Faculty of Computing and Information Technology/Department of Computer Science and Mathematics

BACS3013 Data Science

Tutorial 8 (Unsupervised Learning - part 1)

Q1. Which of the following is required by K-means clustering ?

a) defined distance metric
b) number of clusters
c) initial guess as to cluster centroids
d) All of the Mentioned

Ans:d

Explanation:K-means clustering follows partitioning approach.

Q2. Point out the wrong statement:

a) k-means clustering is a method of vector quantization
b) k-means clustering aims to partition n observations into k clusters
c) k-nearest neighbor is same as k-means
d) None of the Mentioned

Ans:c

Explanation:k-nearest neighbor has nothing to do with k-means.

Q3. Hierarchical clustering should be primarily used for exploration.

a) True
b) False

Ans:a

Explanation:Hierarchical clustering is deterministic.

Q4. Which of the following clustering requires merging approach ?

a) Partitional
b) Hierarchical
c) Naive Bayes
d) None of the Mentioned

Ans:b

Explanation:Hierarchical clustering requires a defined distance as well.

BACS3013 Data Science

Q5. K-means is not deterministic and it also consist of number of iterations.

a) True
b) False

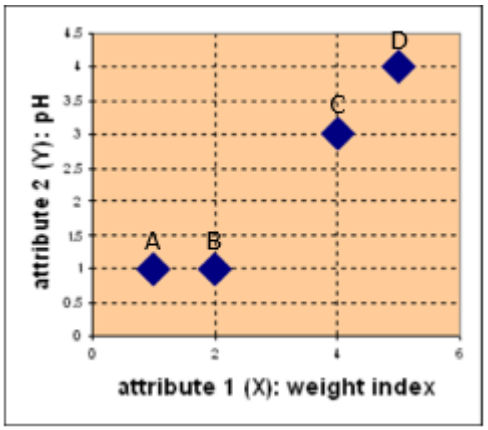
Ans:a

Explanation:K-means clustering produces final estimate of cluster centroids.

Q6. Suppose we have 4 types of medicines and each has two attributes (pH and weight index). The dataset is presented in the Table below. Use K-means with the Euclidean distance metric for clustering analysis by setting *K*=2 and initialising seeds as C1 = A and C2 = B.

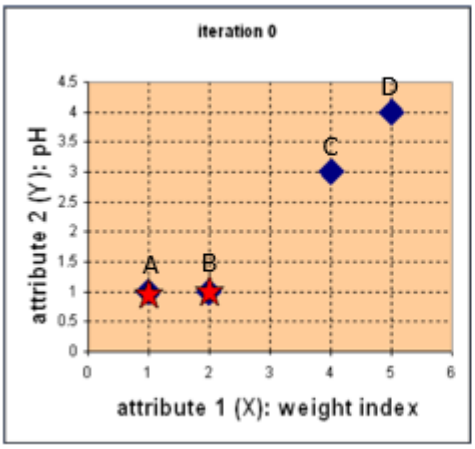
Medicine	Weight	pH-Index
A	1	1
B	2	1
C	4	3
D	5	4

Ans:



Step 1: Use initial seed points for partitioning

BACS3013 Data Science



$c_1 = A, c_2 = B$

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

Euclidean distance

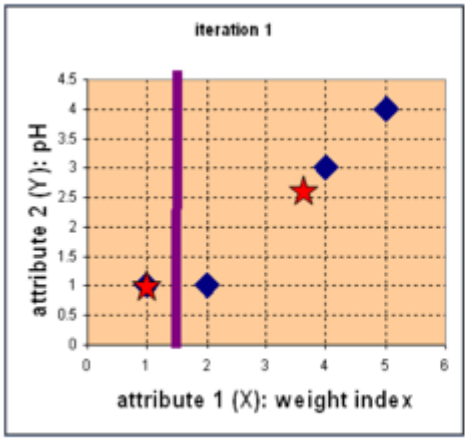
$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & & & \begin{matrix} X \\ Y \end{matrix} \end{matrix}$$

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

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

Assign each object to the cluster with the nearest seed point

Step 2: Compute new centroids of the current partition

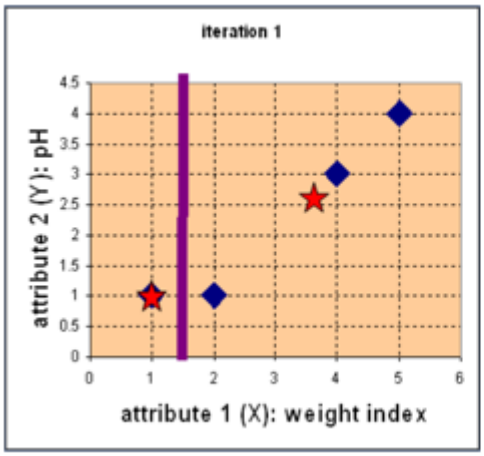


Knowing the members of each cluster, now we compute the new centroid of each group based on these new memberships.

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

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

Step 3: Renew membership based on new centroids



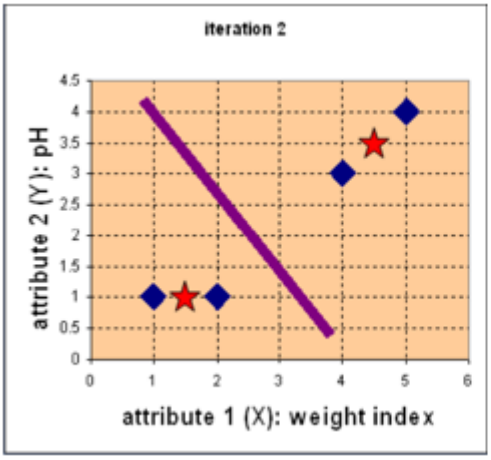
Compute the distance of all objects to the new centroids

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

$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & & & \begin{matrix} X \\ Y \end{matrix} \end{matrix}$$

Assign the membership to objects

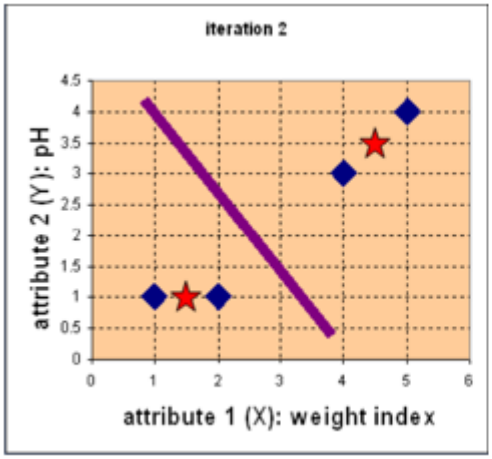
Step 4: Repeat the first two steps until its convergence



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

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

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



Compute the distance of all objects to the new centroids

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \begin{matrix} c_1 = (1\frac{1}{2}, 1) & group-1 \\ c_2 = (4\frac{1}{2}, 3\frac{1}{2}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

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

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

Q7.

a) What problem could occur in SOM learning if you use a very small neighborhood? Why?

Answer: It is possible that neurons in separate areas of the output layer become selective for the same type of input. Because there are no global interactions between neurons in the output layer, separate "islands" of similar selectivity can develop without ever "meeting" each other and merging.

b) SOMs can reduce the dimensionality of a given data space. Explain what that means. Please give an example of how this capability can be used for practical applications.

Answer: The map layer of an SOM is typically one- or two-dimensional, whereas the input space for the SOM usually has many more dimensions. Nevertheless, the SOM tries as much as possible to establish a topologyconserving mapping such that inputs from neighboring regions in the input space will make neighboring neurons (or even the same neuron) in the map layer win the competition. In this way, the high-dimensional input space is mapped onto a lowerdimensional output space. One example is the visual representation of large data sets, for example, image or document databases. Using an SOM, such high-dimensional spaces can be mapped, for example,

onto a two-dimensional space in which similar images or documents are usually close to each other. The user can browse through this space to find a desired item much more easily than through the original, high-dimensional space.

c) Why are SOMs interesting for researchers who study biological nervous systems?

The topology-conserving property of the SOM between its input and output spaces is in fact a very common feature of biological neural networks. For instance, neighboring areas in our sensory cortex respond when our ring finger or our middle finger on our right hand is being touched. However, the area responding to touch in our left foot is far away from the first two areas. The development of this mapping and its adaptation to changes (for example: losing one of the fingers) is assumed to be accomplished by competitive mechanisms as in the SOM. SOMs have been used to predict the development of neural connections in our brain with good accuracy.

(note: part (c) is to let you understand how artificial neural network and biological neural network relate to each other.)