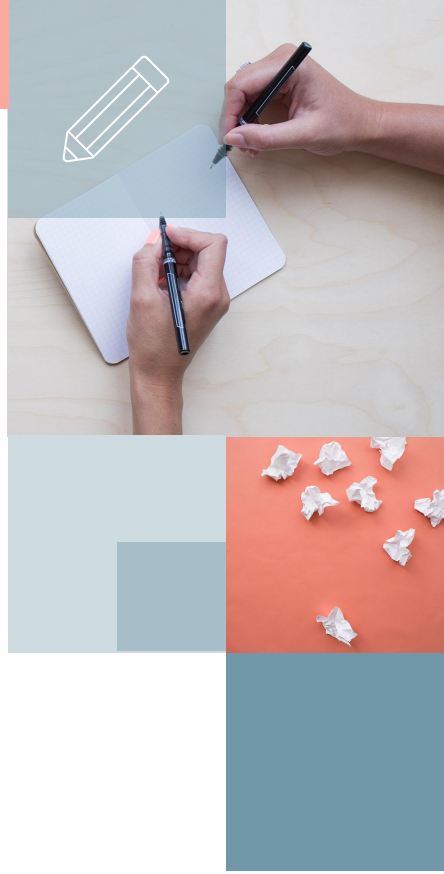
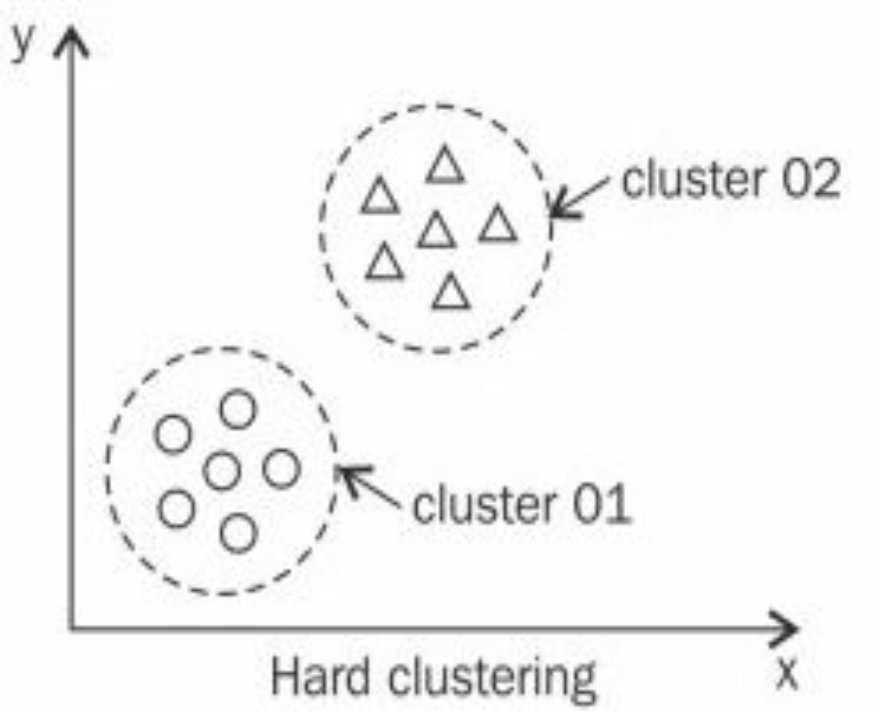
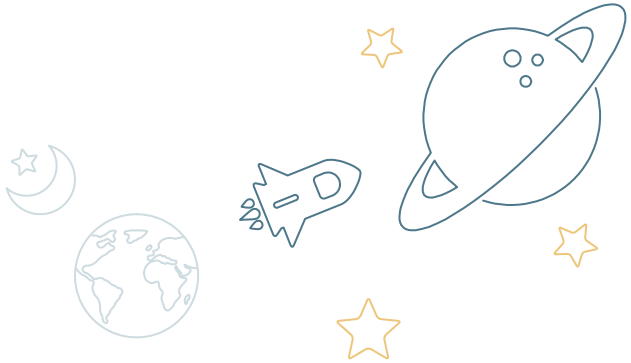

K-Means
Fuzzy C-Means

# Hard Clustering

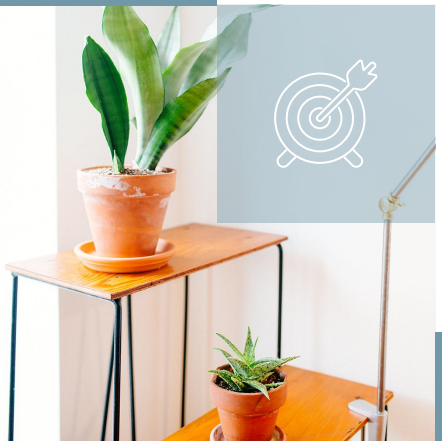
Hard clustering is about grouping the data items such that each item is only assigned to one cluster. For example, we want the algorithm to read all of the tweets and determine if a tweet is a positive or a negative tweet.

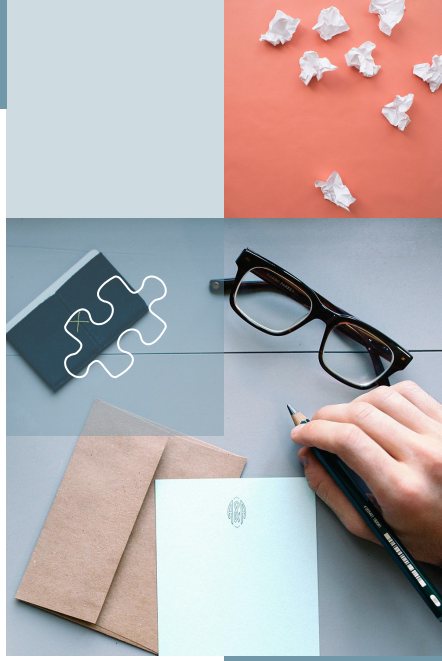



y
cluster 02
cluster 01
Hard clustering
x



# K means

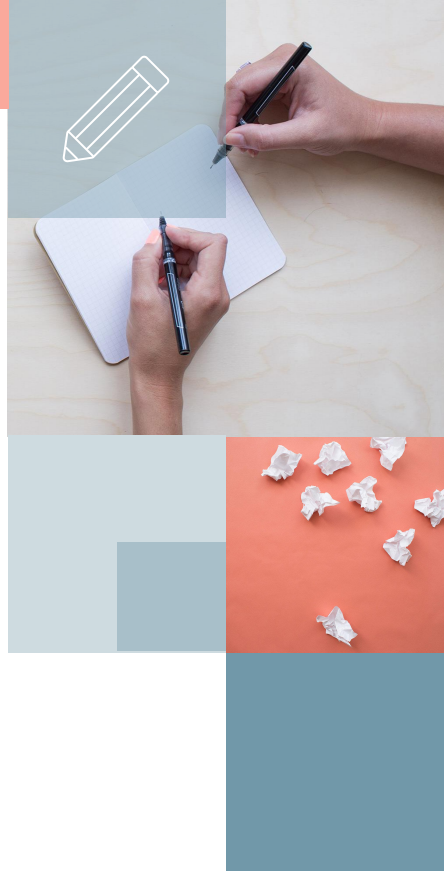


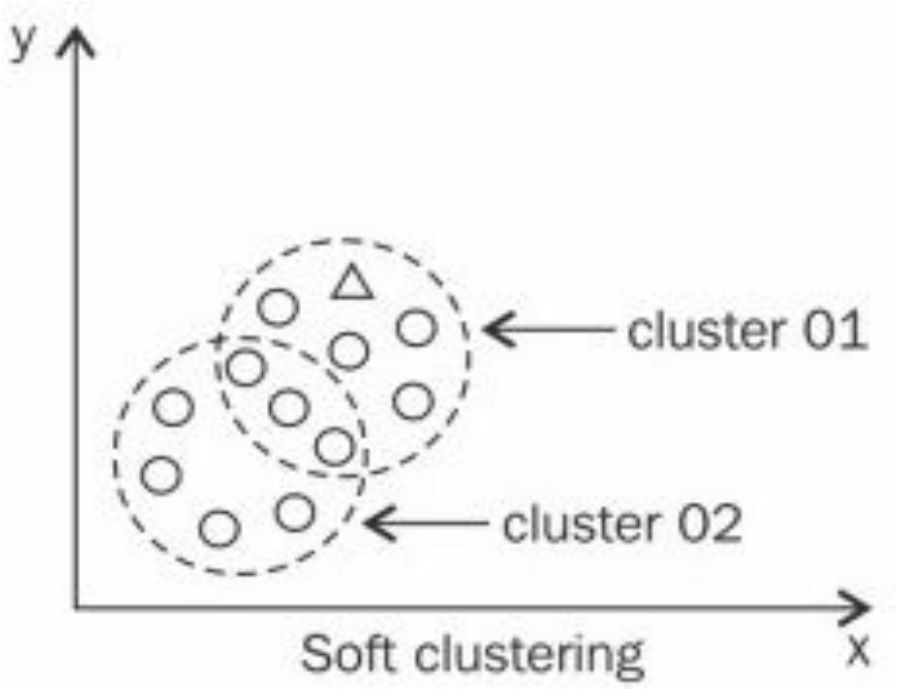


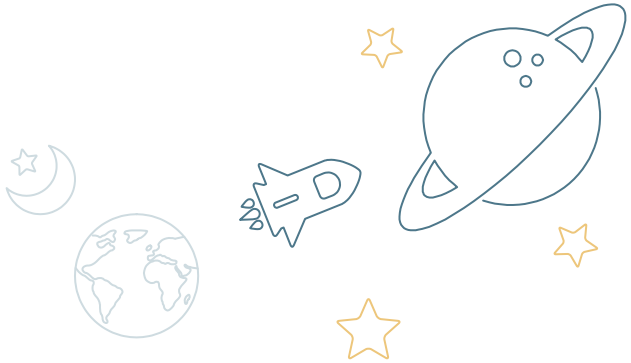
1. Specify the desired number of clusters
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closet cluster centroid
5. Re-compute cluster centroids
6. We repeat steps 4 and 5 until no further improvement is found in the centroid values

# Soft Clustering

Fuzzy clustering (also referred to as soft clustering or soft k-means) is a form of clustering in which each data point can belong to more than one cluster. For example, preferences of customers in a retail store.




y
cluster 01
cluster 02
Soft clustering
x

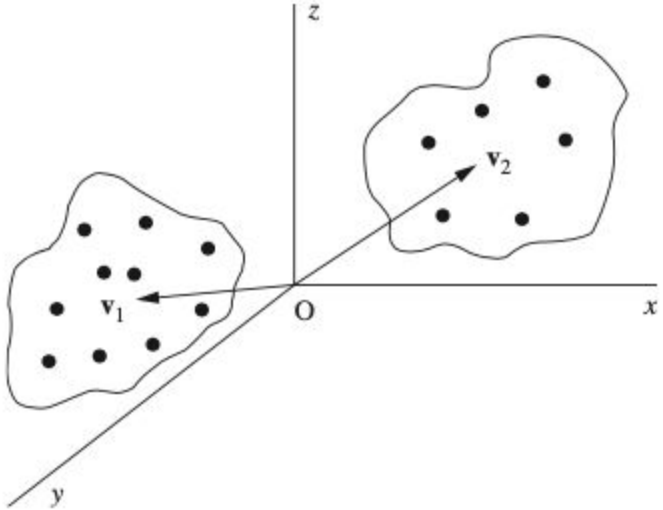


# Fuzzy C means


## C Means

Bezdek (1981) suggested using an objective function approach for clustering the data into hyperspherical clusters.

-> minimize the Euclidean distance between each data point in a cluster and its cluster center (a calculated point).
-> maximize the Euclidean distance between cluster centers.

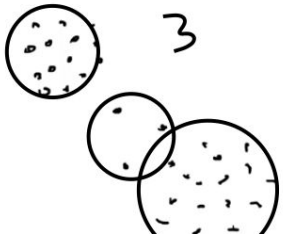


## Fuzzy clustering

Fuzzy clustering is a clustering technique in which each data point belongs to two or more clusters. It is also referred as soft k-means or soft clustering. Clusters are identified using similarity measures.

These similarity measures include connectivity, intensity and Distance.

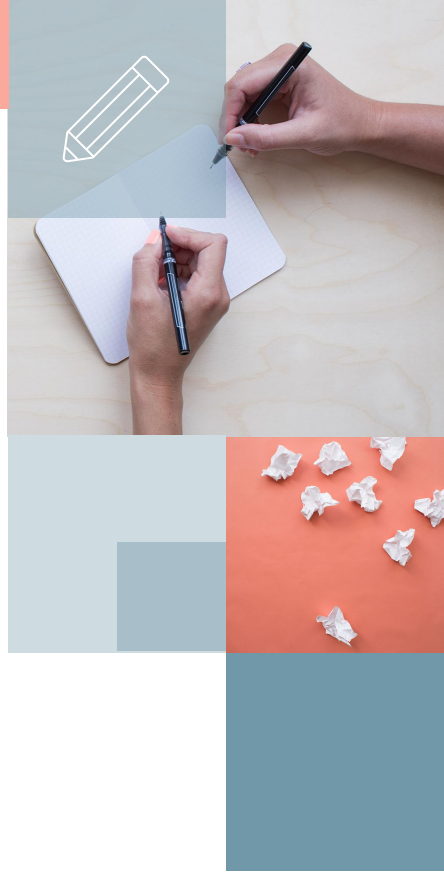
One of the most widely used fuzzy clustering algorithm is Fuzzy C-means Clustering.

## Fuzzy C-Means Clustering (FCM)

Fuzzy c-means clustering (FCM) is a data clustering technique in which the data points is grouped into N clusters with every data point that belongs to other clusters to a certain degree.

**Example**: The data points which lies close to the centre of a cluster will have a high degree of membership in that cluster, and other data points which lies far away from the centre of the cluster will have a low degree of membership to that cluster.



## Steps for fuzzy c-means clustering:

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

Step 2: Find out the centroid of each cluster.

The formula for finding the centroid of cluster

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

Where,

m = fuzziness parameter (generally taken as 2)

xk = data point.

## Steps for fuzzy c-means clustering:

Step 3: Find out the distance of each point from both the centroids using euclidean distance

Step 4: Update the membership values using this equation.

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

Step 5: Repeat the steps (2-4), unless centroids are not changing.

# Example:

Consider data points {(1, 3), (2, 5), (6, 8), (7, 9)}

**Step 1:** Let's consider no of clusters is 2 in which the data is to be divided

| 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 of the cluster

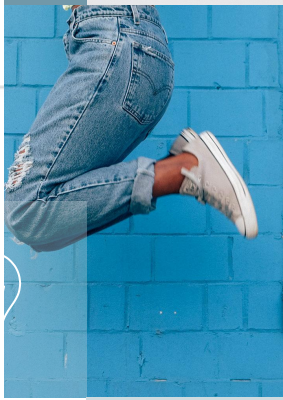
V11 = (0.82 *1 + 0.72 * 2 + 0.22 * 4 + 0.12 * 7) / (0.82 + 0.72 + 0.22 + 0.12) = 1.568

V12 = (0.82 *3 + 0.72 * 5 + 0.22 * 8 + 0.12 * 9) / (0.82 + 0.72 + 0.22 + 0.12) = 4.051

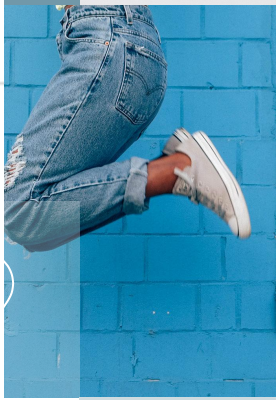
V11 = (0.22 *1 + 0.32 * 2 + 0.82 * 4 + 0.92 * 7) / (0.22 + 0.32 + 0.82 + 0.92) = 5.35

V11 = (0.22 *3 + 0.32 * 5 + 0.82 * 8 + 0.92 * 9) / (0.22 + 0.32 +0.82 + 0.92) = 8.215

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



# Contd..




**Step 3:** Find out the distance of each point from both the centroids.

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 data points is computed from both the centroids.

**Step 4:** Update the membership values.

$$\gamma_{11} = [\{ [(1.2)^2 / (1.2)^2] + [(1.2)^2 / (6.79)^2]\} \wedge \{(1 / (2 - 1))\}]^{-1} = 0.96$$

$$\gamma_{12} = [\{ [(6.79)^2 / (6.79)^2] + [(6.79)^2 / (1.2)^2]\} \wedge \{(1 / (2 - 1))\}]^{-1} = 0.04$$

Similarly, compute all membership values and update the matrix.

**Step 5:** Repeat the steps (2-4), unless centroids are not changing.

# Fuzzy clustering (c means)

Data $X = \{x_1, x_2, \cdots, x_n\}$

Clusters $= \{c_1, c_2, \cdots, c_K\}$

Eg:-

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | |
|---|---|---|---|---|---|
| $x_1$ | .02 | .03 | .01 | .94 | $\Rightarrow \Sigma = 1$ |
| $x_2$ | .8 | .01 | .10 | .09 | $\Rightarrow \Sigma = 1$ |
| $x_3$ | .01 | .01 | .97 | .01 | $\Rightarrow \Sigma = 1$ |

$w_{11} = 0.02$

data <-L-> cluster

Conditions $\rightarrow \sum_{j=1}^{K} w_{ij} = 1$

$\rightarrow 0 < \sum_{i=1}^{n} w_{ij} < n$

Fuzzy c-means (FCM)

Input : Data, K

Output : $w_{ij}$, $c_j$ ; $1 \le j \le K$ ; $1 \le i \le n$ ;

Steps

$$SSE = \sum_{j=1}^{K} \sum_{i=1}^{n} w_{ij}^{p} \, dist^2(x_i, c_j) \quad , p \text{ is } 1 \text{ to } \infty$$

Step 1

-> Initialize data points

Step 2

-> Compute centroids

$$c_j = \frac{\sum_{i=1}^{n} w_{ij}^{p} x_i}{\sum_{i=1}^{n} w_{ij}^{p}}$$

$p = 0$, same as K means

$p = 2, 4, 10 \cdots$, more fuzzification

Here $p = 2$.

**Step 3**

- Find distance bt. pts

**Step 4**

- update fuzzy pseudo - partition

$$w_{ij} = \left(\frac{1}{dist(x_i, c_j)}\right)^{1/p-1} \Big/ \sum_{q=1}^{K} \left(\frac{1}{dist(x_i, c_q)}\right)^{1/p-1}$$

If $p = 2$

$$w_{ij} = \frac{1/dist(x_i, c_j)}{\sum_{q=1}^{K} \frac{1}{dist(x_i, c_q)}}$$



Problem

S1:

Data (1,2) (2,3) (9,4) (10,1)

| | | |
|---|---|---|
| $x_1$ | 1 | 2 |
| $x_2$ | 2 | 3 |
| $x_3$ | 9 | 4 |
| $x_4$ | 10 | 1 |

| | wt. $c_1$ | $c_2$ |
|---|---|---|
| $x_1$ | .4 | .6 |
| $x_2$ | .88 | .12 |
| $x_3$ | .41 | .59 |
| $x_4$ | .27 | .73 |
| $\sum w_{ij}^2$ | 1.18 | 1.25 |

S2:

$$\sum_{i=1}^{\hat{}} w_{i1}^2 = (.4)^2 + (.88)^2 + (.41)^2 + (.27)^2 = 1.18$$

$$\sum_{i=2}^{\hat{}} w_{i2}^2 = (.6)^2 + (.12)^2 + (.59)^2 + (.73)^2 = 1.25$$

two dimensional here, $C_1 = [c_{11}, c_{12}]$ $C_2 = [c_{21}, c_{22}]$

$$C_{11} = \frac{(.4)^2 \times 1 + (.88)^2 \times 2 + (.41)^2 \times 9 + (.27)^2 \times 10}{1.18}$$

$$= 3.97/1.18 = 3.38$$

No.

Date

$$C_{12} = \frac{(.4)^2 \times 2 + (.38)^2 \times 3 + (.41)^2 \times 4 + (.27)^2 \times 1}{1.18}$$

$$\approx 2.88$$

$$C_{21} = \frac{(6)^2 \times 1 + (.12)^2 \times 2 + (.59)^2 \times 9 + (.75)^2 \times 10}{1.25}$$

$$= 7.02$$

$$C_{22} = \frac{(.6)^2 \times 2 + (.12)^2 \times 3 + (.59)^2 \times 4 + (.73)^2 \times 1}{1.25}$$

$$= 2.14$$

$C_1 = [3.38, 2.88]$ $C_2 = [7.02, 2.14]$



S3:

| | cluster | centroid |
|---|---|---|
| $c_1$ | 3.38 | 2.88 |
| $c_2$ | 7.02 | 2.14 |

Distance

| | $c_1$ | $c_2$ |
|---|---|---|
| $x_1$ | 2.54 | 6.03 |
| $x_2$ | 1.38 | 5.10 |
| $x_3$ | 5.73 | 2.71 |
| $x_4$ | 6.88 | 3.19 |

→ Euclidean distance

S4:

$$w_{11} = \left(\frac{1}{dist(x_1, c_1)}\right) \Big/ \frac{1}{dist(x_1, c_1)} + \frac{1}{dist(x_1, c_2)}$$

$$= \frac{1/2.54}{1/2.54 + 1/6.03} = \frac{.39}{.56} = 0.7$$

Date

$$w_{12} = \left(\frac{1}{dist(x_1, c_2)}\right) \Big/ \frac{1}{dist(x_1,c_1)} + \frac{1}{dist(x_1,c_2)}$$

$$= \frac{1/6.03}{1/2.54 + 1/6.03} = \frac{0.17}{0.56} = 0.3$$

$w_{21} = 0.79$ $\quad w_{22} = 0.21$

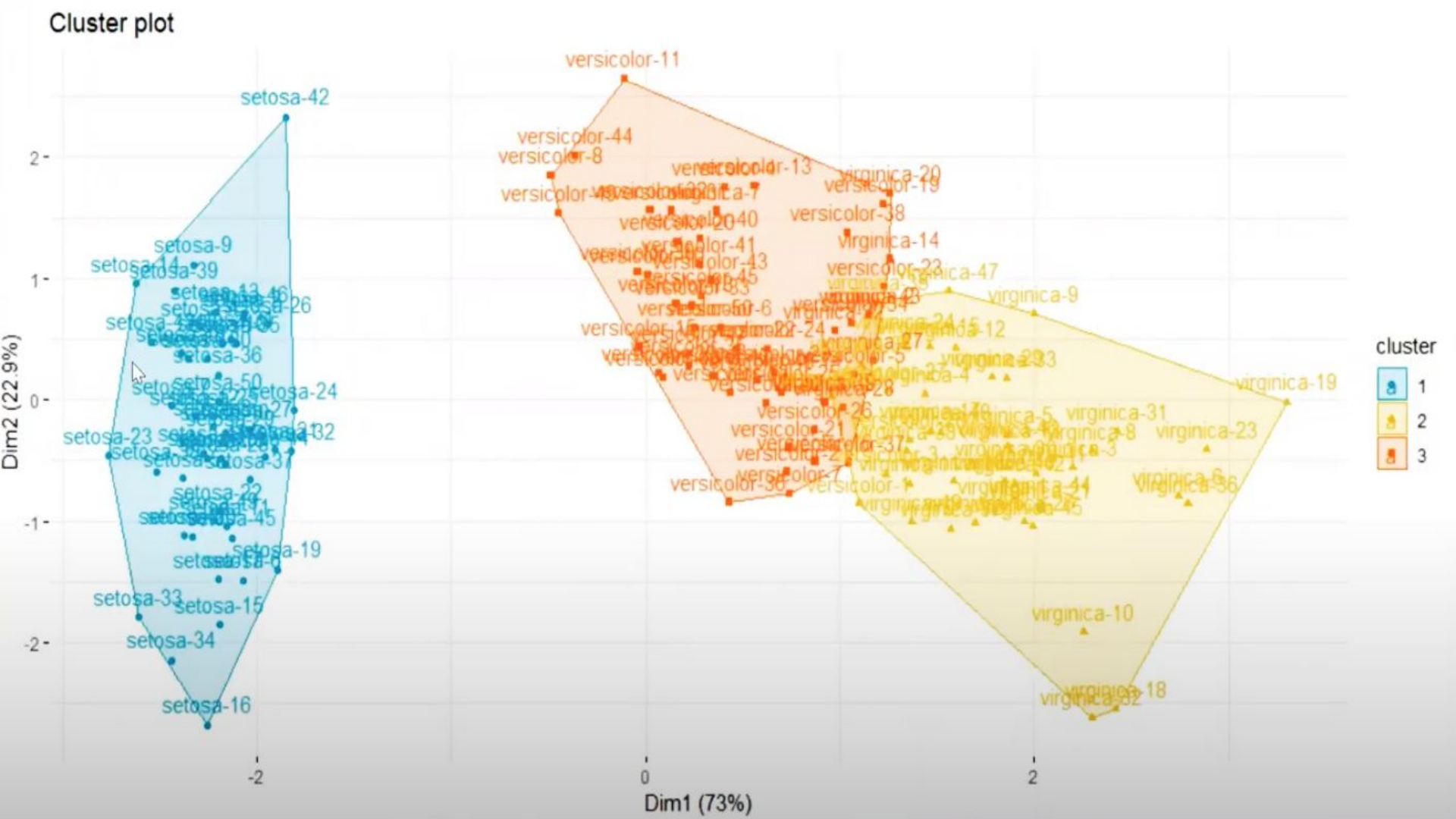
$w_{31} = 0.32$ $\quad w_{32} = 0.68$

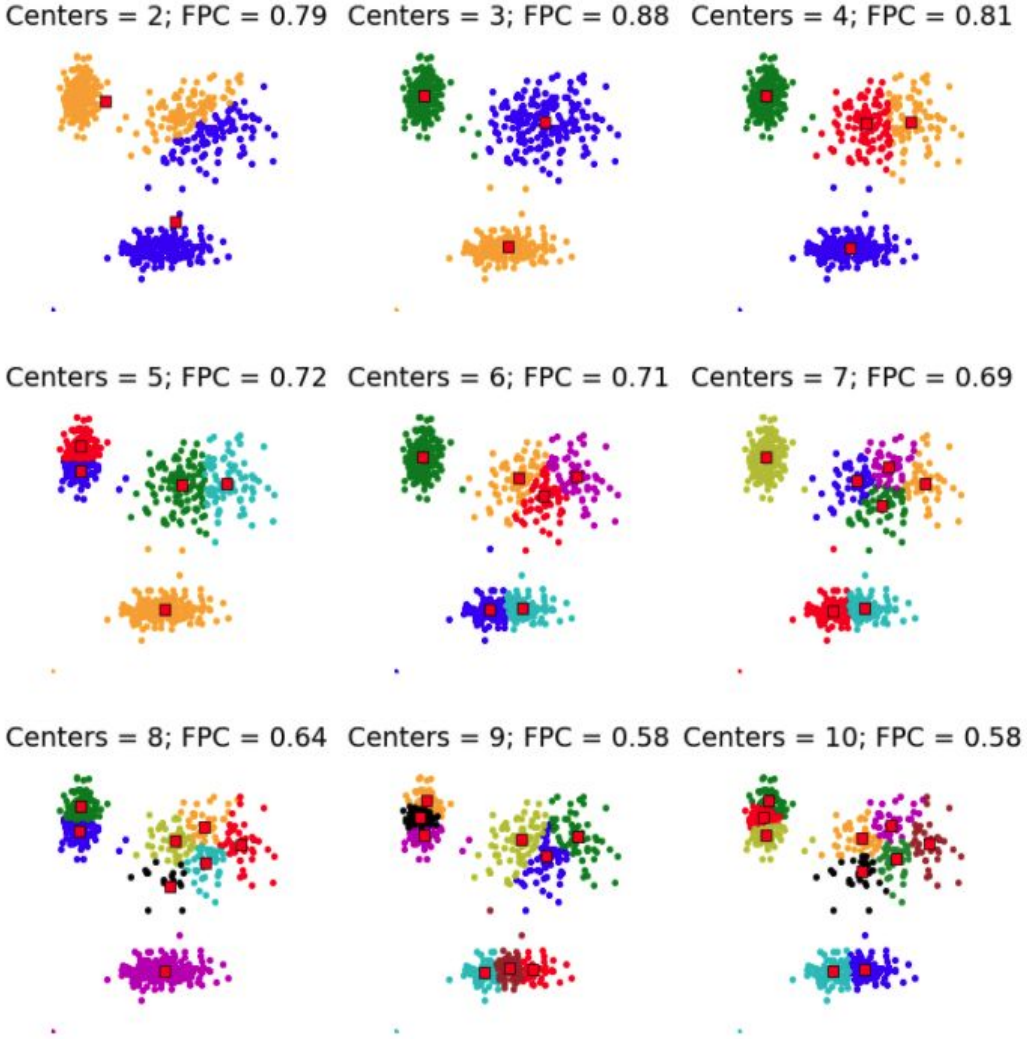
$w_{41} = 0.32$ $\quad w_{42} = 0.68$

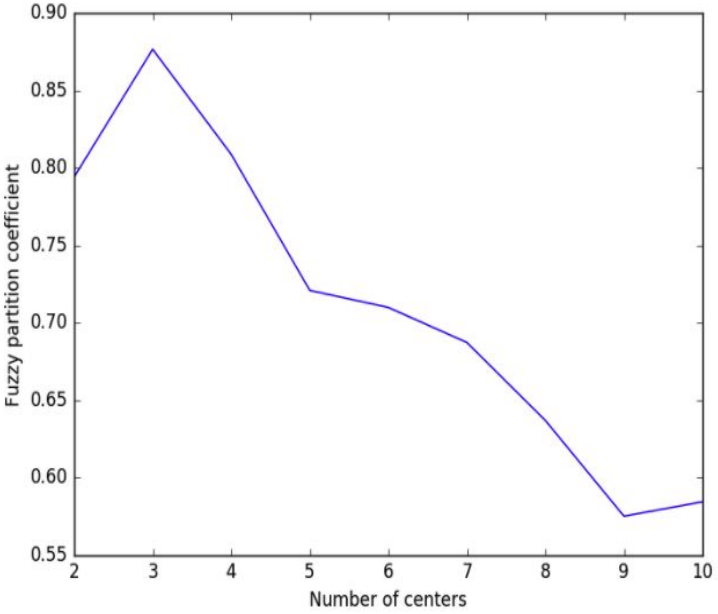
| | $C_1$ | $C_2$ | |
|---|---|---|---|
| $x_1$ | 0.7 | .3 | = 1 |
| $x_2$ | 0.79 | .21 | = 1 |
| $x_3$ | .32 | .68 | = 1 |
| $x_4$ | .32 | .68 | = 1 |

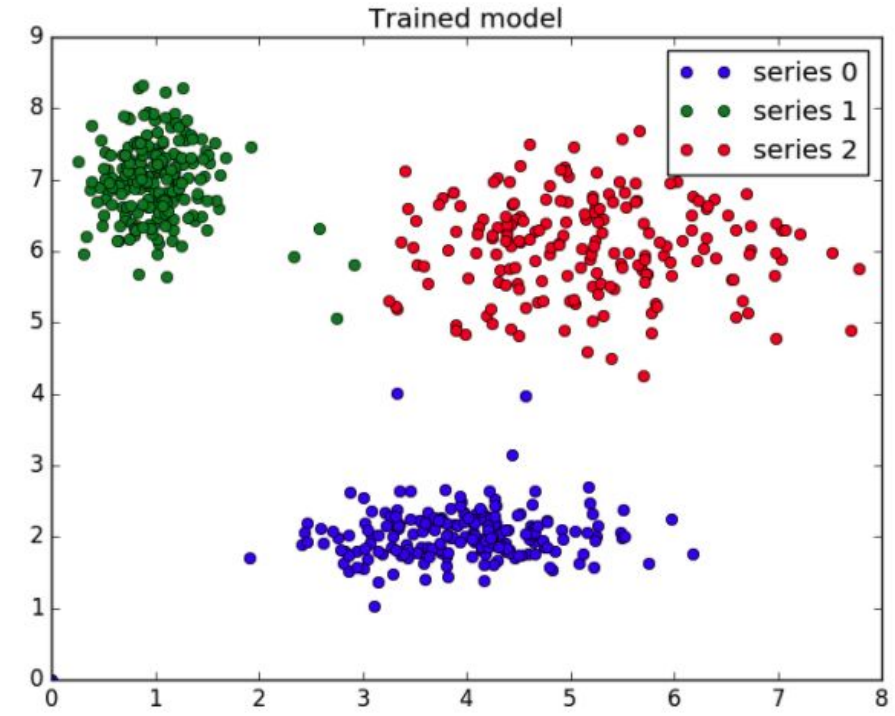
After four itr,

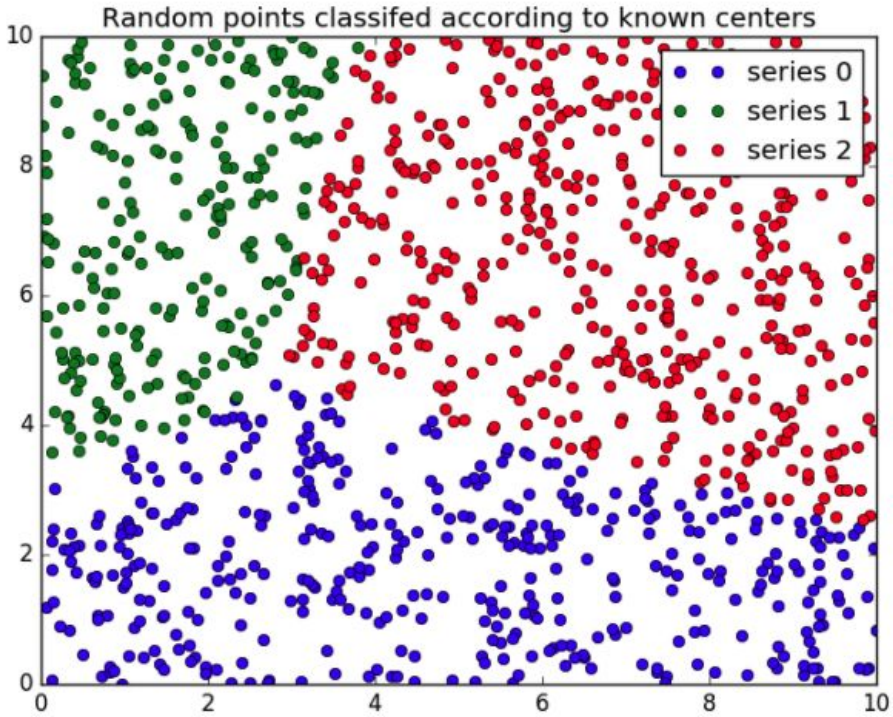
| | $C_1$ | $C_2$ | cluster no |
|---|---|---|---|
| $x_1$ | 0.88 | 0.12 | → $C_1$ |
| $x_2$ | 0.94 | 0.06 | → $C_1$ |
| $x_3$ | 0.17 | 0.83 | → $C_2$ |
| $x_4$ | 0.18 | 0.82 | → $C_2$ |


Cluster plot
Dim2 (22.9%)
Dim1 (73%)
cluster
1
2
3
setosa-42
setosa-9
setosa-39
setosa-36
setosa-24
setosa-23
setosa-37
setosa-22
setosa-19
setosa-33
setosa-15
setosa-34
setosa-16
versicolor-11
versicolor-44
versicolor-8
versicolor-38
versicolor-14
versicolor-23
virginica-9
virginica-19
virginica-31
virginica-23
virginica-10
virginica-18


Centers = 2; FPC = 0.79
Centers = 3; FPC = 0.88
Centers = 4; FPC = 0.81
Centers = 5; FPC = 0.72
Centers = 6; FPC = 0.71
Centers = 7; FPC = 0.69
Centers = 8; FPC = 0.64
Centers = 9; FPC = 0.58
Centers = 10; FPC = 0.58


0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Fuzzy partition coefficient
2
3
4
5
6
7
8
9
10
Number of centers


Trained model
series 0
series 1
series 2


Random points classifed according to known centers
series 0
series 1
series 2

## Advantages

1) Gives best result for overlapped data set and comparatively better then k-means algorithm.

2) Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.

## Disadvantages

1) Apriori specification of the number of clusters.

2) We get the better result but at the expense of  more number of iteration.

3) Euclidean distance measures can unequally weight underlying factors.

Thanks!

# Any questions?