Exercise 3. ***Hierarchical clustering***

Use single and complete link agglomerative clustering to group the data described by the following distance matrix. Show the dendrograms.

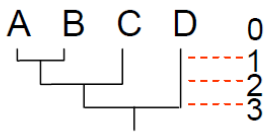
	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Solution:

Agglomerative ➔ initially every point is a cluster of its own and we merge cluster until we end-up with one unique cluster containing all points.

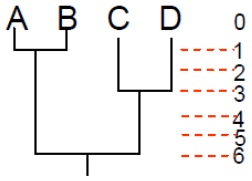
a) single link: distance between two clusters is the shortest distance between a pair of elements from the two clusters.

d	**k**	**K**	**Comments**
0	4	{A}, {B}, {C}, {D}	We start with each point = cluster
1	3	{A, B}, {C}, {D}	Merge {A} and {B} since A & B are the closest: d(A, B)=1
2	2	{A, B, C}, {D}	Merge {A, B} and {C} since B & C are the closest: d(B, C)=2
3	1	{A, B, C, D}	Merge D



b) complete link: distance between two clusters is the longest distance between a pair of elements from the two clusters.

d	k	K	Comments
0	4	{A}, {B}, {C}, {D}	We start with each point = cluster
1	3	{A, B}, {C}, {D}	d(A,B)=1<=1 ➔ merge {A} and {B}
2	3	{A, B}, {C}, {D}	d(A,C)=4>2 so we can't merge C with {A,B} d(A,D)=5>2 and d(B,D)=6>2 so we can't merge D with {A, B} d(C,D)=3>2 so we can't merge C and D
3	2	{A, B}, {C, D}	- d(A,C)=4>3 so we can't merge C with {A,B} - d(A,D)=5>3 and d(B,D)=6>3 so we can't merge D with {A, B} - d(C,D)=3 <=3 so merge C and D
4	2	{A, B}, {C, D}	{C,D} cannot be merged with {A, B} as d(A,D)= 5 >4 (and also d(B,D)= 6 >4) although d(A,C)= 4 <= 4, d(B,C)= 2<=4)
5	2	{A, B}, {C, D}	{C,D} cannot be merged with {A, B} as d(B,D)= 6 > 5
6	1	{A, B, C, D}	{C, D} can be merged with {A, B} since d(B,D)= 6 <= 6, d(A,D)= 5 <= 6, d(A,C)= 4 <= 6, d(B,C)= 2 <= 6



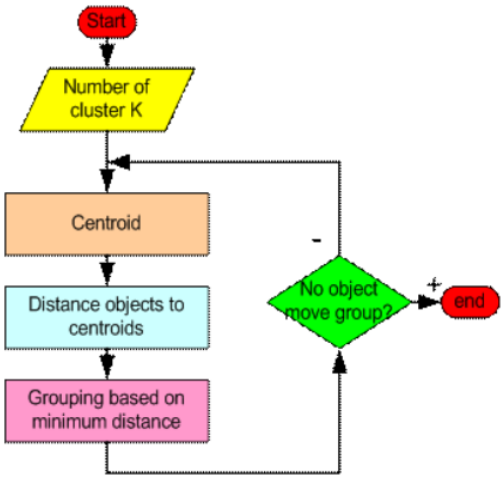
K Means Numerical Example

The basic step of k-means clustering is simple. In the beginning we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids.

Then the K means algorithm will do the three steps below until convergence

Iterate until *stable* (= no object move group):

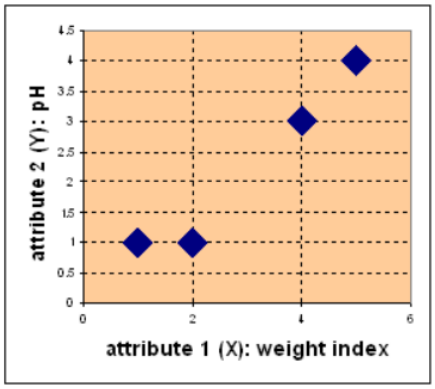
1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance



Suppose we have several objects (4 types of medicines) and each object have two attributes or features as shown in table below. Our goal is to group these objects into K=2 group of medicine based on the two features (pH and weight index).

Object	attribute 1 (X): weight index	attribute 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Each medicine represents one point with two attributes (X, Y) that we can represent it as coordinate in an attribute space as shown in the figure below.



Solution → https://people.revoledu.com/kardi/tutorial/kMean/NumericalExample.htm