MLG381: MACHINE LEARNING

K-Means Clustering

Lesson Objectives

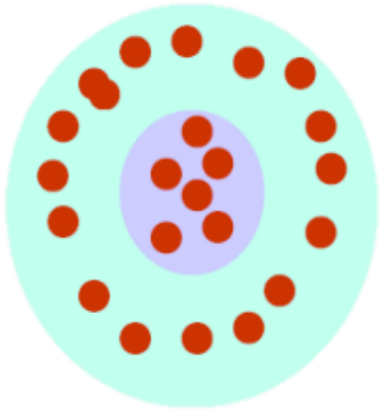
BELGIUM CAMPUS iTversity
It's the way we're wired

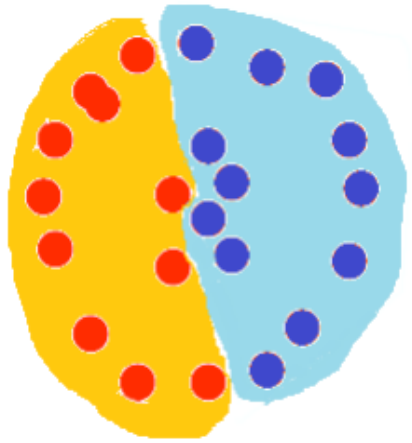
- History of Clustering
- Clustering Techniques
- K-Means Clustering
  - *Plot original data points*
  - *Calculating centroids*
  - *Computing Euclidean distances*
  - *Plot clustered points*
- Examples and Exercises

Introduction

BELGIUM CAMPUS iTversity
It's the way we're wired

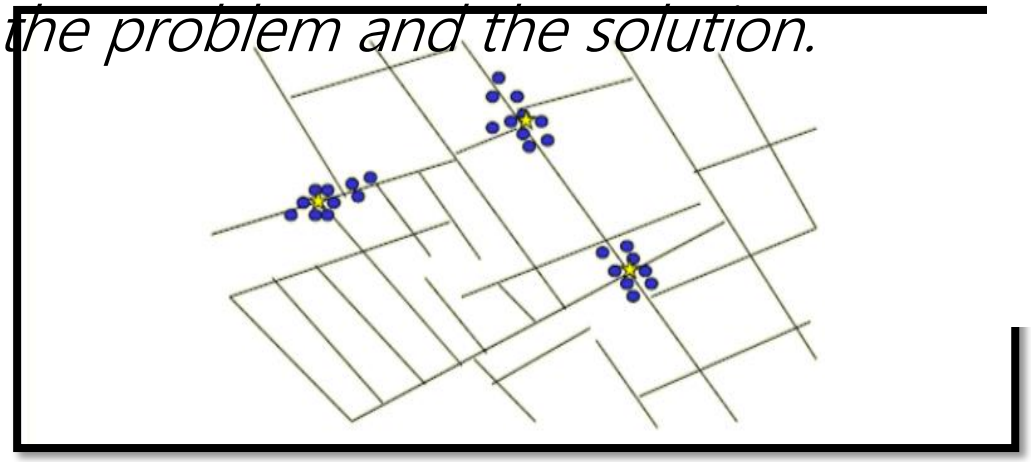
- The organization/segmentation of *unlabeled data* into similarity groups called *clusters*.
- A *cluster* is a collection of data items which are "similar" between them, and "dissimilar" to data items in other clusters.



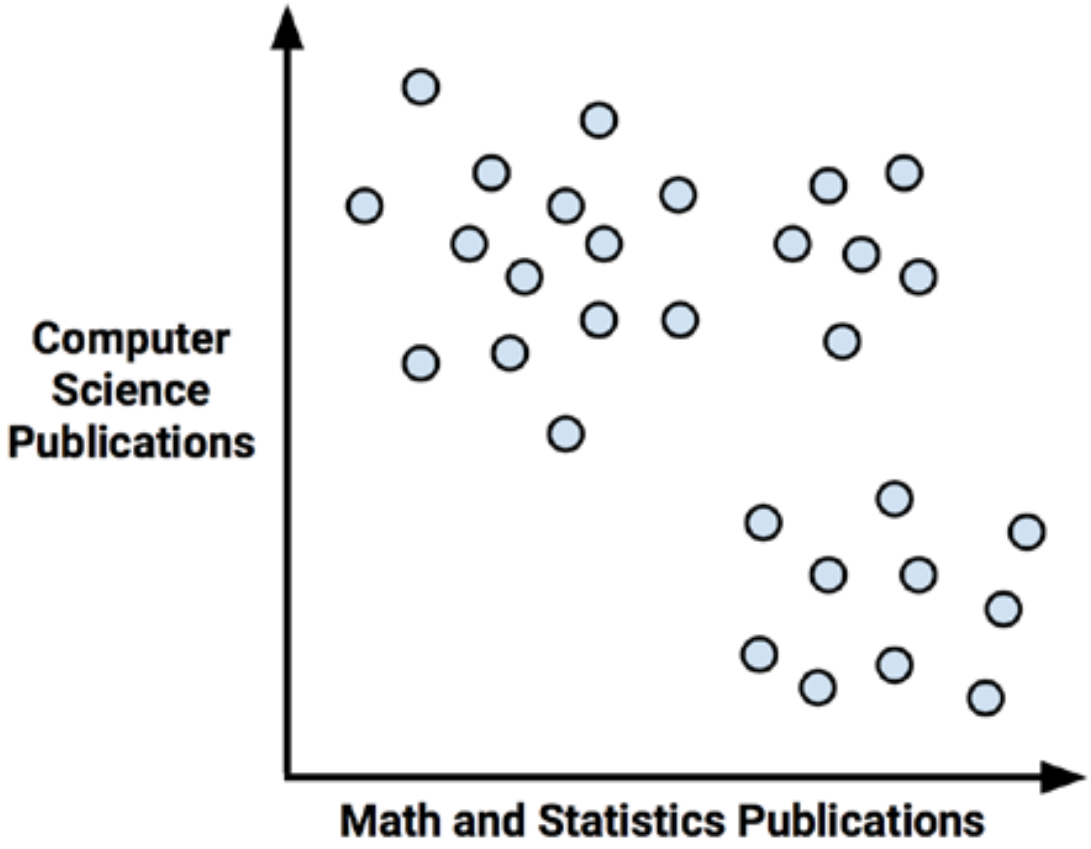


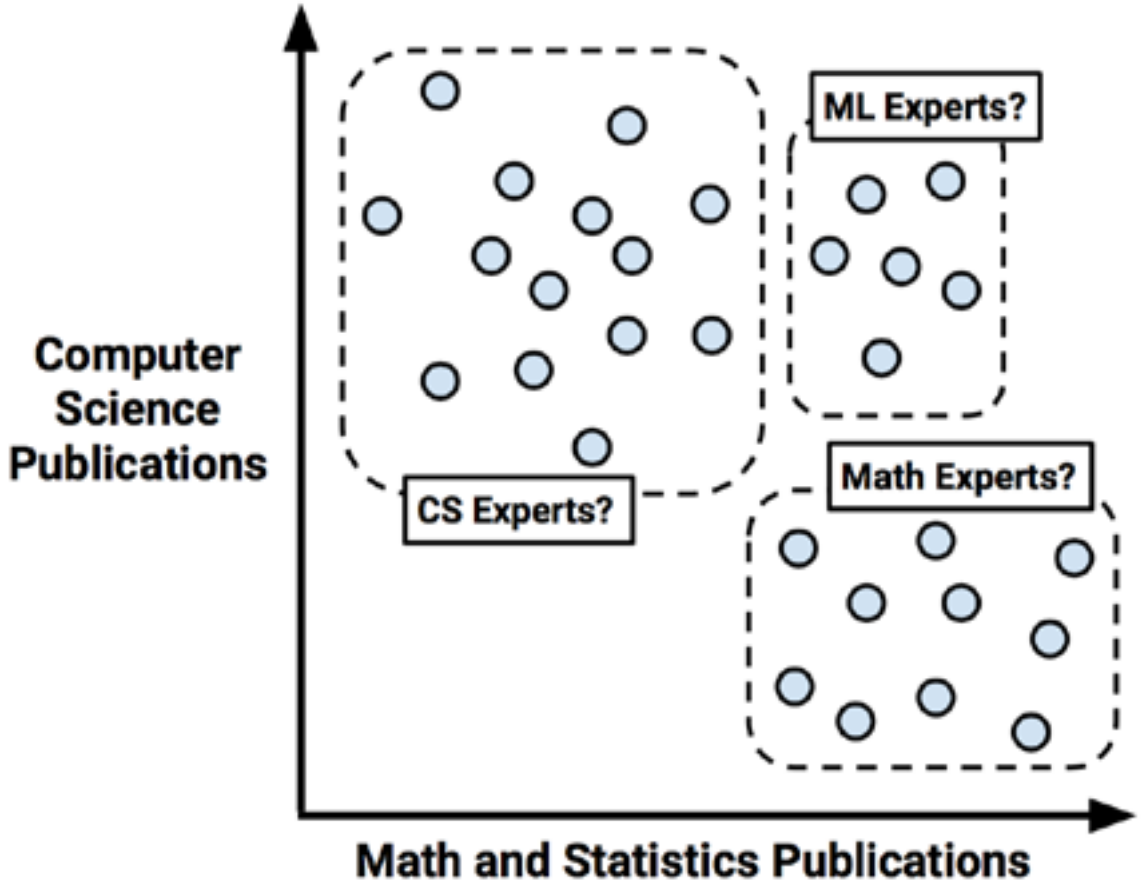
Historical Application of Clustering

- *John Snow*, a London physician plotted location of cholera deaths on a map during an outbreak in 1850s.
- Locations indicated that cases were clustered around certain around certain intersections where there were polluted polluted wells;
  - *thus exposing both the problem and the solution.*



Clustering /Segmentation





Clustering

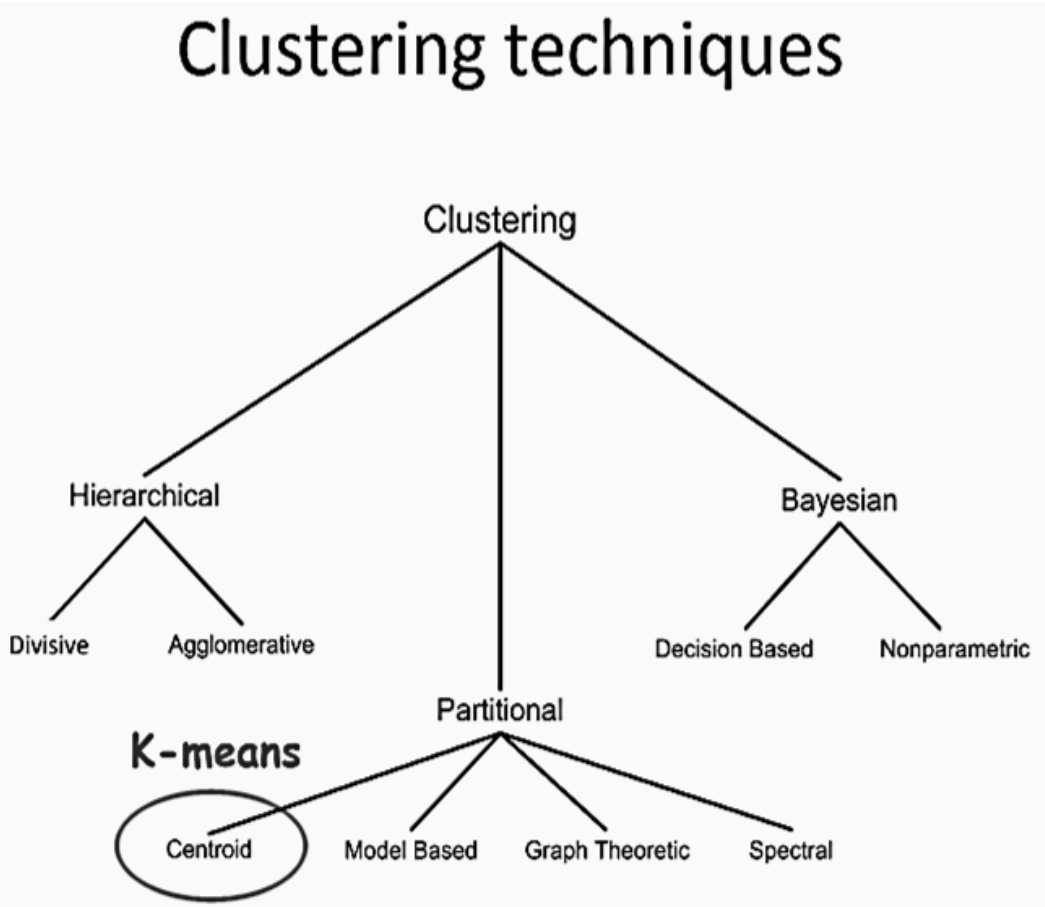
BELGIUM CAMPUS iTversity
It's the way we're wired

- Clustering is a form of machine learning where groups, or classes, are not known ahead of time so groupings are created by looking at similar and shared characteristics among the things being grouped. (homogeneous groups)
- Given a set of potential class members, the task of clustering is to establish the existence of two or more classes or clusters in the members.
- These clusters are often determined through trial and error by grouping things together by looking for similar features, analyzing the results to see how good they are, and repeating this learning process until the groups are deemed acceptable.
- Clustering is an example of unsupervised learning because classes are unknown at the start.

What do we need for Clustering?

BELGIUM CAMPUS iTversity
It's the way we're wired

- *Required:*
  1. Proximity measure.
  2. Criterion function to evaluate a clustering.
  3. Algorithm to compute clustering.



K-Means Clustering

BELGIUM CAMPUS iTversity
It's the way we're wired

- K-means *(MacQueen, 1967)* is a *partitional clustering algorithm.*
- Let the set of data points D be $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$, where $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ir})$ is a vector in $X \subseteq R^r$, and r is the number of dimensions.
- The algorithm partitions the given data into k clusters.
- Each cluster has a cluster centre, called *centroid*., at the beginning you make your own choice.
- k is specified by the user

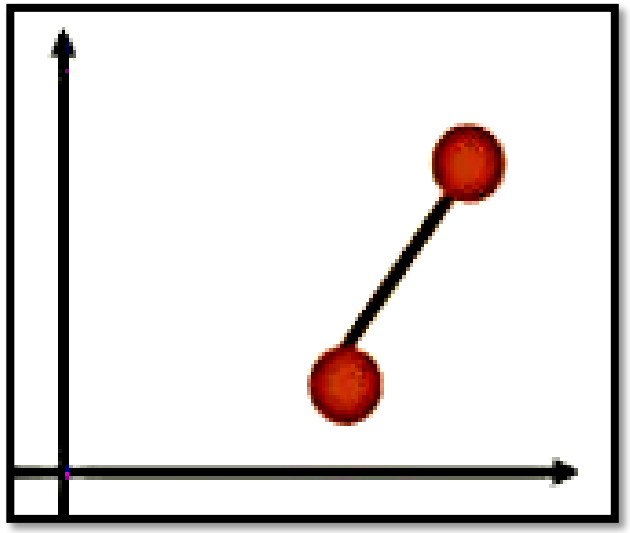
Determining Centroids

BELGIUM CAMPUS iTversity
It's the way we're wired

- Centroids are to be determined using *Euclidean distance* formula:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} \left(x_i^{(k)} - x_j^{(k)} \right)^2}$$

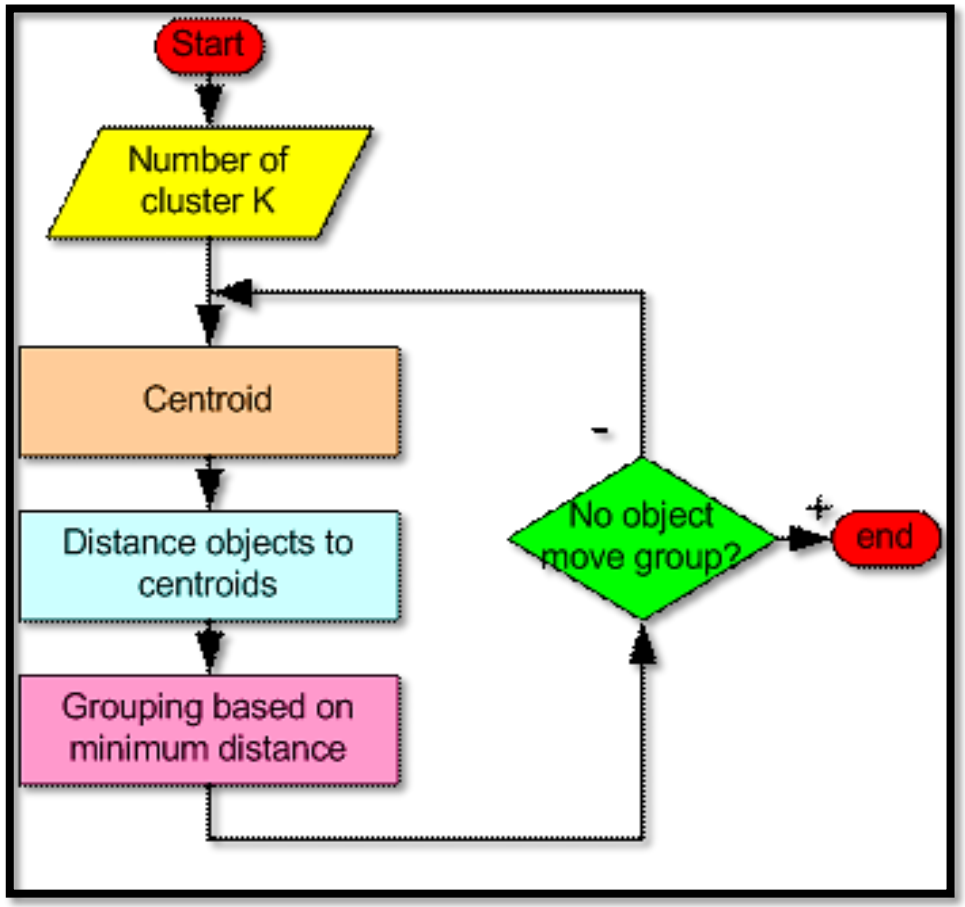
- where k represents number of clusters.
- $d(x_i; y_j) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$



K-means Algorithm

BELGIUM CAMPUS iTversity
It's the way we're wired

- We can take any random objects as the initial centroids.
- ***Given k, the k-means algorithm works as follows:***
  1. Determine the centroid coordinate.
  2. Determine the distance of each object to the centroids(D^n-matrix).
  3. Group the objects based on minimum distance(G^n-matrix).



Cost function for K-Means

BELGIUM CAMPUS iTversity
It's the way we're wired

- The goal is to minimize this cost function, which represents the overall within-cluster variation.
- During the iterative process of K-Means, centroids are updated to minimize this cost function until convergence is achieved, and the clustering solution stabilizes.

$$J(c, \mu) = \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

- $J(c, \mu)$ is the cost function.
- m is the total number of data points.
- $x^{(i)}$ represents the i^{th} data point.
- $\mu_{c^{(i)}}$ is the centroid to which $x^{(i)}$ is assigned.
- $||x^{(i)} - \mu_{c^{(i)}}||^2$ calculates the squared Euclidean distance between $x^{(i)}$ and its assigned centroid.

K-means Algorithm Example

BELGIUM CAMPUS iTversity
It's the way we're wired

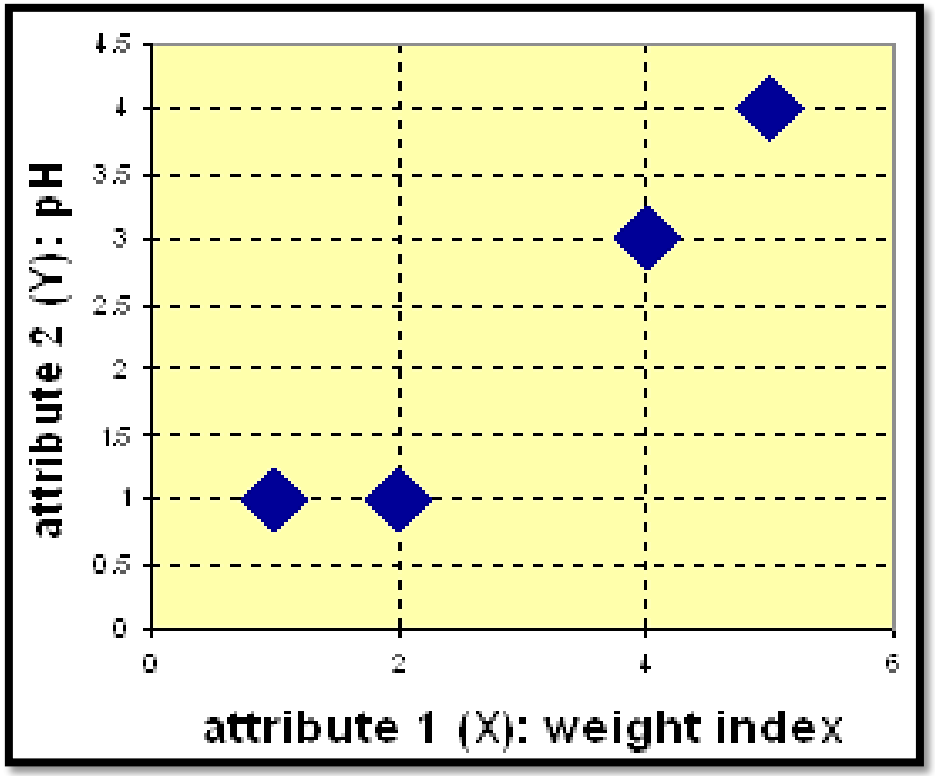
- Suppose we have 4 objects as your training data point and each object have 2 attributes.
- Each attribute represents coordinate of the object.
- Thus, we also know before hand that these objects belong to two groups of medicine (cluster 1 and cluster 2).
- **Q:** Determine which medicines belong to cluster 1 and which medicines belong to the other cluster?
- Each medicine represents one point with two components coordinate as shown below.

Object	Attribute 1: Weight index	Attribute 2: pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Solution

BELGIUM CAMPUS iTversity
It's the way we're wired

- Each medicine represents one point with two features (X, Y) that we can represent it as coordinate in a feature space as shown below;

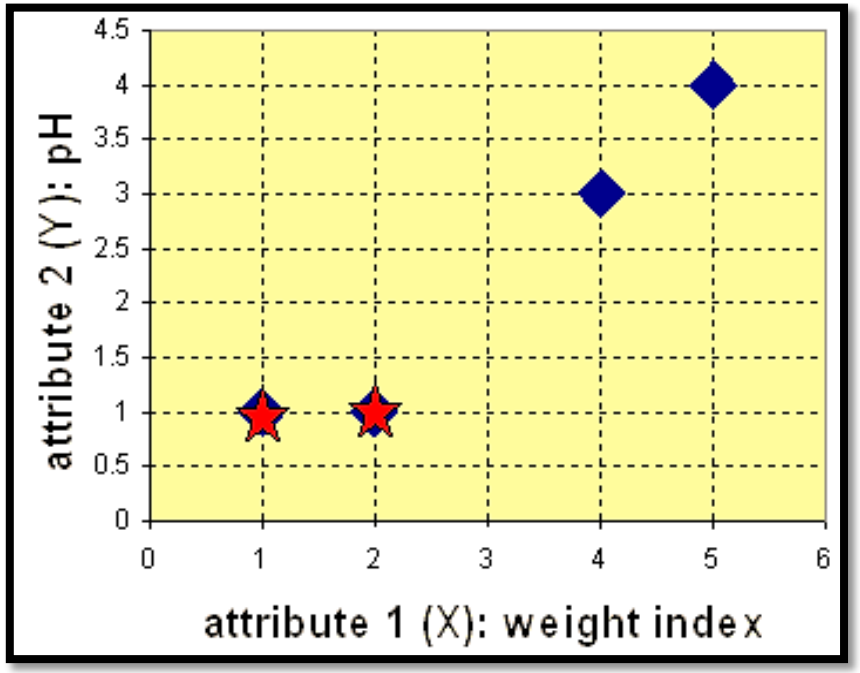


Iteration-0

BELGIUM CAMPUS iTversity
It's the way we're wired

1. Initial value of Centroids:

- Suppose we use medicine A and medicine B as the first centroids.
- Let c_1 and c_2 denote the coordinate of the centroids, then $c_1 = (1; 1)$ & $c_2 = (2; 1)$.



Iteration-0 (Continued)

BELGIUM CAMPUS iTversity
It's the way we're wired

2. <u>Objects-Centroids Distance:</u>

- Using ***Euclidean distance***, we have distance matrix;

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1;1)\ group - 1 \\ c_2 = (2;1)\ group - 2 \end{matrix}$$

$$\boldsymbol{A} \quad \boldsymbol{B} \quad \boldsymbol{C} \quad \boldsymbol{D}$$

- $\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{matrix}$,

$$d(C;c_1) = \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$
$$d(C;c_2) = \sqrt{(4-2)^2 + (3-1)^2} = 2.83$$

etc.

- Each column in the distance matrix symbolises the object.
- The first row of the distance matrix corresponds to the distance of each object to the first centroid and,

Iteration-0 (Continued)

BELGIUM CAMPUS iTversity
It's the way we're wired

3. <u>Objects Clustering:</u>

- We assign each object based on the minimum distance.
- The element of Group matrix below is 1 if and only if the object is assigned to that group.

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix}$$

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$A \quad B \quad C \quad D$$

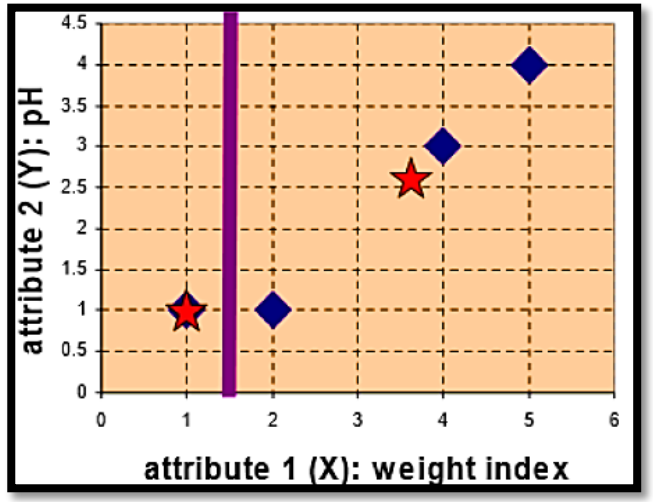
Iteration-1

BELGIUM CAMPUS iTversity
It's the way we're wired

4. <u>Determine New Centroids:</u>

- We compute the *<u>new centroid</u>* of each group based on these new memberships.
- Group 1 only has one member thus the centroid *remains unchanged*; $c_1 = (1; 1)$.
- Group 2 now has *three* members, thus the centroid is the average coordinate among the three members:

$$c_2 = \left(\frac{2+4+5}{3}; \frac{1+3+4}{3}\right)$$
$$= \left(\frac{11}{3}; \frac{8}{3}\right)$$



Iteration-1 (Continued)

BELGIUM CAMPUS iTversity
It's the way we're wired

5. *Objects-Centroids Distances:*

- The next step is to compute the distance of all objects to the *new centroids*.
- Similar to step 2, we have distance matrix:

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} c_1 = (1;1) \ \ group - 1 \\ c_2 = \left(\frac{11}{3};\frac{8}{3}\right) \ \ group - 2 \end{matrix}$$

$$A \quad B \quad C \quad D$$

6. *Objects Clustering:*

- We assign each object based on the minimum distance.
- Based on the new distance matrix, we move the *medicine B* to Group 1 while all the other objects remain.

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group - 1 \\ group - 2 \end{matrix}$$

$$A \quad B \quad C \quad D$$

Iteration-2

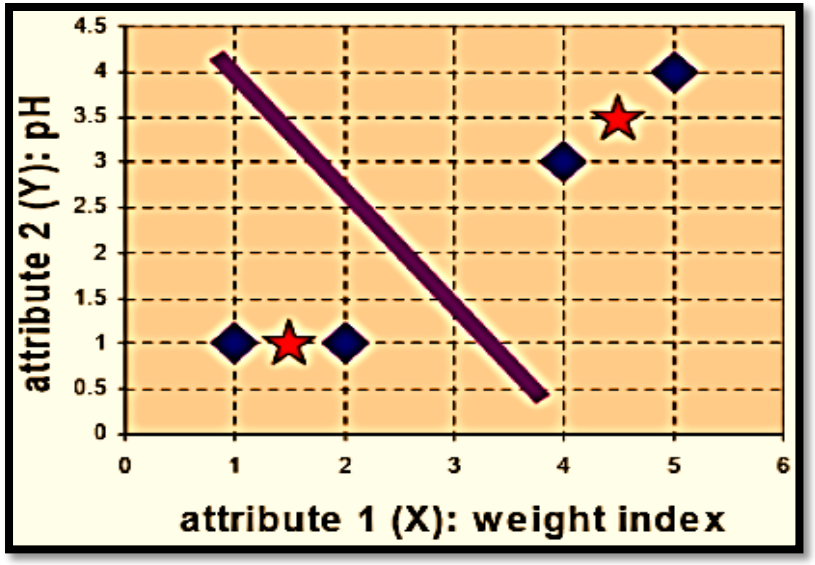
BELGIUM CAMPUS iTversity
It's the way we're wired

7. Determine Centroids:

- Now we repeat step 4 to calculate the new centroids coordinate based on the clustering of previous iteration.
- Group1 and group 2 both has two members, thus the new centroids are;

$$c_1 = \left(\frac{1+2}{2}; \frac{1+1}{2}\right) = \left(1\frac{1}{2}; 1\right) \text{ and,}$$

$$c_2 = \left(\frac{4+5}{2}; \frac{3+4}{2}\right) = \left(4\frac{1}{2}; 3\frac{1}{2}\right)$$



Iteration-2 (Continued)

BELGIUM CAMPUS iTversity
It's the way we're wired

8. *Objects-Centroids Distances:*

- Repeat step 2 again, we have new distance matrix at iteration 2 as

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix}$$

$$c_1 = \left(1\frac{1}{2}; 1\right) \quad group - 1$$

$$c_2 = \left(4\frac{1}{2}; 3\frac{1}{2}\right) \quad group - 2$$

$$A \quad B \quad C \quad D$$

9. *Objects Clustering:*

- We obtain result that $G^2 = G^1$.
- Thus, the computation of the *k-mean clustering* has reached its stability(has Converged) and no more iteration is needed.

$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group - 1 \\ group - 2 \end{matrix}$$

$$A \quad B \quad C \quad D$$

Final Result:

BELGIUM CAMPUS iTversity
It's the way we're wired

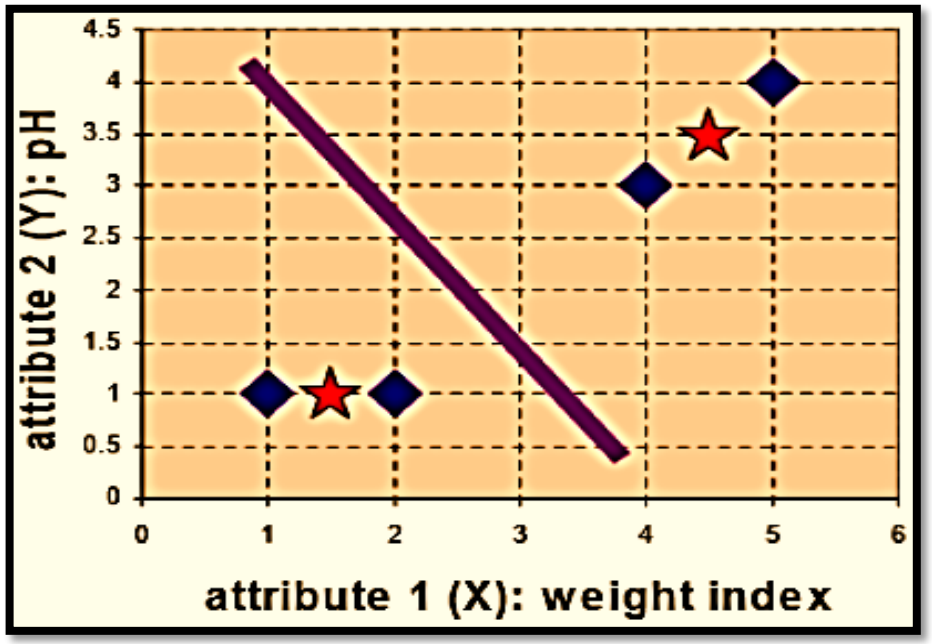
- We get the final grouping as the results;
  - Group 1: Medicine A and B
  - Group 2: Medicine C and D.

Object	Feature 1: Weight index	Feature 2: pH	Group (Results)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Final Solution

BELGIUM CAMPUS iTversity
It's the way we're wired

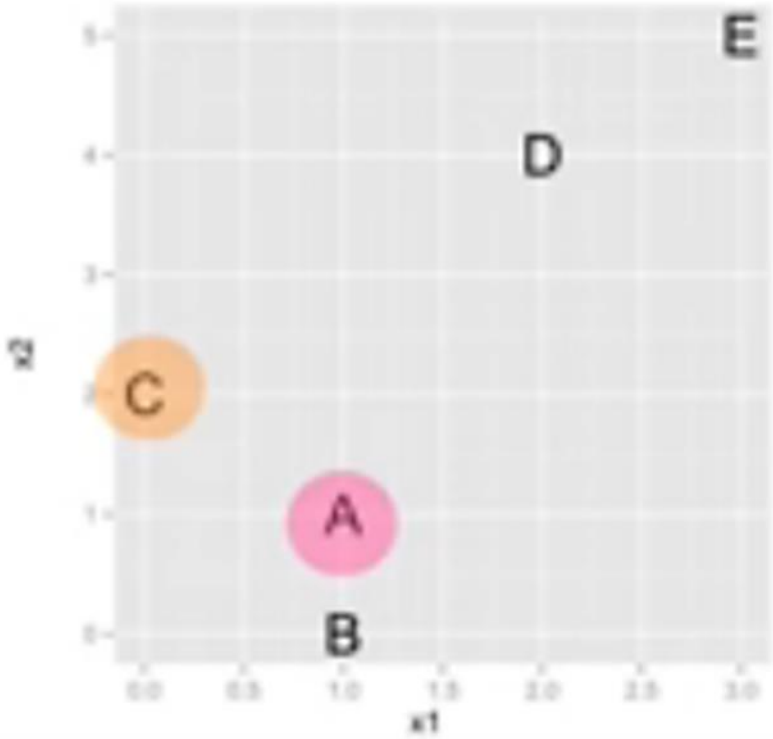
- Group1{A; B} c1(1.5; 1)
- Group2{C; D} c2(4.5; 3.5)



Exercise 1

BELGIUM CAMPUS iTversity
It's the way we're wired

	i	X_1	X_2
$\bar{X}_1^0$	A	1	1
	B	1	0
$\bar{X}_2^0$	C	0	2
	D	2	4
	E	3	5



Cluster Analysis Exercise

BELGIUM CAMPUS iTversity
It's the way we're wired

- Cluster analysis embraces a variety of techniques, the main objective of which is to group observations or variables into homogeneous and distinct clusters.
- The daily expenditures on food (X1) and clothing (X2) of five persons is given below. Use the k means algorithm to cluster them.

Person	X_1	X_2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1

(k-means method)

BELGIUM CAMPUS iTversity
It's the way we're wired

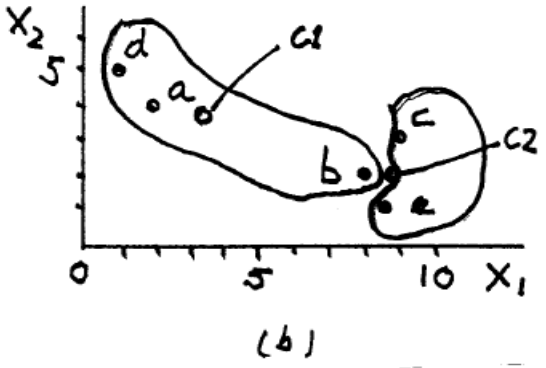
Person	X_1	X_2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1

Suppose two clusters are to be formed
First assign a, b and d to Cluster 1,
Next assign c and e to Cluster 2.
Calculate cluster centroids

Cluster 1			Cluster 2		
Obs.	X_1	X_2	Obs.	X_1	X_2
a	2	4	c	9	3
b	8	2	e	8.5	1
d	1	5			
Ave.	3.67	3.67	Ave.	8.75	2

(a)

Centroid of Cluster 1 is the point (X1 = 3:67, X2 = 3:67),



(b)

The cluster centroid is the point with coordinates equal to the average values of the variables for the observations in that cluster

calculate the distance between **a** and the two centroids:

$$D(a, abd) = \sqrt{(2-3.67)^2 + (4-3.67)^2} = 1.702,$$
$$D(a, ce) = \sqrt{(2-8.75)^2 + (4-2)^2} = 7.040.$$

a is closer to the centroid of Cluster 1, to which it is currently assigned. a is not reassigned

calculate the distance between b and the two cluster centroids:

$$D(b, abd) = \sqrt{(8-3.67)^2 + (2-3.67)^2} = 4.641$$
$$D(b, ce) = \sqrt{(8-8.75)^2 + (2-2)^2} = 0.750.$$

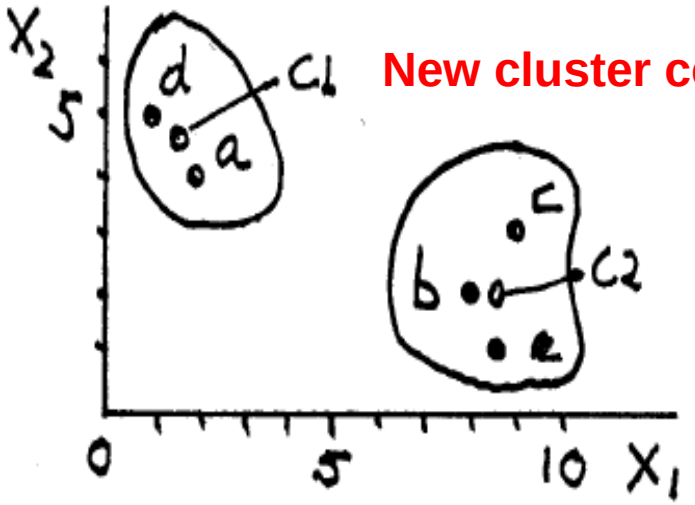
Since b is closer to Cluster 2's centroid than to that of Cluster 1, it is reassigned to Cluster 2.

(k-means method)

BELGIUM CAMPUS iTversity
It's the way we're wired

Cluster 1			Cluster 2		
Obs.	X_1	X_2	Obs.	X_1	X_2
a	2	4	c	9	3
d	1	5	e	8.5	1
			b	8	2
Ave.	1.5	4.5	Ave.	8.5	2

(a)



(b)

New cluster centroids

Obs.	Distance from Cluster 1	Distance from Cluster 2
a	0.707*	6.801
b	6.964	0.500*
c	7.649	1.118*
d	0.707*	8.078
e	7.826	1.000*

Every observation belongs to the cluster to the centroid of which it is nearest, and the k-means method stops.

Exercises

BELGIUM CAMPUS iTversity
It's the way we're wired

Q1: Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: $A_1 = (2; 10)$, $A_2 = (2; 5)$, $A_3 = (8; 4)$, $A_4 = (5; 8)$,$A_5 = (7; 5)$, $A_6 = (6; 4)$, $A_7 = (1; 2)$, $A_8 = (4; 9)$. Suppose that the initial seeds (centres of each cluster) are A_1 , A_4 and A_7 . Run the k −means algorithm for 1 epoch only. At the end of this epoch show:

a. The new clusters (i.e. the examples belonging to each cluster).
b. The centres of the new cluster.
c. Draw a 10 by 10 feature space with all the 8 points and show the clusters after the first epoch and the new centroids.
d. How many more iterations are needed to converge? Draw the result for each epoch.

Q2: Use the K-means algorithm with **K=3** to cluster the same data set that is copied below for your convenience.

10; 20; 40; 80; 85; 121; 160; 168; 195

Suppose that the points 160, 168, and 195 were selected as the initial cluster means. Work from these initial values *to determine the final clustering for the data*. Show your work so that it will be easy to see each step you took to get from the initial values to your final clustering.

Exs. No. 3

BELGIUM CAMPUS iTversity
It's the way we're wired

	X	Y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

THANK YOU

BELGIUM CAMPUS
iTversity
It's the way we're wired

info@belgiumcampus.ac.za +27 10 593 53 68

/belgiumcampusSA

#Belgium Campus

/belgiumcampus