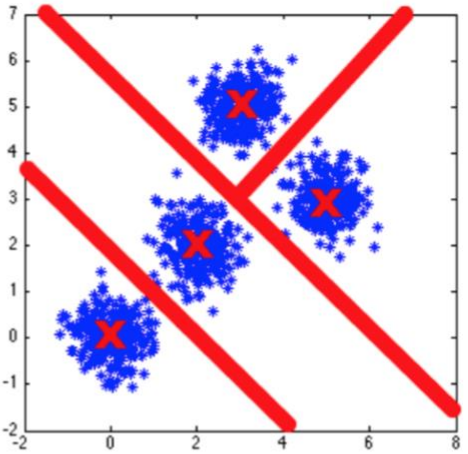
1. Given the following artificial datasets with 1000 2-D points each. We want to find 4 clusters in each of them by using k-mean clustering. Give two examples for each of them to illustrate their sensitivity to initialization.

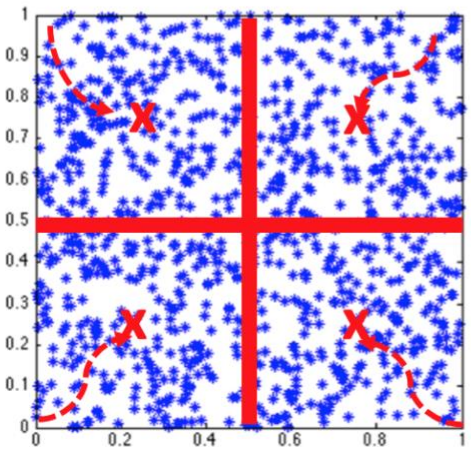


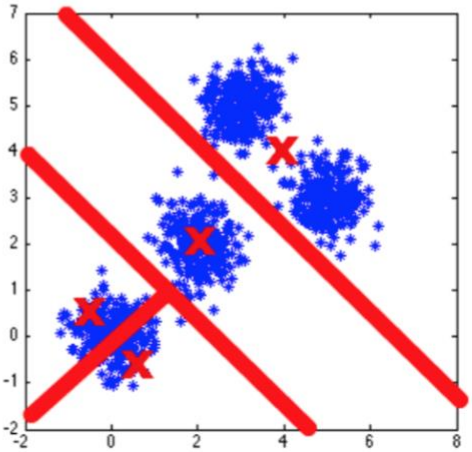


Answer:









2. Given the following medical data records where all attributes except *gender* are asymmetric.

Name	*Gender*	*Fever*	*Cough*	*Test-1*	*Test-2*	*Test-3*	*Test-4*
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
Nick	M	N	N	N	P	N	N
Elaine	F	Y	N	N	N	N	N

Name	~~*Gender*~~	*Fever*	*Cough*	*Test-1*	*Test-2*	*Test-3*	*Test-4*
Jack	~~M~~	1	0	1	0	0	0
Mary	~~F~~	1	0	1	0	1	0
Jim	~~M~~	1	1	0	0	0	0
Nick	~~M~~	0	0	0	1	0	0
Elaine	~~F~~	1	0	0	0	0	0

a) Compute the missing Jaccard coefficients to complete the matrix above.
Answer

	Jack	*Mary*	*Jim*	*Nick*	*Elaine*
Jack	0	–	–	–	–
Mary	0.33	0	–	–	–
Jim	0.67	0.75	0	–	–
Nick	1	1	1	0	–
Elaine	0.5	0.67	0.5	1	0

b) Cluster the data records using the single-link agglomerative clustering algorithm and the Jaccard coefficient matrix computed in part (a). Make your own assumption(s) if necessary.

Merging Jack and Mary (d=0.33), we have

	J & M	*Jim*	*Nick*	*Elaine*
J & M	0	–	–	–
Jim	0.67	0	–	–
Nick	1	1	0	–
Elaine	0.50	0.50	1	0

If merging of more than 2 records is allowed, J&M, Jim and Elaine should be merged next. Thus, the last record being grouped is Nick.

c) Based on the result of part (b), divide the records into two clusters. Could we obtain three clusters?

Since 2 groups are required, they are formed when the inter-cluster distance using Jaccard coefficient is larger 0.5 but less than 1.
Group 1: Jack, Mary, Jim, Elaine
Group 2: Nick
It is not reasonable to split the data into 3 clusters.

3. Given the following web page content database records.

URL	Web Page ID	Keywords Found					
		Popstar	Actor	Actress	Music	Movie	Holly-wood
Jackchan.com	P100	√	√			√	√
Nictsz.com	P200	√	√		√		
Faywang.com	P300			√	√	√	√
Allantam.com	P400		√		√	√	
SammyChen.com	P500	√		√	√	√	

By considering the occurrence of a keyword as a symmetric binary attribute, a partially filled simple matching coefficient matrix is depicted below. Here, the present of a keyword is set to 1 while its absent is set to 0.

	*P*100	*P*200	*P*300	*P*400	*P*500
*P*100	0	–	–	–	–
*P*200	0.5	0	–	–	–
*P*300			0	–	–
*P*400		0.33		0	–
*P*500				0.5	0

a) Compute and fill in the missing simple matching coefficients in the matrix above.

Answer:

URL	Web Page ID	Keywords Found					
		Popstar	Actor	Actress	Music	Movie	Holly-wood
Jackchan.com	P100	1	1	0	0	1	1
Nictsz.com	P200	1	1	0	1	0	0
Faywang.com	P300	0	0	1	1	1	1
Allantam.com	P400	0	1	0	1	1	0
SammyChen.com	P500	1	0	1	1	1	0

	*P*100	*P*200	*P*300	*P*400	*P*500
*P*100	0	–	–	–	–
*P*200	0.5	0	–	–	–
*P*300	0.66	0.83	0	–	–
*P*400	0.5	0.33	0.5	0	–
*P*500	0.66	0.5	0.33	0.5	0

b) Based on the coefficient matrix completed in part (a), cluster the data records using the single-link agglomerative hierarchical clustering algorithm.

Answer:

1st round: Merging P200 & P400 (distance=0.33)
2nd round: Merging P300 & P500 (distance=0.33)
3rd round: Merging C1(P200,P400) to C2(P300, P500) (distance=0.5) or
Merging C1(P200,P400) to P100 (distance=0.5)
4th round: Merging the remaining two clusters
Detail steps are omitted here.