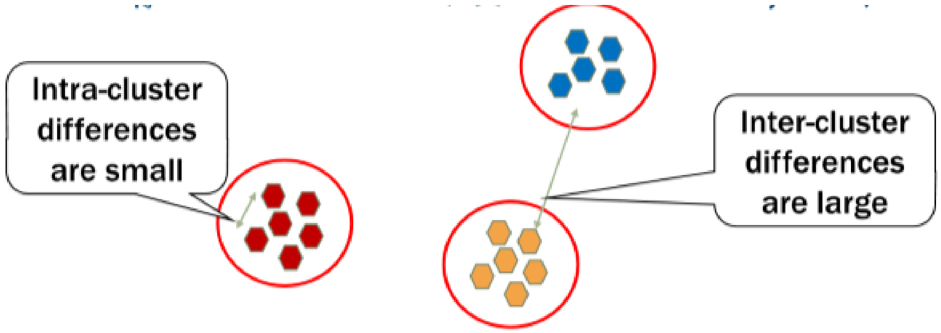
1155173051
Jirapong SAELOR
1155173051@link.cuhk.edu.hk

Lecture 7 Clustering

1. Introduction to clustering

What is clustering?

- Clustering is a process to group similar objects together based on their common properties.
- Clustering is performed by identifying the pattern of the object such that object in the same group will be similar to each other(small intra-cluster difference) and different from the other group.
- 
Intra-cluster differences are small
Inter-cluster differences are large
-

Why do we need to perform clustering?

- Clustering is beneficial to manage and organize a database especially for better understanding of data distribution and to summarize the database.
- In biology, clustering can be used to group genes with different expressions or further understand how the gene in each sample is related.

How to perform clustering

Clustering requires three components.

1. Data to be clustered.
2. Measurement or index identification to separate and label the data.
3. Clustering algorithm to identify the distance to distinguish data into a different cluster.

2. Similarity and dissimilarity

Similarity is a measure of similarity of the two datasets. The higher similarity means two objects are more similar.

Dissimilarity represents the difference between two data sets. The higher dissimilarity value refers to the less similarity of two data sets.

Example of similarity measurement method.

- Cosine similarity

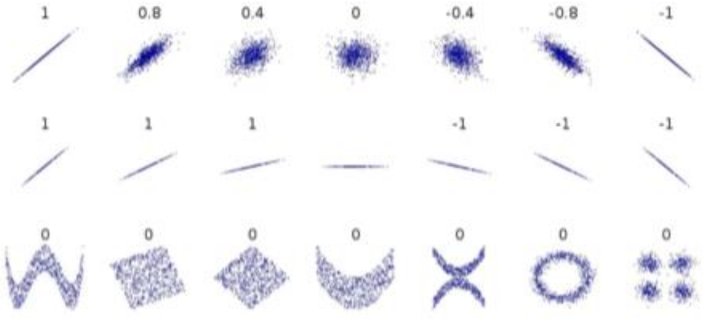
- Cosine similarity calculates the similarity of two vectors in term of angle difference. The closer the two vectors, the more Cosine similarity value(theta is zero). Likewise, the less similar the closer value to 0 (theta is 90).
- Cosine similarity can be calculated by dividing dot product of two vectors by the multiplication of two vectors norm.

$$➢ cos(d_1, d_2) = \frac{d_1 \cdot d_2}{(|d_1| * |d_2|)}$$

- Correlation
  - Correlation holds a value of linear relationship between two objects. The value of correlation ranging from 0 to 1 when 0 represents no correlation of two objects and 1 means it totally correlated. Thus, correlation is a measure of similarity.
  - The correlation can be calculated by dividing the covariance by standard deviation of two dataset. Covariance is a summation of any datapoint subtracting the mean.

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

  - Example of different correlation



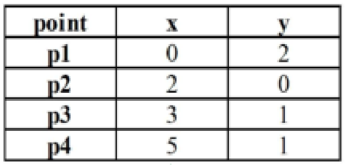
Yu Li

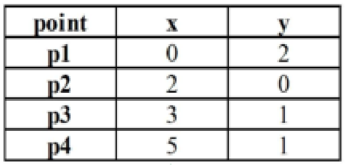
Measurement of dissimilarity

- Euclidean distance is a function to calculate the distance of any two data points with m dimensions.

$$Ed(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{k=1}^{m} (p_k - q_k)^2}$$

  - Intuitively, it is just the normal distance calculation with a dimension of m and Pm is a component in P vector of m dimension.
  - The operation of Euclidean distance function takes place with different data range. We can scale the data range from different sources by normalizing it to make it more comparable.

- Minkowski distance
  - Minkowski distance is a measured distance of two points in N dimensional space.
  - Minkowski distance introduced one more variable to the distance measurement system which is the r parameter.
  - $$dist(p,q)=\left(\sum_{k=1}^{m}|p_k-q_k|^r\right)^{\frac{1}{r}}$$
  - Euclidean distance is just a subset of Minkowski distance with r = 2.
  - **When r equal to one**, Minkowski distance indicates summation of difference between two data points.
    - 

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1
    - Dist(P1,P2) = (P1x-P2x) + (P1y-P2y) = 2+2 = 4
  - **When r is approaching infinity**, Minkowski distance, or Chebyshev distance, will be a supremum or the maximum difference between any component of the vector.
    - 

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1
    - Minkowski distance of P1P2 when r is approaching infinity will be a max(|P1x-P2x|, |P1Y-P2Y|)
      - Max(|0 – 5|, |2-1|) = 5

3. Hierarchical clustering

A hierarchical clustering nest a multiple cluster together as a hierarchical tree. It can be visualized as a dendrogram which illustrates a hierarchical structure of a data set.

Hierarchical clustering method

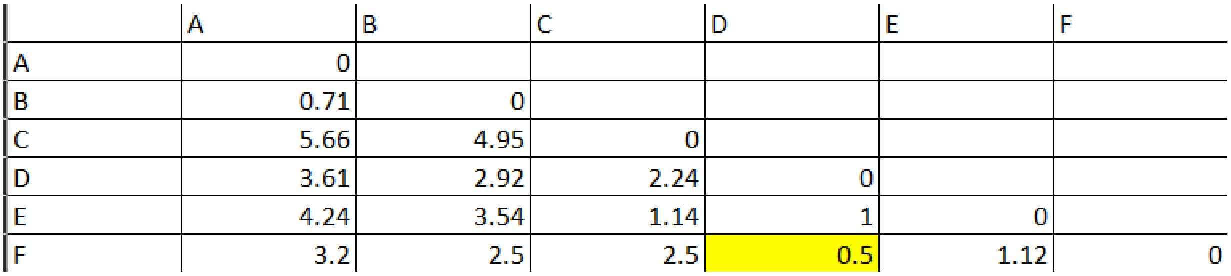
1. Calculate a similarity or distance matrix of all data points.
2. Treat all data points as a cluster.
3. Merge two closet data points as a same cluster.
4. Recalculate the distance between new data points(merged data points with other data points) and update the similarity or distance matrix.
   a. Distance can be recalculated by several method including,
      i. Maximum distance of two clusters.
      ii. Minimum distance of two clusters.
      iii. Average of the group.
      iv. Distance between centroids of two clusters.
      v. Others
5. Repeat step three and four until only a single cluster remains.

1155173051
Jirapong SAELOR
1155173051@link.cuhk.edu.hk

Hierarchical clustering example:

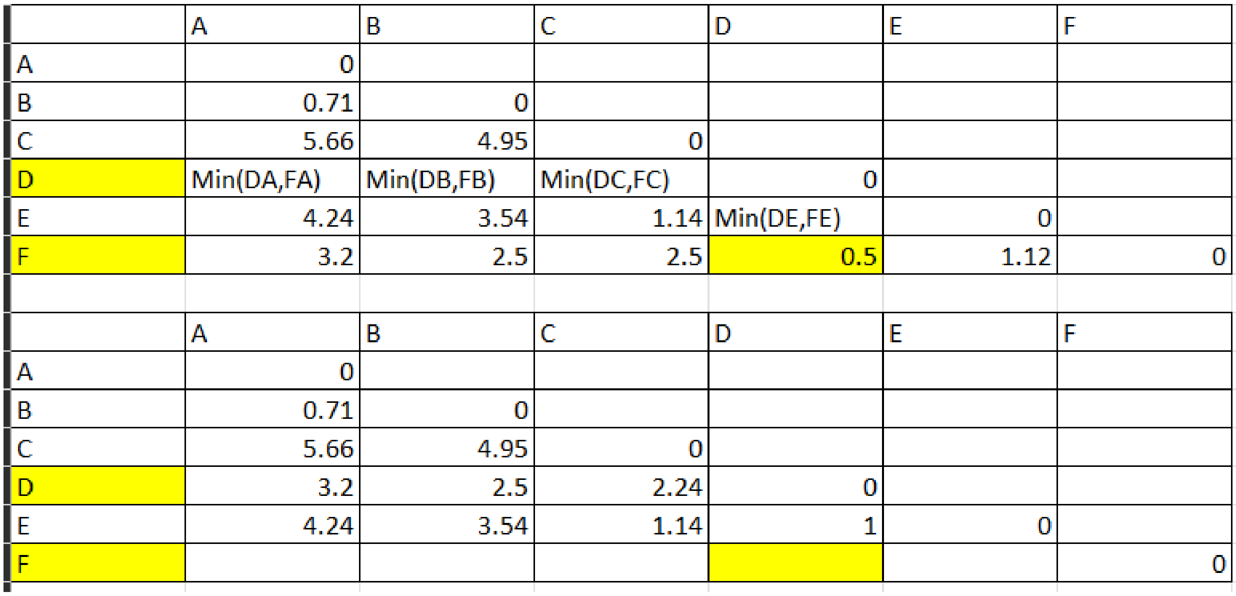
Assume that recalculate distance matrix using the minimum value of two data points. Given a similarity matrix here below.

1. Treat every data point as a cluster and now we are merging data with the closet distance (data point D and F).



	A	B	C	D	E	F
A	0					
B	0.71	0				
C	5.66	4.95	0			
D	3.61	2.92	2.24	0		
E	4.24	3.54	1.14	1	0	
F	3.2	2.5	2.5	0.5	1.12	0

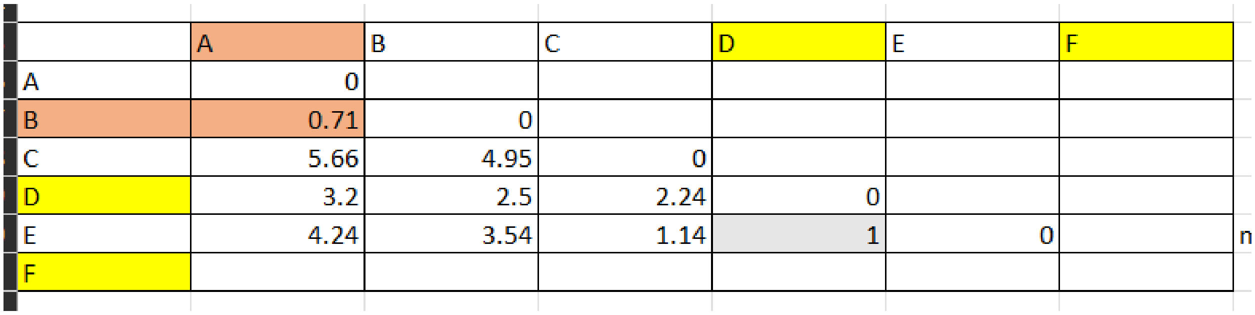
2. We need to recalculate the DA-FA, DB-FB, DC-FC, and DE-FE.



	A	B	C	D	E	F
A	0					
B	0.71	0				
C	5.66	4.95	0			
D	Min(DA,FA)	Min(DB,FB)	Min(DC,FC)	0		
E	4.24	3.54	1.14	Min(DE,FE)	0	
F	3.2	2.5	2.5	0.5	1.12	0

	A	B	C	D	E	F
A	0					
B	0.71	0				
C	5.66	4.95	0			
D	3.2	2.5	2.24	0		
E	4.24	3.54	1.14	1	0	
F						0

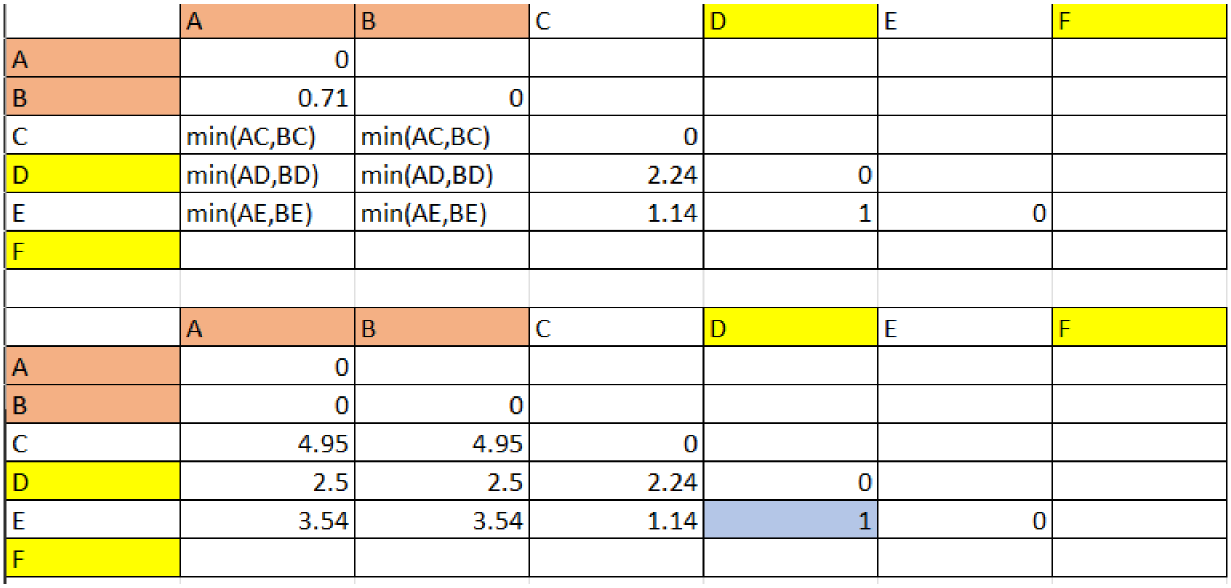
3. Then identify a new closest value and merge them. In this case we will merge A and B since the distance is 0.71.



	A	B	C	D	E	F	
A	0						
B	0.71	0					
C	5.66	4.95	0				
D	3.2	2.5	2.24	0			
E	4.24	3.54	1.14	1	0		n
F							

1155173051
Jirapong SAELOR
1155173051@link.cuhk.edu.hk

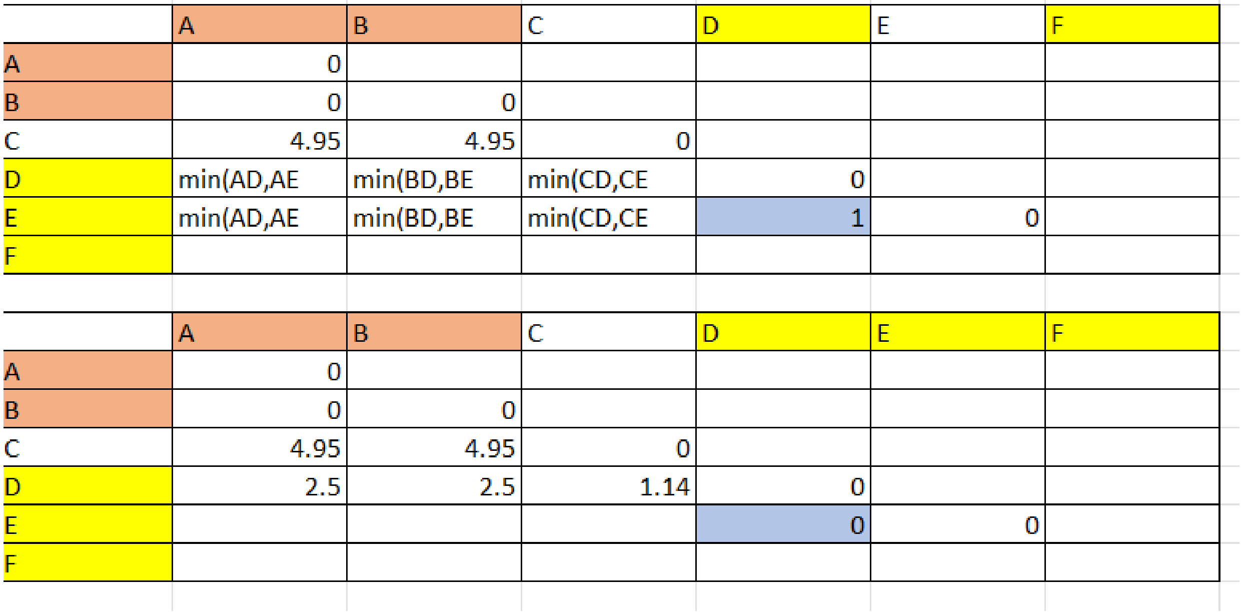
4. Recalculated distance of clsuter AB with C, E, and cluster DF



	A	B	C	D	E	F
A	0					
B	0.71	0				
C	min(AC,BC)	min(AC,BC)	0			
D	min(AD,BD)	min(AD,BD)	2.24	0		
E	min(AE,BE)	min(AE,BE)	1.14	1	0	
F						

	A	B	C	D	E	F
A	0					
B	0	0				
C	4.95	4.95	0			
D	2.5	2.5	2.24	0		
E	3.54	3.54	1.14	1	0	
F						

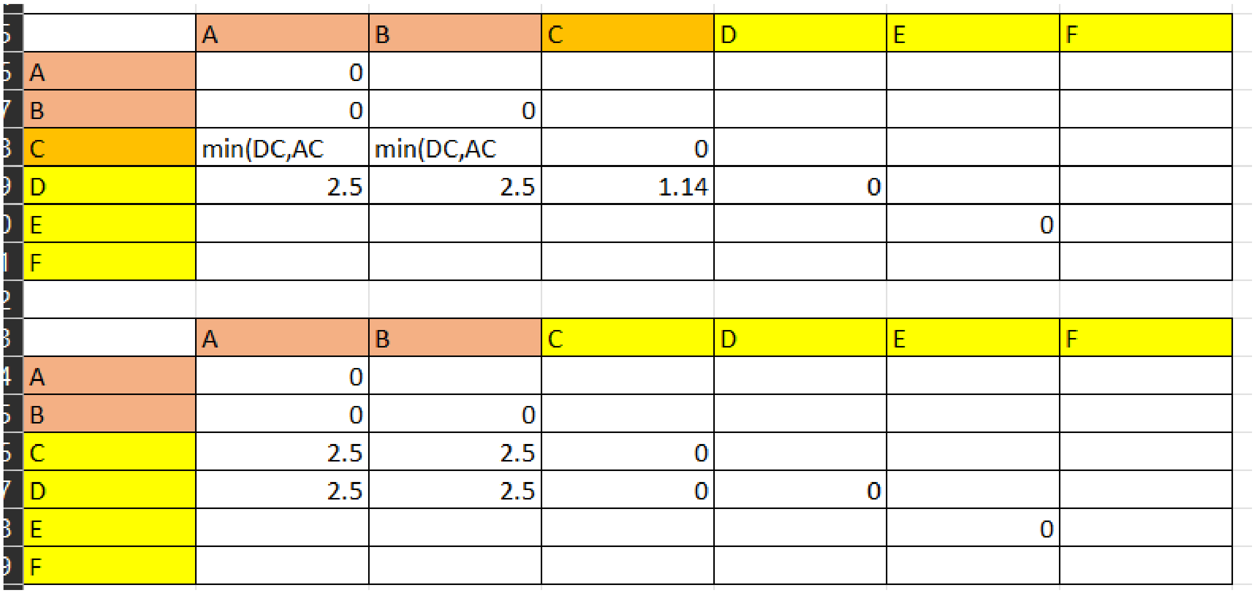
5. Then identify a new closest value and merge them. In this case we will merge E and cluster DF since the distance is 1. We need to recalculate the distance of E with cluster AB and cluster D.



	A	B	C	D	E	F
A	0					
B	0	0				
C	4.95	4.95	0			
D	min(AD,AE	min(BD,BE	min(CD,CE	0		
E	min(AD,AE	min(BD,BE	min(CD,CE	1	0	
F						

	A	B	C	D	E	F
A	0					
B	0	0				
C	4.95	4.95	0			
D	2.5	2.5	1.14	0		
E				0	0	
F						

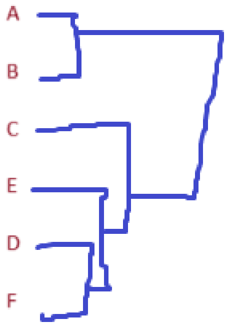
6. The next closet distance is C and cluster DEF. We need to recalculate the distance between cluster AB with CDEF.



	A	B	C	D	E	F
A	0					
B	0	0				
C	min(DC,AC	min(DC,AC	0			
D	2.5	2.5	1.14	0		
E					0	
F						

	A	B	C	D	E	F
A	0					
B	0	0				
C	2.5	2.5	0			
D	2.5	2.5	0	0		
E					0	
F						

This cluster can be illustrated with this dendrogram diagram.



Mahalanobis distance

Mahalanobis distance consider the data distribution into a distance calculation process.