# Data Mining

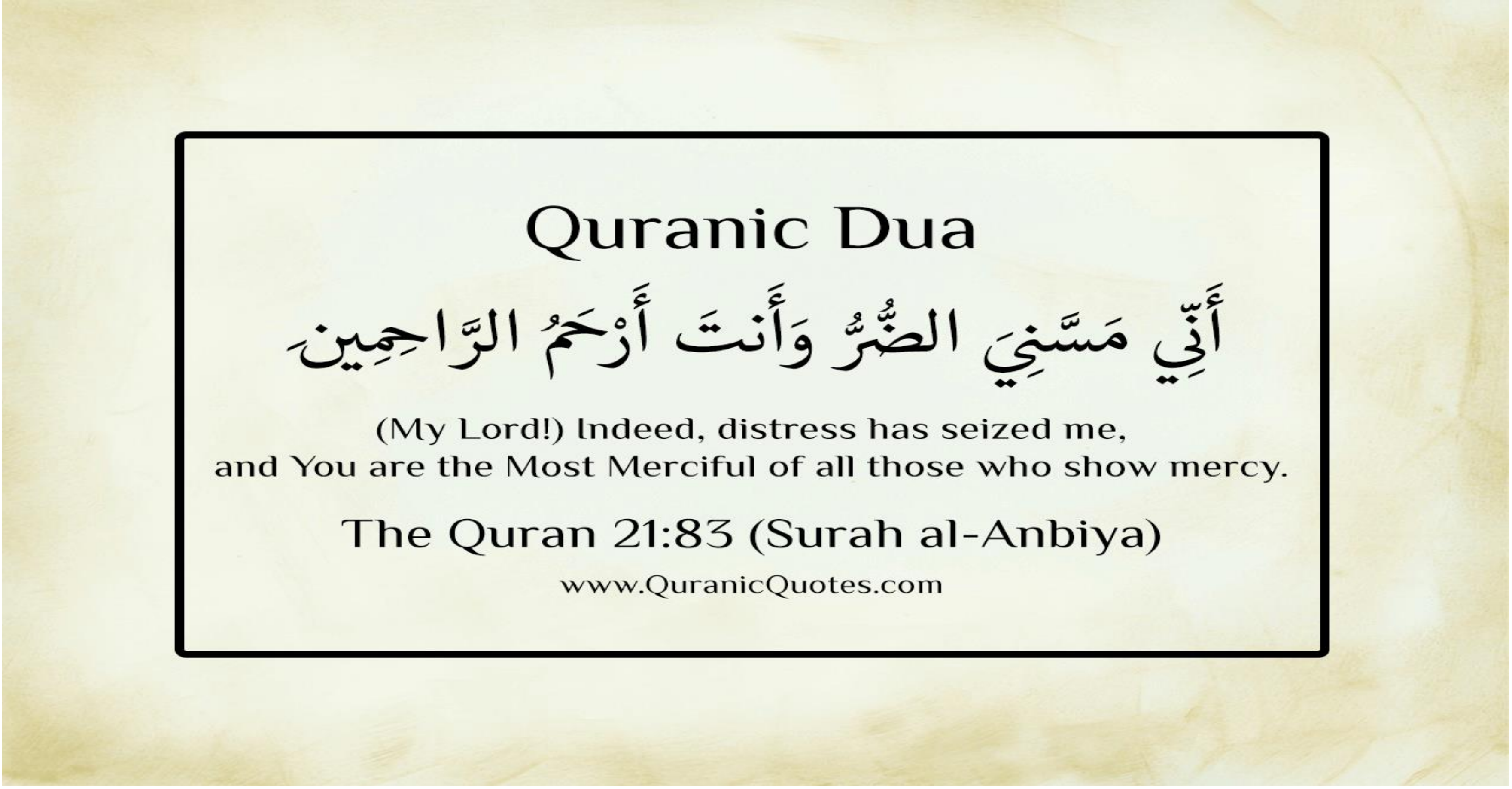
## CLUSTERING
## K-MEANS ALGORITHM

Prof. Dr. Hikmat Ullah Khan
Department of Information Technology

UNIVERSITY OF SARGODHA, SARGODHA

# Lesson from Holy Quran





# What is Un-Supervised Learning?



- Categorization of objects into different groups
- No concept of Class Labels
- The partitioning of a data set into subsets (clusters),
- Data in each subset share some common trait
  - according to some defined distance measure.
- Aim:
  - Maximum Intra Cluster Similarity
  - Minimum Inter Cluster Similarity (Maximum Inter Cluster Dis-similarity)

# Types of clustering



1. **Hierarchical algorithms**: these find successive clusters using previously established clusters.

   1. **Agglomerative ("bottom-up"):**

   Begin with each element as a separate cluster and merge them into successively larger clusters.

   2. **Divisive ("top-down"):**

   Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

2. **Partitional clustering:**

   Determine all clusters at once.

   - ***K*-means and derivatives**

# Common Distance measures



- Determine how the *similarity* of two elements is calculated. They include:

1. The Euclidean distance :
2. The Manhattan distance :

$$d(x,y) = \sqrt[2]{\sum_{i=1}^{p} |x_i - y_i|^2}$$

$$d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$

# Common Distance measures



3.The maximum norm is given by:

$$d(x, y) = \max_{1 \le i \le p} | x_i - y_i |$$

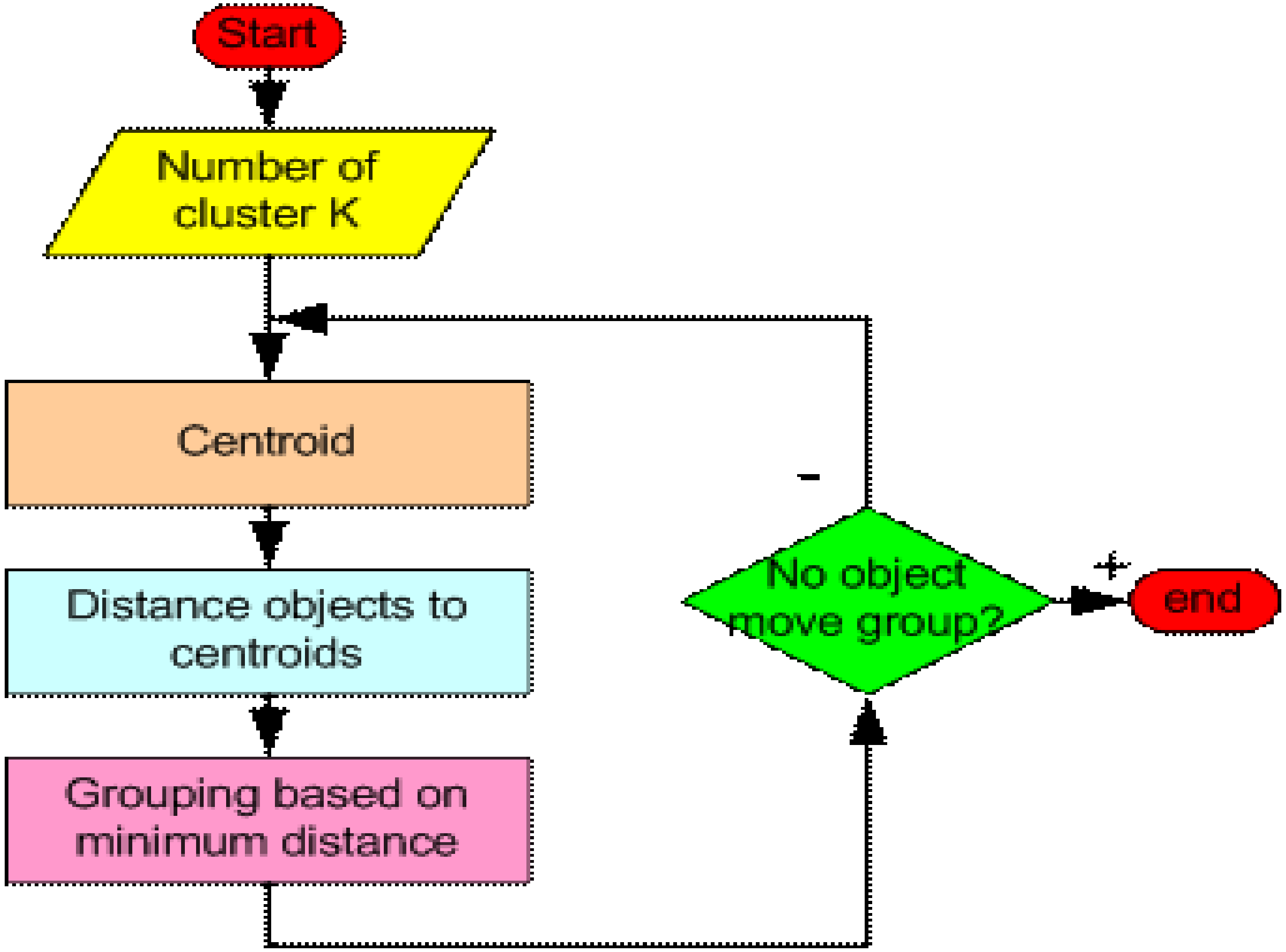
# K-MEANS CLUSTERING



- The **k-means algorithm** is an algorithm to cluster $n$ objects based on attributes into $k$ partitions, where $k < n$.

# How the K-Mean Clustering algorithm works?





# Steps



- **Step 1:** Begin with a decision on the value of k = number of clusters .
- **Step 2**: Take any k elements as Centre element for k cluster
- **Step 3:**
  - Compute the distance of each element from the centroid of each of the clusters.
  - Find the new Centroid by averaging the element values.
- **Step 4 .** Repeat step 3 until convergence is achieved,

# A Simple example showing the implementation of k-means algorithm (using K=2)



| Individual | Variable 1 | Variable 2 |
|---|---|---|
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

- **Step 1**:
- Randomly choose two centroids (k=2) for two clusters.
- For instance, m1=(1.0,1.0) and m2=(5.0,7.0).

| Individual | Variable 1 | Variable 2 |
|---|---|---|
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

| | Individual | Mean Vector |
|---|---|---|
| Group 1 | 1 | (1.0, 1.0) |
| Group 2 | 4 | (5.0, 7.0) |

## NOTE the Data Points

1. Calculate The Euclidean Distance of each Element from Each Centroid Point:
2. Prepare a table showing distances of Each element from Each Centroid.
3. For instance, The distance of 2nd element from both centroids are calculated as follows:

| Value | Distance from Centroid 1 | Distance from Centroid 2 |
| --- | --- | --- |
| | | |

$$d(m_1,2)=\sqrt{|1.0-1.5|^2+|1.0-2.0|^2}=1.12$$

$$d(m_2,2)=\sqrt{|5.0-1.5|^2+|7.0-2.0|^2}=6.10$$

- Have you got the same result as follows:

| Individual | Centroid 1 | Centroid 2 |
|---|---|---|
| 1 | 0 | 7.21 |
| 2 (1.5, 2.0) | 1.12 | 6.10 |
| 3 | 3.61 | 3.61 |
| 4 | 7.21 | 0 |
| 5 | 4.72 | 2.5 |
| 6 | 5.31 | 2.06 |
| 7 | 4.30 | 2.92 |

**Step 2:**

- Thus, we obtain two clusters containing:

  {1,2,3} and {4,5,6,7}.

- New centroids are:

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12, 5.38)$$

| Individual | Centroid 1 | Centroid 2 |
|---|---|---|
| 1 | 0 | 7.21 |
| 2 (1.5, 2.0) | 1.12 | 6.10 |
| 3 | 3.61 | 3.61 |
| 4 | 7.21 | 0 |
| 5 | 4.72 | 2.5 |
| 6 | 5.31 | 2.06 |
| 7 | 4.30 | 2.92 |

**Step 3:**

- Now using these centroids, we compute the Euclidean distance of each object, as shown in table.

- Therefore, the new clusters are: {1,2} and {**3**,4,5,6,7}
- New Centroid?
- Next centroids are: m1=(1.25,1.5) and m2 = (3.9,5.1)

| Individual | Centroid 1 | Centroid 2 |
|---|---|---|
| 1 | 1.57 | 5.38 |
| 2 | 0.47 | 4.28 |
| 3 | 2.04 | 1.78 |
| 4 | 5.64 | 1.84 |
| 5 | 3.15 | 0.73 |
| 6 | 3.78 | 0.54 |
| 7 | 2.74 | 1.08 |

- Step 4 :

  The clusters obtained are:

  {1,2} and {3,4,5,6,7}

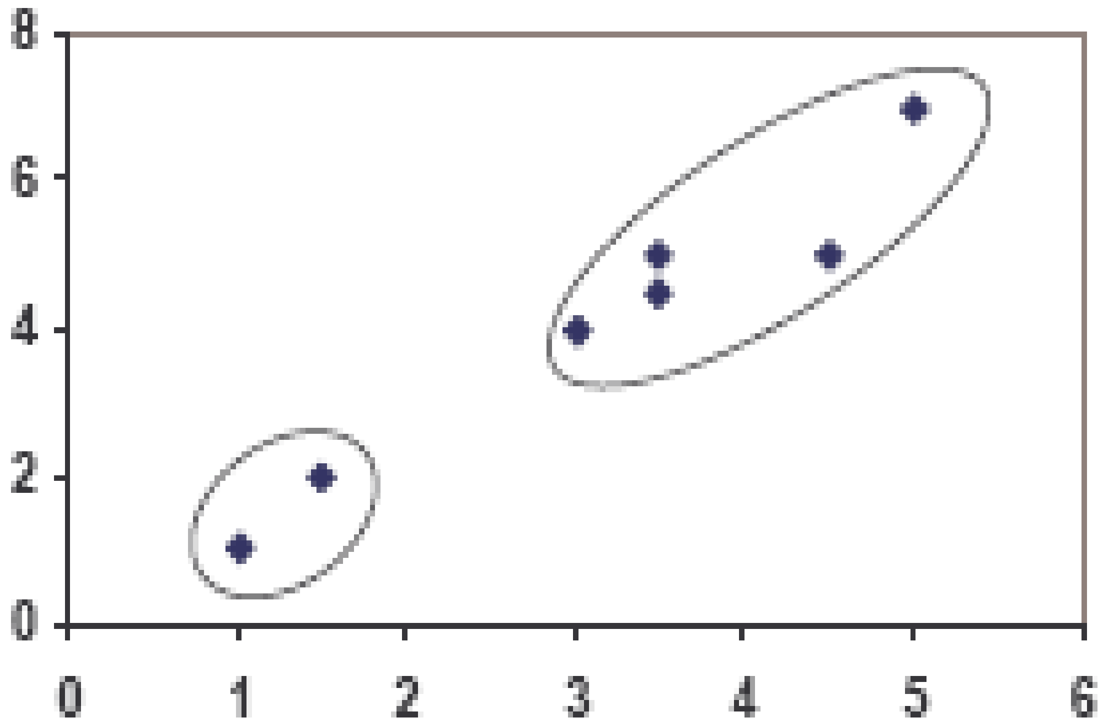
- Therefore, there is no change in the cluster.
- Thus, the algorithm comes to a halt here and final result consist of 2 clusters {1,2} and {3,4,5,6,7}.

| Individual | Centroid 1 | Centroid 2 |
|---|---|---|
| 1 | 0.56 | 5.02 |
| 2 | 0. 56 | 3.92 |
| 3 | 3.05 | 1.42 |
| 4 | 6.66 | 2.20 |
| 5 | 4.16 | 0.41 |
| 6 | 4.78 | 0.61 |
| 7 | 3.75 | 0.72 |

# PLOT



- (PLOT for each Iteration will help you how K-means works)



# Value of K



- Now us consider the same example values
- Take k-3
  - We are interested in finding three clusters
- Re-apply K-Means Algorithms again

# (with K=3)



- For instance, Take FIRST THREE ELEMENTS AS THREE CENTROIDS

| Individual | Variable 1 | Variable 2 |
|---|---|---|
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

**Step 1**

# (with K=3)



| Individual | $m_1 = 1$ | $m_2 = 2$ | $m_3 = 3$ | cluster |
|---|---|---|---|---|
| 1 | 0 | 1.11 | 3.61 | 1 |
| 2 | 1.12 | 0 | 2.5 | 2 |
| 3 | 3.61 | 2.5 | 0 | 3 |
| 4 | 7.21 | 6.10 | 3.61 | 3 |
| 5 | 4.72 | 3.61 | 1.12 | 3 |
| 6 | 5.31 | 4.24 | 1.80 | 3 |
| 7 | 4.30 | 3.20 | 0.71 | 3 |

$C_3$

clustering with initial centroids (1, 2, 3)

**Step 1**

# (with K=3)



| Individual | $m_1 = 1$ | $m_2 = 2$ | $m_3 = 3$ | cluster |
|---|---|---|---|---|
| 1 | 0 | 1.11 | 3.61 | 1 |
| 2 | 1.12 | 0 | 2.5 | 2 |
| 3 | 3.61 | 2.5 | 0 | 3 |
| 4 | 7.21 | 6.10 | 3.61 | 3 |
| 5 | 4.72 | 3.61 | 1.12 | 3 |
| 6 | 5.31 | 4.24 | 1.80 | 3 |
| 7 | 4.30 | 3.20 | 0.71 | 3 |

$C_3$

clustering with initial centroids (1, 2, 3)

**Step 1**

| Individual | $m_1$ (1.0, 1.0) | $m_2$ (1.5, 2.0) | $m_3$ (3.9,5.1) | cluster |
|---|---|---|---|---|
| 1 | 0 | 1.11 | 5.02 | 1 |
| 2 | 1.12 | 0 | 3.92 | 2 |
| 3 | 3.61 | 2.5 | 1.42 | 3 |
| 4 | 7.21 | 6.10 | 2.20 | 3 |
| 5 | 4.72 | 3.61 | 0.41 | 3 |
| 6 | 5.31 | 4.24 | 0.61 | 3 |
| 7 | 4.30 | 3.20 | 0.72 | 3 |

**Step 2**

# Real-Life Numerical Example



- We have 4 medicines as our training data points object and each medicine has 2 attributes. Each attribute represents coordinate of the object. We have to determine which medicines belong to cluster 1 and which medicines belong to the other cluster.

# Real-Life Numerical Example

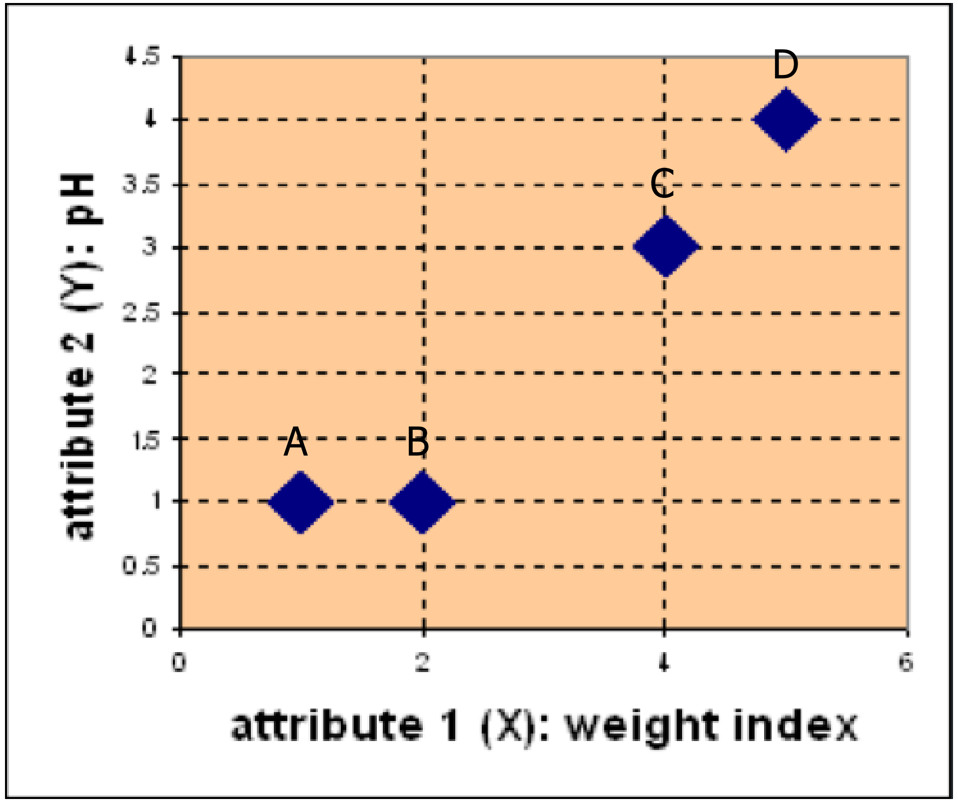


| Medicine | Weight | pH-Index |
|---|---|---|
| A | 1 | 1 |
| B | 2 | 1 |
| C | 4 | 3 |
| D | 5 | 4 |

# Example



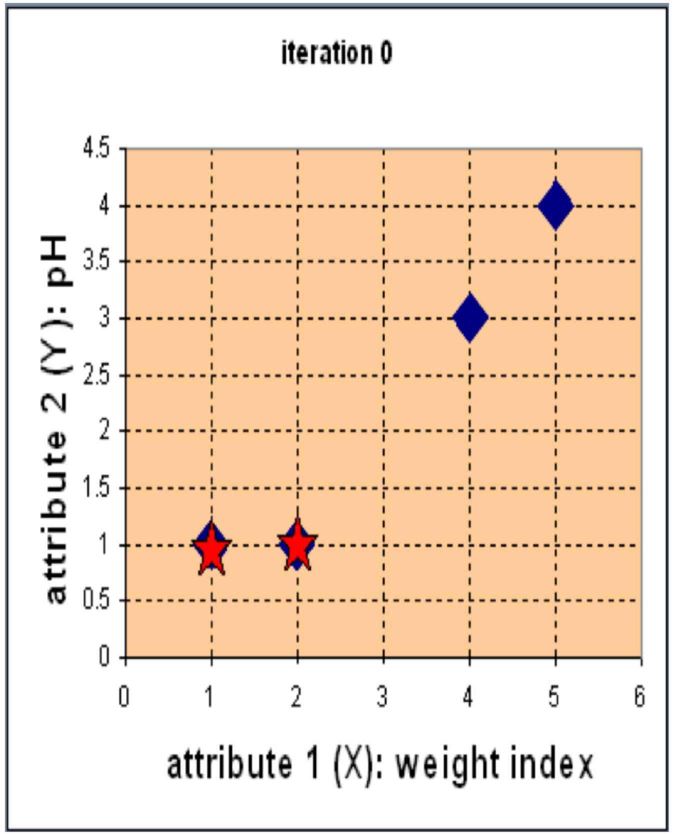
| Medicine | Weight | pH-Index |
|---|---|---|
| A | 1 | 1 |
| B | 2 | 1 |
| C | 4 | 3 |
| D | 5 | 4 |



# Example



- **Step 1: Use initial seed points for partitioning**




$$c_1 = A, c_2 = B$$

$$\mathbf{D}^0 = \begin{array}{c} \begin{bmatrix} 0 & 1 & 3.61 & \boxed{5} \\ 1 & 0 & 2.83 & \boxed{4.24} \end{bmatrix} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \quad \begin{array}{ll} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{array}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{array}{c} X \\ Y \end{array}$$

$$d(D,c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

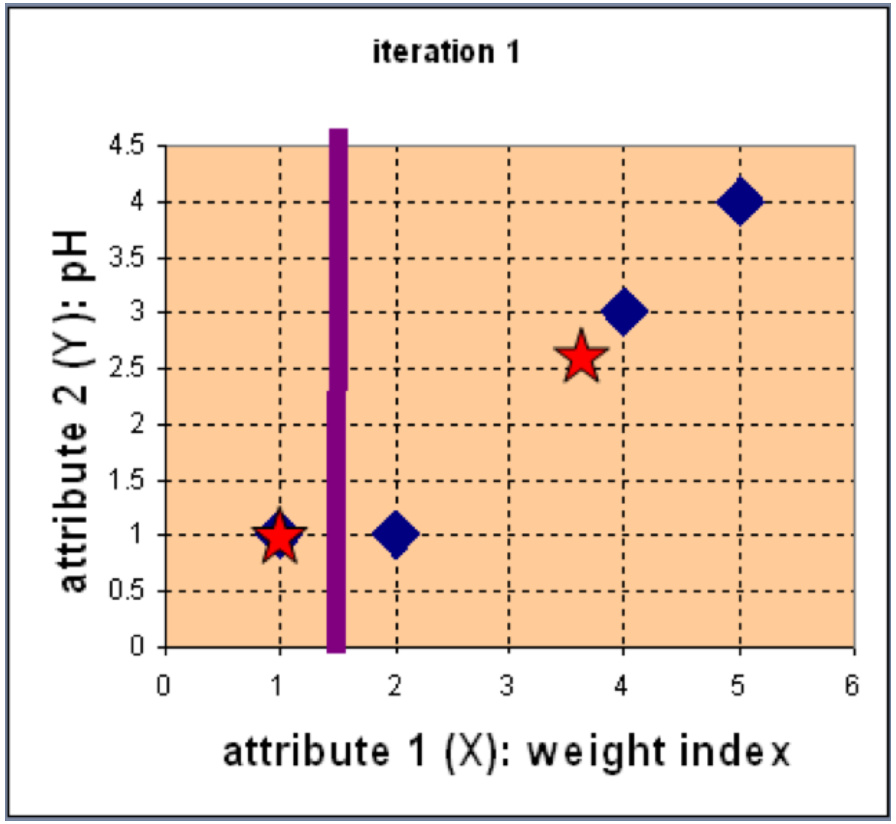
$$d(D,c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

Assign each object to the cluster with the nearest seed point

# Example



- **Step 2: Compute new centroids of the current partition**




Knowing the members of each cluster, now we compute the new
centroid of each group based on
these new memberships.

$$c_1 = (1,\ 1)$$

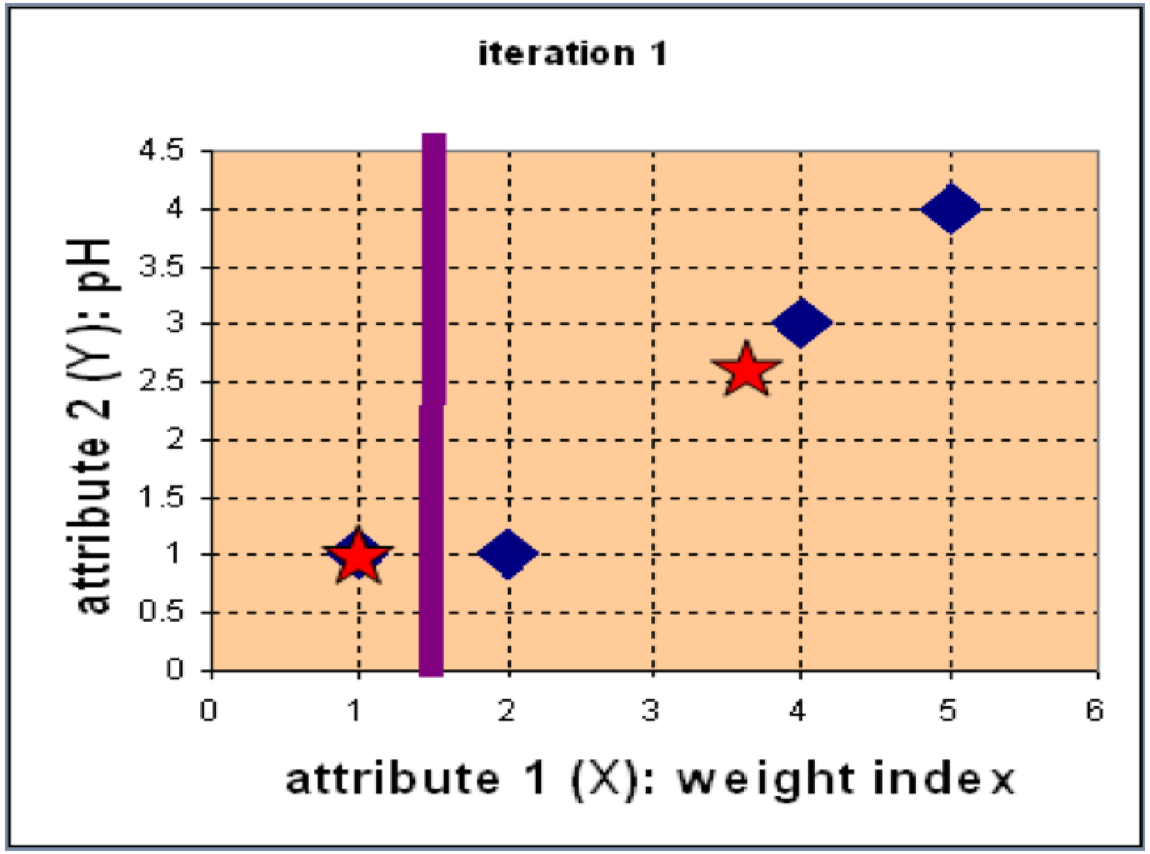
$$c_2 = \left( \frac{2+4+5}{3},\ \frac{1+3+4}{3} \right)$$

$$= (\frac{11}{3},\ \frac{8}{3})$$

# Example



- **Step 2: Renew membership based on new centroids**




Compute the distance of all objects to the new centroids

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \begin{array}{ll} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{array}$$

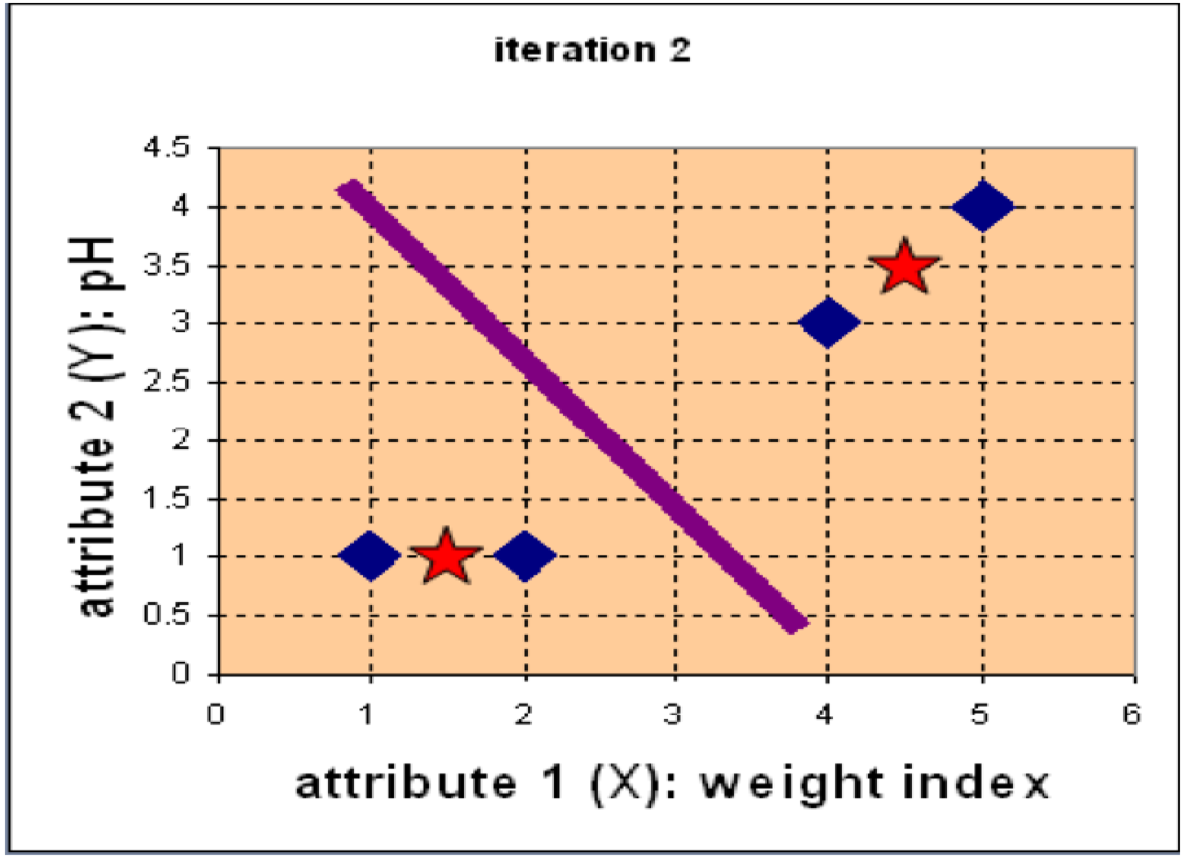
$$\begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \end{array} \begin{array}{l} \\ X \\ Y \end{array}$$

Assign the membership to objects

# Example



- **Step 3: Repeat the first two steps until its convergence**




Knowing the members of each cluster, now we compute the new
centroid of each group based on
these new memberships.

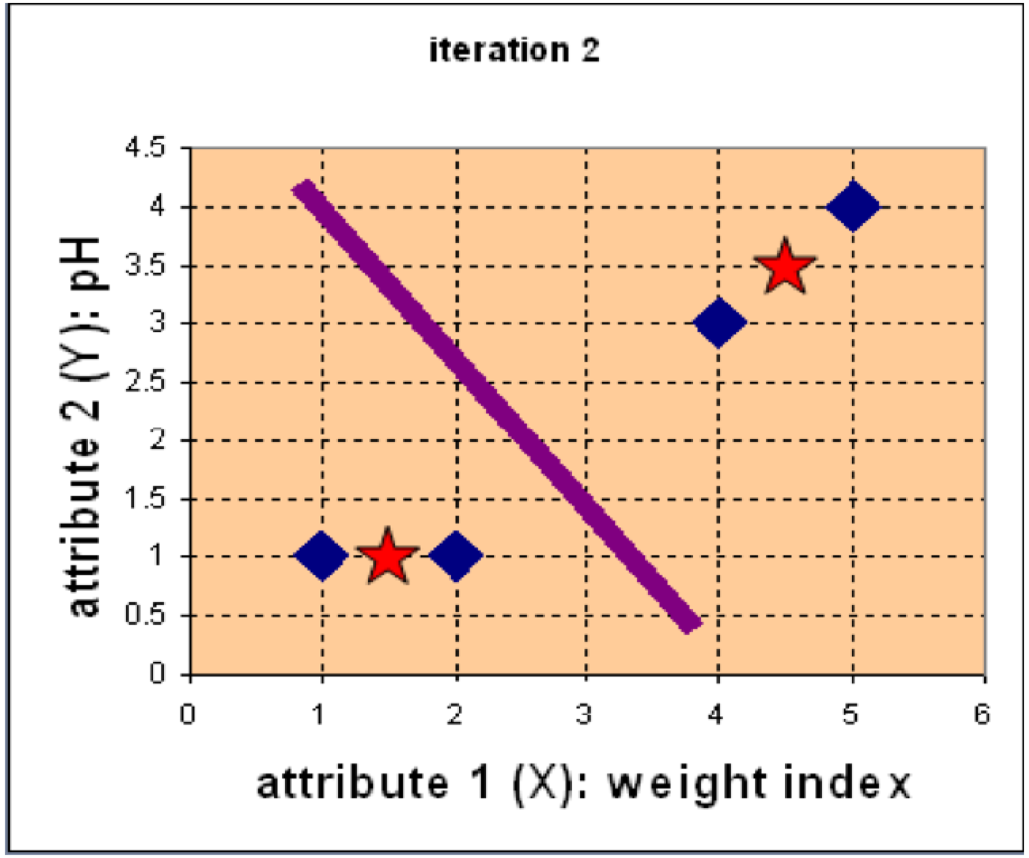
$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1\frac{1}{2}, 1)$$

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

# Example



- **Step 3: Repeat the first two steps until its convergence**




Compute the distance of all objects to the new centroids

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{array}{l} \mathbf{c}_1 = (1\tfrac{1}{2}, 1) \quad group-1 \\ \mathbf{c}_2 = (4\tfrac{1}{2}, 3\tfrac{1}{2}) \quad group-2 \end{array}$$

$$\begin{array}{cccc} A & B & C & D \end{array}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{array}{l} X \\ Y \end{array}$$

Stop due to no new assignment
Membership in each cluster no longer change

# Example



- We get the final grouping as the results as:

| Object | Feature1(X): weight index | Feature2 (Y): pH | Group (result) |
|---|---|---|---|
| Medicine A | 1 | 1 | 1 |
| Medicine B | 2 | 1 | 1 |
| Medicine C | 4 | 3 | 2 |
| Medicine D | 5 | 4 | 2 |

# Relevant Issues



- Other problems
  - Need to specify $K$, the *number* of clusters, in advance
  - Unable to handle noisy data and outliers
  - Applicable only when mean is defined, then what about categorical data
    - NOT Applicable for Categorical Data

# Advantages and Applications of K-Mean Clustering



- Advantages:
  - It is *efficient algorithm (Fast Computation).*
  - Simple , easy to implement and understand and apply
  - *Widely used in Machine Learning, Data mining, etc*
- *Applications*
  - Speech Recognition and Signal Processing
  - Data compression
  - Noise reduction
  - Digital Image Segmentation
  - *Web Studies for grouping of People based on Similar Characteristics (Homophily)*

# Exercise



*1. Take any Ten Elements, each of two attributes*

- *Take different values of K*
  - *(for instance, k=2, k=3, and k=4)*
- *Apply K-Means Algorithm*
- *Take Different Starting Centroids for different values of k*

- *2. Try to take sample/downloaded data and run K-means using Weka, other tools (discussed in class) and/OR Languages (R or Python)*

# TEAMWORK

coming together is a beginning
keeping together is progress
working together is success

- Henry Ford