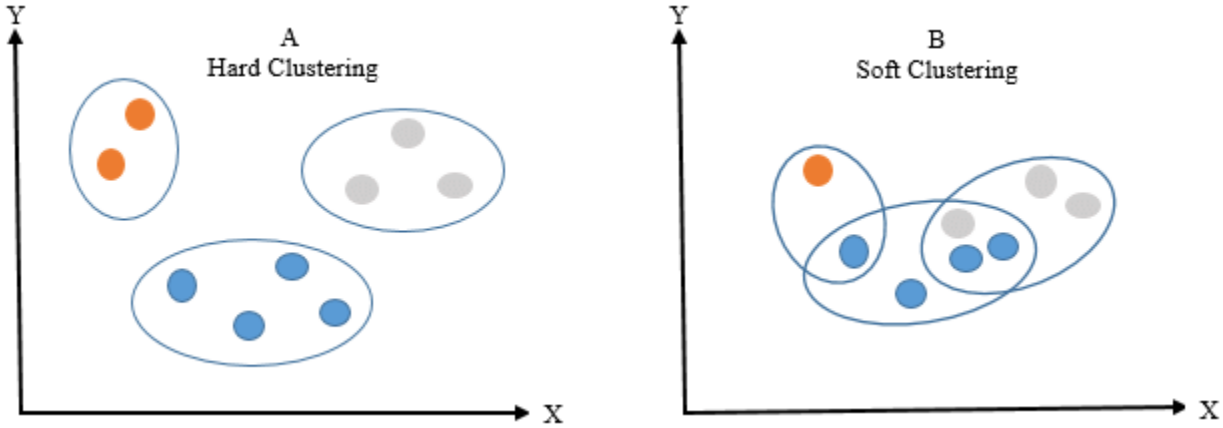
1. Thuật toán phân cụm

- Phân cụm: là một thuật toán học máy không giám sát dựa trên khoảng cách. Trong đó, các điểm dữ liệu gần nhau được nhóm thành một số cụm hoặc nhóm nhất định.
- Phân loại:
    - Phân cụm cứng (Hard Clustering): mỗi điểm dữ liệu dược gán cho một cụm, tức nó thuộc về một cụm cụ thể nào đó. Thuật toán K-Means là một thuật toán phân cụm cứng.
    - Phân cụm mềm (Soft Clustering): mỗi điểm dữ liệu thuộc về các cụm có xác suất nhất định còn được gọi là membership value. Ví dụ thuật toán Fuzzy C-Means.



1.1 Thuật toán phân cụm cứng K-Means

1.2 Thuật toán phân cụm mềm Fuzzy C-Means

- Bước 1: Thuật toán được cung cấp các điểm dữ liệu dựa trên số cụm đã khởi tạo bảng membership với các giá trị ngẫu nhiên. Ví dụ: Cung cấp 4 điểm dữ liệu: {(1,3), (2,5), (6,8), (7,9)}. Mỗi điểm dữ liệu chứa 2 thành phần ở đây có thể gọi là 2 features. Giả sử khởi tạo 2 cụm

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

- Bước 2: Tìm tâm cụm dựa theo công thức:

$$v_{ij} = \frac{\Sigma_{k=1}^{n} \gamma_{ik}^{m} * x_k}{\Sigma_{k=1}^{n} \gamma_{ik}^{m}}$$

    - Trong đó:
        - γ : Fuzzy membership value

- m : Fuzziness parameter generally taken as 2
- x_k is the data point

  - Tính tâm cụm 1
    - $v_{11} = \frac{(0{,}8^2 * 1 + 0{,}7^2 * 2 + 0{,}2^2 * 4 + 0{,}1^2 * 7)}{(0{,}8^2 + 0{,}7^2 + 0{,}2^2 + 0{,}1^2)} = 1{,}568$
    - $v_{12} = \frac{(0{,}8^2 * 3 + 0{,}7^2 * 5 + 0{,}2^2 * 8 + 0{,}1^2 * 9)}{(0{,}8^2 + 0{,}7^2 + 0{,}2^2 + 0{,}1^2)} = 4{,}051$
  - Tính tâm cụm 2
    - $v_{21} = \frac{(0{,}2 * 1 + 0{,}3^2 * 2 + 0{,}8^2 * 4 + 0{,}9^2 * 7)}{(0{,}2^2 + 0{,}3^2 + 0{,}8^2 + 0{,}9^2)} = 5{,}35$
    - $v_{22} = \frac{(0{,}2 * 3 + 0{,}3 * 5 + 0{,}8^2 * 8 + 0{,}9^2 * 9)}{(0{,}2^2 + 0{,}3^2 + 0{,}8^2 + 0{,}9^2)} = 8{,}215$
  - Suy ra, Tâm của 2 cụm lần lượt là: (1.568, 4.051) và (5.35, 8.215)
- Bước 3: Tính khoảng cách giữa các điểm dữ liệu đến các tâm cụm.
  - Sử dụng công thức tính khoảng cách Euclide.

$$D_{11} = \sqrt{(1 - 1.568)^2 + (3 - 4.051)^2} = 1.2$$

$$D_{12} = \sqrt{(1 - 5.35)^2 + (3 - 8.215)^2} = 6.79$$

$$D_{21} = \sqrt{(2 - 1.568)^2 + (5 - 4.051)^2} = 1.04$$

$$D_{22} = \sqrt{(2 - 5.35)^2 + (5 - 8.215)^2} = 4.64$$

$$D_{31} = \sqrt{(4 - 1.568)^2 + (8 - 4.051)^2} = 4.63$$

$$D_{32} = \sqrt{(4 - 5.35)^2 + (8 - 8.215)^2} = 1.36$$

$$D_{41} = \sqrt{(7 - 1.568)^2 + (9 - 4.051)^2} = 7.34$$

$$D_{42} = \sqrt{(7 - 5.35)^2 + (9 - 8.215)^2} = 1.82$$

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

- Bước 4: Cập nhật Membership values
  - Công thức tính membership values của điểm thứ k cho cụm thứ i:
    - $\gamma_{ki} = \left(\Sigma_{j=1}^{n} \left\{ \frac{d_{ki}^2}{d_{kj}^2} \right\}^{\frac{1}{m-1}} \right)^{-1}$
  - Tính new membership values cho điểm thứ nhất:
    - $\gamma_{11} = \left(\left\{ \frac{(1.2)^2}{(1.2)^2} + \frac{(1.2)^2}{(6.79)^2} \right\}^{\frac{1}{2-1}} \right)^{-1} = 0.97$
    - $\gamma_{12} = \left(\left\{ \frac{(6.79)^2}{(1.2)^2} + \frac{(6.79)^2}{(6.79)^2} \right\}^{\frac{1}{2-1}} \right)^{-1} = 0.03$
  - Tính new membership values cho điểm thứ 2:
    - $\gamma_{21} = \left(\left\{ \frac{(1.04)^2}{(1.04)^2} + \frac{(1.04)^2}{(4.64)^2} \right\}^{\frac{1}{2-1}} \right)^{-1} = 0.95$
    - $\gamma_{21} = \left(\left\{ \frac{(4.64)^2}{(1.04)^2} + \frac{(4.64)^2}{(4.64)^2} \right\}^{\frac{1}{2-1}} \right)^{-1} = 0.05$
  - Tính new membership values cho điểm thứ 3:
    - $\gamma_{31} = \left(\left\{ \frac{(4.63)^2}{(4.63)^2} + \frac{(4.63)^2}{(1.36)^2} \right\}^{\frac{1}{2-1}} \right)^{-1} = 0.08$
    - $\gamma_{31} = \left(\left\{ \frac{(1.36)^2}{(4.63)^2} + \frac{(1.36)^2}{(1.36)^2} \right\}^{\frac{1}{2-1}} \right)^{-1} = 0.92$
  - Tính new membership values cho điểm thứ 4:
    - $\gamma_{31} = \left(\left\{ \frac{(7.34)^2}{(7.34)^2} + \frac{(7.34)^2}{(1.82)^2} \right\}^{\frac{1}{2-1}} \right)^{-1} = 0.06$
    - $\gamma_{31} = \left(\left\{ \frac{(1.82)^2}{(7.34)^2} + \frac{(1.82)^2}{(1.82)^2} \right\}^{\frac{1}{2-1}} \right)^{-1} = 0.94$
  - Suy ra bảng membership values mới

Cluster	(1,3)	(2,5)	(4,8)	(7,9)
1	0.8 -> 0.97	0.7 -> 0.95	0.2 -> 0.08	0.1 -> 0.06
2	0.2 -> 0.03	0.3 -> 0.05	0.8 -> 0.92	0.9 -> 0.94

- Bước 5: Lặp lại bước 2-4, cho tới khi giá trị dung sai nhỏ hơn tolerance value. Giả sử tolerance = 0.01 thì $\gamma_t - \gamma_{t-1} < 0.01$