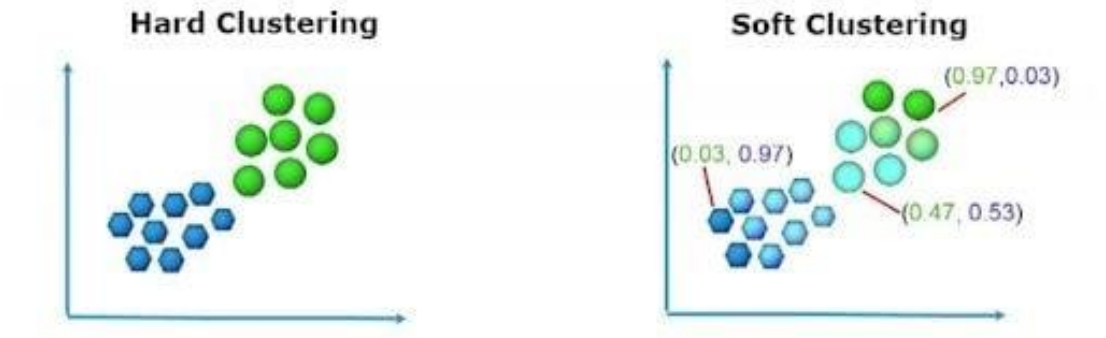
Clustering

Clustering is used to identify some segments or groups in your dataset. Clustering can be divided into two sub-groups:



What Is Hard Clustering?

In hard clustering, each data point is clustered or grouped to any one cluster. For each data point, it may either completely belong to a cluster or not. As observed in the above diagram, the data points are divided into two clusters, each point belonging to either of the two clusters.

K-means clustering is a hard clustering algorithm. It clusters data points into k-clusters.

What Is Soft Clustering?

In soft clustering, instead of putting each data point into separate clusters, a probability of that point is assigned to probable clusters. In soft clustering or fuzzy clustering, each data point can belong to multiple clusters along with its probability score or likelihood.

One of the widely used soft clustering algorithms is the fuzzy c-means clustering (FCM) Algorithm.

Fuzzy Clustering

Fuzzy Clustering is a type of clustering algorithm in machine learning that allows a data point to belong to more than one cluster with different degrees of membership. Unlike traditional clustering algorithms, such as k-means or hierarchical clustering, which assign each data point to a single cluster, fuzzy clustering assigns a membership degree between 0 and 1 for each data point for each cluster.

The steps to perform the algorithm are:

Step 1: Initialize the data points into the desired number of clusters randomly.

Let us assume there are 2 clusters in which the data is to be divided, initializing the data point randomly. Each data point lies in both clusters with some membership value which can be assumed anything in the initial state.

The table below represents the values of the data points along with their membership (gamma) in each cluster.

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1)	0.8	0.7	0.2	0.1
2)	0.2	0.3	0.8	0.9

Step 2: Find out the centroid.
The formula for finding out the centroid (V) is:

$$V_{ij} = (\sum_{1}^{n}(\gamma_{ik}^{m} * x_k) / \sum_{1}^{n} \gamma_{ik}^{m}$$

Where, μ is fuzzy membership value of the data point, m is the fuzziness parameter (generally taken as 2), and xk is the data point.
Here,

V11 = (0.8^2 *1 + 0.7^2 * 2 + 0.2^2 * 4 + 0.1^2 * 7) / ((0.8^2 + 0.7^2 + 0.2^2 + 0.1^2) = 1.568
V12 = (0.8^2 *3 + 0.7^2 * 5 + 0.2^2 * 8 + 0.1^2 * 9) / ((0.8^2 + 0.7^2 + 0.2^2 + 0.1^2) = 4.051
V21 = (0.2^2 *1 + 0.3^2 * 2 + 0.8^2 * 4 + 0.9^2 * 7) / ((0.2^2 + 0.3^2 + 0.8^2 + 0.9^2) = 5.35
V22 = (0.2^2 *3 + 0.3^2 * 5 + 0.8^2 * 8 + 0.9^2 * 9) / ((0.2^2 + 0.3^2 + 0.8^2 + 0.9^2) = 8.215
Centroids are: (1.568, 4.051) and (5.35, 8.215)

Step 3: Find out the distance of each point from the centroid.

$D11 = ((1 - 1.568)^2 + (3 - 4.051)^2)^{0.5} = 1.2$
$D12 = ((1 - 5.35)^2 + (3 - 8.215)^2)^{0.5} = 6.79$

Similarly, the distance of all other points is computed from both the centroids.

Step 4: Updating membership values.

$$\gamma = \sum_{1}^{n} (d_{ki}^2 / d_{kj}^2)^{1/m-1}]^{-1}$$

For point 1 new membership values are:

γ_{11}= [{ $[(1.2)^2 / (1.2)^2] + [(1.2)^2 / (6.79)^2]$} ∧ {(1 / (2 – 1))}] $^{-1}$ = 0.96

γ_{12}= [{ $[(6.79)^2 / (6.79)^2] + [(6.79)^2 / (1.2)^2]$} ∧ {(1 / (2 – 1))}] $^{-1}$ = 0.04

Alternatively,

$$\gamma_{12} = 1 - \gamma_{11} = 0.04$$

Similarly, compute all other membership values, and update the matrix.

Step 5: Repeat the steps(2-4) until the constant values are obtained for the membership values or the difference is less than the tolerance value (a small value up to which the difference in values of two consequent updations is accepted).

Solved Example

Fuzzy C-Means Clustering – Steps

- **Step 1:** Given the data points based on the number of clusters required initialize the membership table with random values.
- Suppose the given data points are **{(1, 3), (2, 5), (6, 8), (7, 9)}**

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

- **Step 2: Find out the centroid.**
- The formula for finding out the centroid (V) is:

- $V_{ij} = \frac{\sum_{k=1}^{n} \gamma_{ik}^{m} * x_k}{\sum_{k=1}^{n} \gamma_{ik}^{m}}$

  - γ: Fuzzy membership value
  - m: Fuzziness parameter generally taken as 2 and
  - x_k is the data point

- $V_{11} = \frac{(0.8^2 * 1 + 0.7^2 * 2 + 0.2^2 * 4 + 0.1^2 * 7)}{(0.8^2 + 0.7^2 + 0.2^2 + 0.1^2)} = 1.568$
- $V_{12} = \frac{(0.8^2 * 3 + 0.7^2 * 5 + 0.2^2 * 8 + 0.1^2 * 9)}{(0.8^2 + 0.7^2 + 0.2^2 + 0.1^2)} = 4.051$

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

- $V_{21} = \frac{(0.2^2 * 1 + 0.3^2 * 2 + 0.8^2 * 4 + 0.9^2 * 7)}{(0.2^2 + 0.3^2 + 0.8^2 + 0.9^2)} = 5.35$
- $V_{22} = \frac{(0.2^2 * 3 + 0.3^2 * 5 + 0.8^2 * 8 + 0.9^2 * 9)}{(0.2^2 + 0.3^2 + 0.8^2 + 0.9^2)} = 8.215$

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

Centroids are: (1.568, 4.051) and (5.35, 8.215)

- **Step 3: Find out the distance of each point from the centroid.**

- $D_{11} = \sqrt{(1 - 1.568)^2 + (3 - 4.051)^2} = 1.2$
- $D_{12} = \sqrt{(1 - 5.35)^2 + (3 - 8.215)^2} = 6.79$
- $D_{21} = \sqrt{(2 - 1.568)^2 + (5 - 4.051)^2} = 1.04$
- $D_{22} = \sqrt{(2 - 5.35)^2 + (5 - 8.215)^2} = 4.64$
- $D_{31} = \sqrt{(4 - 1.568)^2 + (8 - 4.051)^2} = 4.63$
- $D_{32} = \sqrt{(4 - 5.35)^2 + (8 - 8.215)^2} = 1.36$
- $D_{31} = \sqrt{(7 - 1.568)^2 + (9 - 4.051)^2} = 7.34$
- $D_{32} = \sqrt{(7 - 5.35)^2 + (9 - 8.215)^2} = 1.82$

Centroids are:
(1.568, 4.051) and
(5.35, 8.215)

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9
	1	1	2	2

- **Step 4: Updating membership values.**

- $\gamma_{ki} = \left(\sum_{j=1}^{n} \left\{ \frac{d_{ki}^2}{d_{kj}^2} \right\}^{\left(\frac{1}{(m-1)}\right)} \right)^{-1}$

$D_{11} = 1.2, \quad D_{12} = 6.79$
$D_{21} = 1.04, \quad D_{22} = 4.64$
$D_{31} = 4.63, \quad D_{32} = 1.36$
$D_{31} = 7.34, \quad D_{32} = 1.82$

- For point 1 new membership values are:

- $\gamma_{11} = \left(\left\{ \frac{(1.2)^2}{(1.2)^2} + \frac{(1.2)^2}{(6.79)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.97$
- $\gamma_{12} = \left(\left\{ \frac{(6.79)^2}{(1.2)^2} + \frac{(6.79)^2}{(6.79)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.03$

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	**0.97**	0.7	0.2	0.1
2	**0.03**	0.3	0.8	0.9

- **Step 4: Updating membership values.**

- $\gamma_{ki} = \left(\sum_{j=1}^{n} \left\{ \frac{d_{ki}^2}{d_{kj}^2} \right\}^{\left(\frac{1}{(m-1)}\right)} \right)^{-1}$

$D_{11} = 1.2, \quad D_{12} = 6.79$
$D_{21} = 1.04, \quad D_{22} = 4.64$
$D_{31} = 4.63, \quad D_{32} = 1.36$
$D_{31} = 7.34, \quad D_{32} = 1.82$

- For point 2 new membership values are:

- $\gamma_{21} = \left(\left\{ \frac{(1.04)^2}{(1.04)^2} + \frac{(1.04)^2}{(4.64)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.95$
- $\gamma_{22} = \left(\left\{ \frac{(4.64)^2}{(1.04)^2} + \frac{(4.64)^2}{(4.64)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.05$

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	**0.97**	**0.95**	0.2	0.1
2	**0.03**	**0.05**	0.8	0.9

- **Step 4: Updating membership values.**

- $\gamma_{ki} = \left(\sum_{j=1}^{n} \left\{ \frac{d_{ki}^2}{d_{kj}^2} \right\}^{\left(\frac{1}{(m-1)}\right)} \right)^{-1}$

$D_{11} = 1.2, \quad D_{12} = 6.79$
$D_{21} = 1.04, \quad D_{22} = 4.64$
$D_{31} = 4.63, \quad D_{32} = 1.36$
$D_{31} = 7.34, \quad D_{32} = 1.82$

- For point 3 new membership values are:

- $\gamma_{31} = \left(\left\{ \frac{(4.63)^2}{(4.63)^2} + \frac{(4.63)^2}{(1.36)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.08$
- $\gamma_{32} = \left(\left\{ \frac{(1.36)^2}{(4.63)^2} + \frac{(1.36)^2}{(1.36)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.92$

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.97	0.95	0.08	0.1
2	0.03	0.05	0.92	0.9

- **Step 4: Updating membership values.**

- $\gamma_{ki} = \left(\sum_{j=1}^{n} \left\{ \frac{d_{ki}^2}{d_{kj}^2} \right\}^{\left(\frac{1}{(m-1)}\right)} \right)^{-1}$

$D_{11} = 1.2, \quad D_{12} = 6.79$
$D_{21} = 1.04, \quad D_{22} = 4.64$
$D_{31} = 4.63, \quad D_{32} = 1.36$
$D_{31} = 7.34, \quad D_{32} = 1.82$

- For point 4 new membership values are:

- $\gamma_{41} = \left(\left\{ \frac{(7.34)^2}{(7.34)^2} + \frac{(7.34)^2}{(1.82)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.06$
- $\gamma_{42} = \left(\left\{ \frac{(1.82)^2}{(7.34)^2} + \frac{(1.82)^2}{(1.82)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.94$

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.97	0.95	0.08	0.06
2	0.03	0.05	0.92	0.94

- **Step 5:** Repeat the steps (2-4) until the constant values are obtained for the membership values or the difference is less than the tolerance value

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

$t = 0.01$

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.97	0.95	0.08	0.06
2	0.03	0.05	0.92	0.94

When to apply this technique?

Fuzzy c-means clustering gives better results for overlapped data sets compared to k-means clustering. In other words, clusters are formed in a way that: Data points in the same cluster are close to each other and are very similar. Data points in different clusters are far apart and are different from each other. The performance of the fuzzy c-means algorithm gives better performance than k-mean, both when using thresholding with mean and median methods. Better performance of fuzzy c-means requires additional time when compared to k-means

Applications in several fields of Fuzzy clustering :

1. **Image segmentation**: Fuzzy clustering can be used to segment images by grouping pixels with similar properties together, such as color or texture.

2. **Pattern recognition**: Fuzzy clustering can be used to identify patterns in large datasets by grouping similar data points together.

3. **Marketing:** Fuzzy clustering can be used to segment customers based on their preferences and purchasing behavior, allowing for more targeted marketing campaigns.

4. **Medical diagnosis**: Fuzzy clustering can be used to diagnose diseases by grouping patients with similar symptoms together.

5. **Environmental monitoring:** Fuzzy clustering can be used to identify areas of environmental concern by grouping together areas with similar pollution levels or other environmental indicators.

6. **Traffic flow analysis**: Fuzzy clustering can be used to analyze traffic flow patterns by grouping similar traffic patterns together, allowing for better traffic management and planning.

7. **Risk assessment:** Fuzzy clustering can be used to identify and quantify risks in various fields, such as finance, insurance, and engineering.

Advantages of Fuzzy Clustering:

1. **Flexibility:** Fuzzy clustering allows for overlapping clusters, which can be useful when the data has a complex structure or when there are ambiguous or overlapping class boundaries.

2. **Robustness:** Fuzzy clustering can be more robust to outliers and noise in the data, as it allows for a more gradual transition from one cluster to another.

3. **Interpretability:** Fuzzy clustering provides a more nuanced understanding of the structure of the data, as it allows for a more detailed representation of the relationships between data points and clusters.

Disadvantages of Fuzzy Clustering:

1. C**omplexity**: Fuzzy clustering algorithms can be computationally more expensive than traditional clustering algorithms, as they require optimization over multiple membership degrees.

2. **Model selection:** Choosing the right number of clusters and membership functions can be challenging, and may require expert knowledge or trial and error.

Reference

1. https://www.youtube.com/watch?v=X7co6-U4BJY
2. https://www.geeksforgeeks.org/ml-fuzzy-clustering/
3. https://builtin.com/data-science/c-means