Fuzy c-Mean Clustering

- **Types of clustering algorithms:**
  - Hard clustering:
    - Data points to only one cluster
    - Examples: K-Means
  - Soft clustering:
    - Each data point has a degree of membership (or probability) of belonging to each cluster.
    - Example: Fuzzy C-Means
- **Overview:**
  - In 1965 Professor Lotfi A. Zadeh introduced the concept of the fuzzy theory that deals with uncertain concepts.
  - Developed by Dunn in 1973 and improved vy Bezdek in 1981.
  - Fuzzy c-Mean clustering is an extension of k-means clustering algorithm.
  - Fuzzy c-means allows data points to be assigned into more than one cluster:
    - Each data point has a degree of membership (or probability) of belonging to each cluster.
    - One data point can potentially belong to multiple clusters.
  - It gives better results for overlapped dataset and comparatively better then K-mean cluster algorithm.
  - Example:
    - Gene classification using RNA sequencing.
- **Algorithms:**
  - Given:

    n: number of data points

vj: jth cluster center
xi = ith of d-dimensional measured data
m: fuzziness index m >1
c: Number of clusters
μ_{ii} Membership of ith data to jth cluster center
k: number of iterations.

- Randomly select 'c' cluster centers.
- Repeat:
  - Calculate the cluster membership probability μ_{ii} for each ith data in the jth cluster:

$$u_{ij} = \frac{\left(\frac{1}{\|x_i - c_j\|}\right)^{\frac{1}{(m-1)}}}{\sum_{k=1}^{c}\left(\frac{1}{\|x_i - c_k\|}\right)^{\frac{1}{(m-1)}}}$$

  - Compute the centroids c_j for each cluster:

$$c_j = \frac{\sum_{i=1}^{n}(u_{ij})^m x_i}{\sum_{i=1}^{n}(u_{ij})^m}$$

  - Until minimal ε achieved. (ε is between 0 and 1)

$$\varepsilon > \{|u_{ij}^{k+1} - u_{ij}^{k}|\}$$

- Notes:
  - Each cluster center is the mean of all the samples weighted by the membership value μ_{ii} (membership value).

- The algorithm adjusts the center centroids and the membership values, to minimize the weighted sum of square Euclidian distances between the centers and the data points.

- Disadvantages:
  - Need to define c, the number of clusters.
  - Need to determine membership cutoff value.
  - With a lower value of threshold we get better results but the expense of more number of iterations.
  - Clusters are sensitive to the initial assignment of centroids.
  - Fuzzy c-means is not a deterministic algorithm.
- Example:
  - Consider data points in two-dimensional space using the fuzzy C-Means algorithm.
  - Input:
    - number of objects 6
    - Number of clusters = 2
    - Fuzzification parameter m = 2
    - Threshold= 0.01
    - Max iteration = 2

Point(x,y)
(1,6)
(2,5)
(3,8)
(4,4)
(5,7)
(6,9)

  - Step 1: Randomly initialize the membership of each data point:

Point(x,y)	μ(C1)	μ(C2)
(1,6)	0.8	0.2
(2,5)	0.9	0.1

(3,8)	0.7	0.1
(4,4)	0.3	0.7
(5,7)	0.5	0.5
(6,9)	0.2	0.8

Note: The sum of the probabilities of clusters is 1

- Step 2: Calculate the centroid using Centroid equation with m=2

$$C_{11} = \frac{(1*0.8)^2+2*(0.9)^2+3*(0.7)^2+4*(0.3)^2+5*(0.5)^2+6*(0.2)^2}{(0.8)^2+(0.9)^2+(0.7)^2+(0.3)^2+(0.5)^2+(0.2)^2}$$

$$C_{12} = \frac{(6*0.8)^2+5*(0.9)^2+8*(0.7)^2+4*(0.3)^2+7*(0.5)^2+9*(0.2)^2}{(0.8)^2+(0.9)^2+(0.7)^2+(0.3)^2+(0.5)^2+(0.2)^2}$$

C1 = (2.4,6.1)

$$C_{21} = \frac{(1 * 0.2)^2 + 2 * (0.1)^2+3 * (0.3)^2+4 * (0.7)^2+5 * (0.5)^2+6 * (0.8)^2}{(0.2)^2 + (0.1)^2+(0.3)^2+(0.7)^2 + (0.5)^2+(0.8)^2}$$

$$C_{22} = \frac{(6*0.2)^2+5*(0.1)^2+8*(0.3)^2+4*(0.7)^2+7*(0.5)^2+9*(0.8)^2}{(0.2)^2+(0.1)^2+(0.3)^2+(0.7)^2+(0.5)^2+(0.8)^2}$$

C2= (4.8,6.8)

- Step 3: Calculate distance between the data points and centroid using Euclidean distance.

$$D_i = \sqrt{(x_i - c_1)^2 + (y_i - c_2)^2}$$

Centroid 1:

(1,6)(2.4,6.1) $D_1 = \sqrt{(1-2.4)^2 + (6-6.1)^2} = 1.40$

(2,5)(2.4,6.1) $D_2 = \sqrt{(2-2.4)^2 + (5-6.1)^2} = 1.17$

(3,8)(2.4,6.1) $D_3 = \sqrt{(3-2.4)^2 + (8-6.1)^2} = 1.99$

(4,4)(2.4,6.1) $D_4 = \sqrt{(4-2.4)^2 + (4-6.1)^2} = 2.64$

(5,7)(2.4,6.1) $D_5 = \sqrt{(5-2.4)^2 + (7-6.1)^2} = 2.75$

(6,8)(2.4,6.1) $D_6 = \sqrt{(6-2.4)^2 + (9-6.1)^2} = 4.62$

Centroid 2:

(1,6)(4.8,6.8) $D_1 = \sqrt{(1-4.8)^2 + (6-6.8)^2} = 3.88$

(2,5)(4.8,6.8) $D_2 = \sqrt{(2-4.8)^2 + (5-6.8)^2} = 3.32$

(3,8)(4.8,6.8) $D_3 = \sqrt{(3-4.8)^2 + (8-6.8)^2} = 2.16$

(4,4)(4.8,6.8) $D_4 = \sqrt{(4-4.8)^2 + (4-6.8)^2} = 2.91$

(5,7)(4.8,6.8) $D_5 = \sqrt{(5-4.8)^2 + (7-6.8)^2} = 0.28$

(6,8)(4.8,6.8) $D_6 = \sqrt{(6-4.8)^2 + (9-6.8)^2} = 2.50$

Points	Distance(C1)	Distance(C2)
(1,6)	1.40	3.88
(2,5)	1.17	3.32
(3,8)	1.99	2.16
(4,4)	2.64	2.91
(5,7)	2.75	0.28
(6,9)	4.62	2.50

- Step 4: Update the new membership matrix using the equation.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{\frac{2}{m-1}}}$$

Cluster 1:

$$\mu_{11} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{11}}{d_{k1}})^{\frac{2}{2-1}}} = \frac{1}{(\frac{d_{11}}{d_{11}})^{\frac{2}{2-1}} + (\frac{d_{11}}{d_{21}})^{\frac{2}{2-1}}}$$

$$\mu_{11} = \frac{1}{(\frac{1.4}{1.4})^2 + (\frac{1.4}{3.88})^2} = 0.89$$

Cluster 2:

$$\mu_{21} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{21}}{d_{k1}})^{\frac{2}{2-1}}} = \frac{1}{(\frac{d_{21}}{d_{11}})^{\frac{2}{2-1}} + (\frac{d_{21}}{d_{21}})^{\frac{2}{2-1}}}$$

$$\mu_{11} = \frac{1}{(\frac{3.88}{1.40})^2 + (\frac{3.88}{3.88})^2} = 0.11$$

New Membership=μ^1	
Point(x,y)	μ(C1)
(1,6)	0.89
(2,5)	0.9
(3,8)	0.54
(4,4)	0.55
(5,7)	0.01
(6,9)	0.23

Old Membership μ^0	
Point(x,y)	μ(C1)
(1,6)	0.8
(2,5)	0.9
(3,8)	0.7
(4,4)	0.3
(5,7)	0.5
(6,9)	0.2

Max$\lvert(\mu^1 - \mu^0)\rvert$
0.09
0
0.16
0.25
0.49
0.03

- Step 5:

If Max$\lvert(\mu^1 - \mu^0)\rvert$ =0.49 <=0.01 or iteration 1 = ma iteration(2)

Stop; //(none of the conditions is true since we are in iteration

// 1 and 0.49 is greater that 0.01 (threshold)

Else

// then repeat

Return to step 2 iteration = iteration + 1;