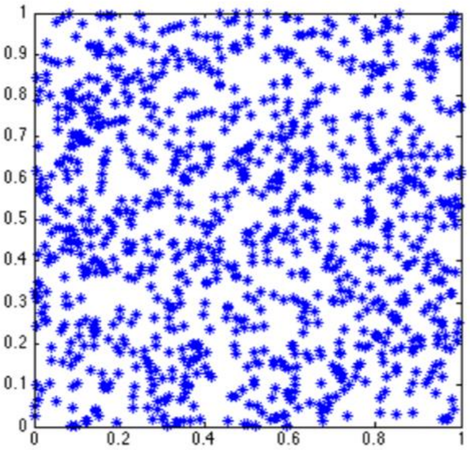
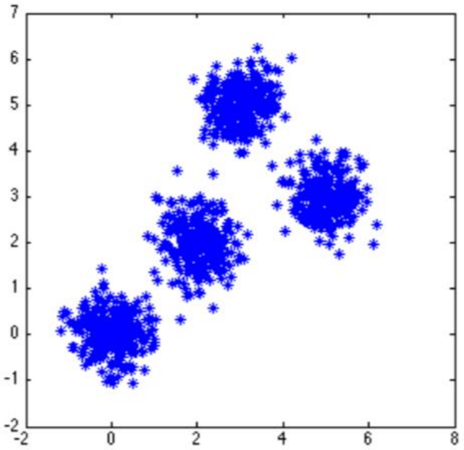
COMP4433 Data Mining and Data Warehousing

FAQ on Clustering I

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.





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

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

$$\begin{array}{c|ccccc} & \textit{Jack} & \textit{Mary} & \textit{Jim} & \textit{Nick} & \textit{Elaine} \\ \hline \textit{Jack} & 0 & - & - & - & - \\ \textit{Mary} & 0.33 & 0 & - & - & - \\ \textit{Jim} & 0.67 & 0.75 & 0 & - & - \\ \textit{Nick} & 1 & & & 0 & - \\ \textit{Elaine} & 0.5 & & & & 0 \end{array}$$

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.

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

3. Given the following web page content database records.

URL	Web Page ID	Keywords Found					
		Popstar	Actor	Actress	Music	Movie	Hollywood
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.

$$\begin{array}{c|ccccc} & P100 & P200 & P300 & P400 & P500 \\ \hline P100 & 0 & - & - & - & - \\ P200 & 0.5 & 0 & - & - & - \\ P300 & & & 0 & - & - \\ P400 & & 0.33 & & 0 & - \\ P500 & & & & 0.5 & 0 \end{array}$$

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

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