Clustering

Objectives

After completing this module, you should be able to:

Describe what is clustering?

Understand how K Means clustering works.

Find out patterns using K Means Clustering

Understand the concept of Fuzzy C Means Clustering

Clustering Analysis

How many clusters do you expect?

3

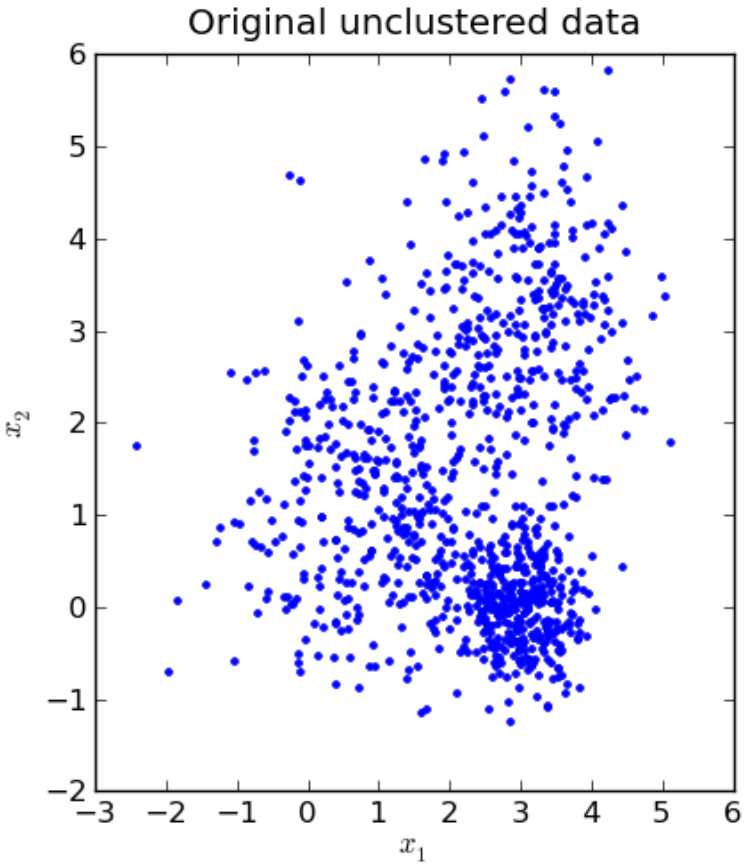
4

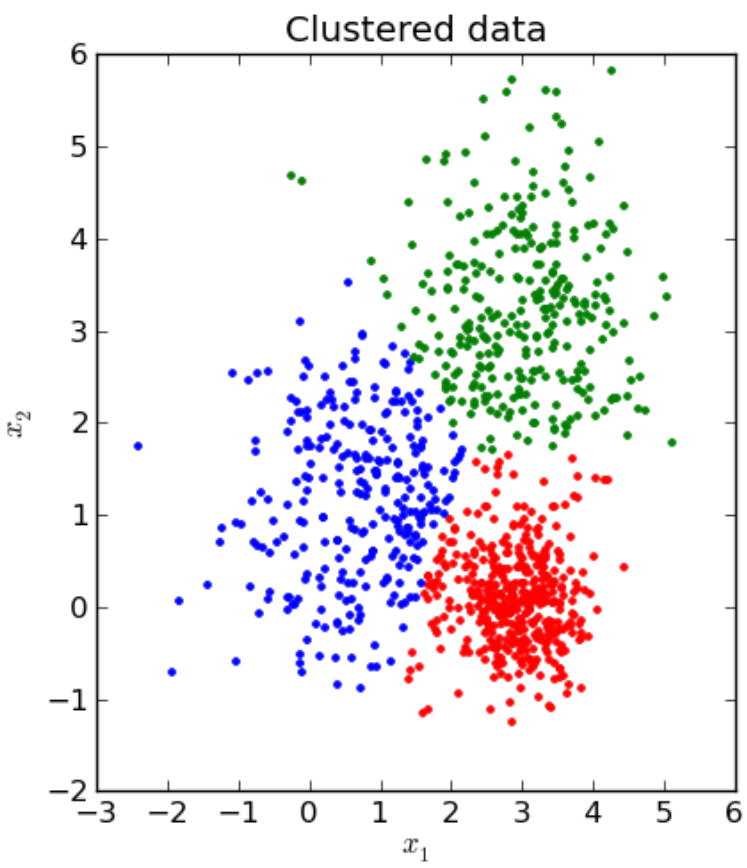
Clustering

Clustering

Clustering is the categorisation of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.

6





7

Hierarchical clustering

Agglomerative (Bottom-up)

Compute all pair-wise pattern-pattern similarity coefficients

Place each of *n* patterns into a class of its own

Merge the two most similar clusters into one

Replace the two clusters into the new cluster

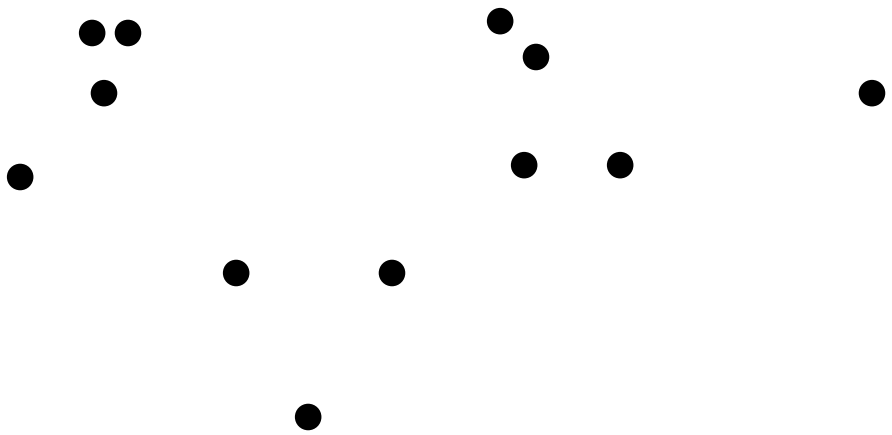
Re-compute inter-cluster similarity scores w.r.t. the new cluster

Repeat the above step until there are *k* clusters left (*k* can be 1)

8

Hierarchical clustering

Agglomerative (Bottom up)

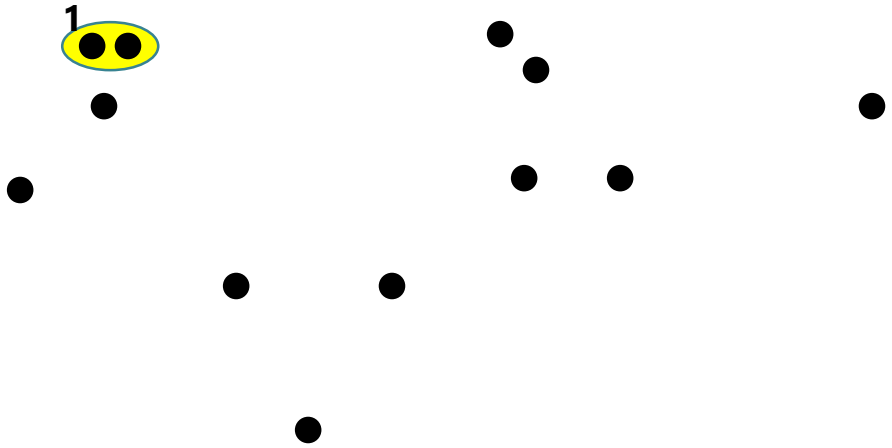


9

Hierarchical clustering

Agglomerative (Bottom up)

1st iteration

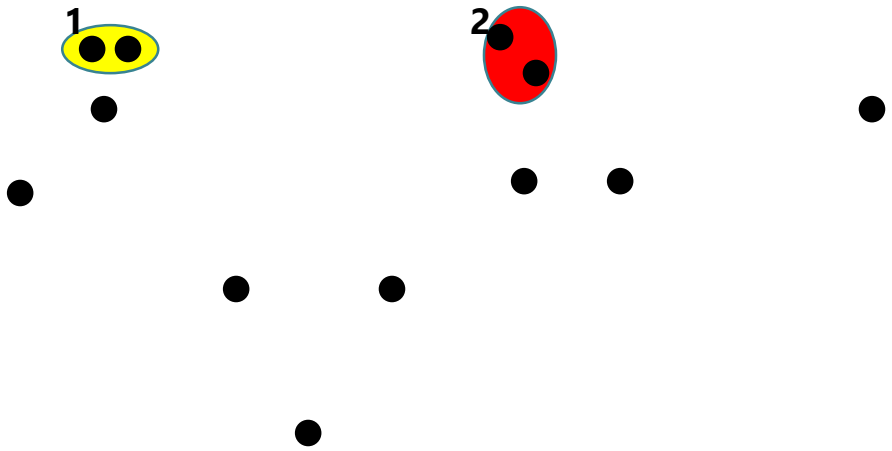


10

Hierarchical clustering

Agglomerative (Bottom up)

2nd iteration

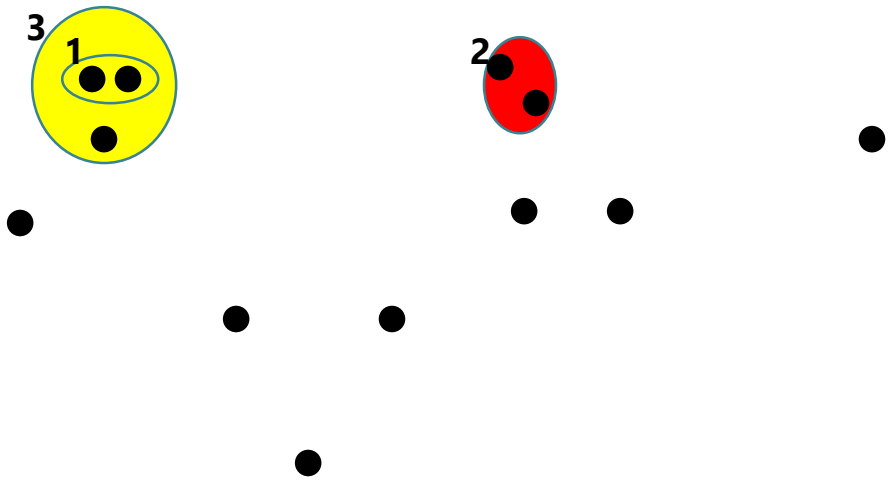


11

Hierarchical clustering

Agglomerative (Bottom up)

3rd iteration

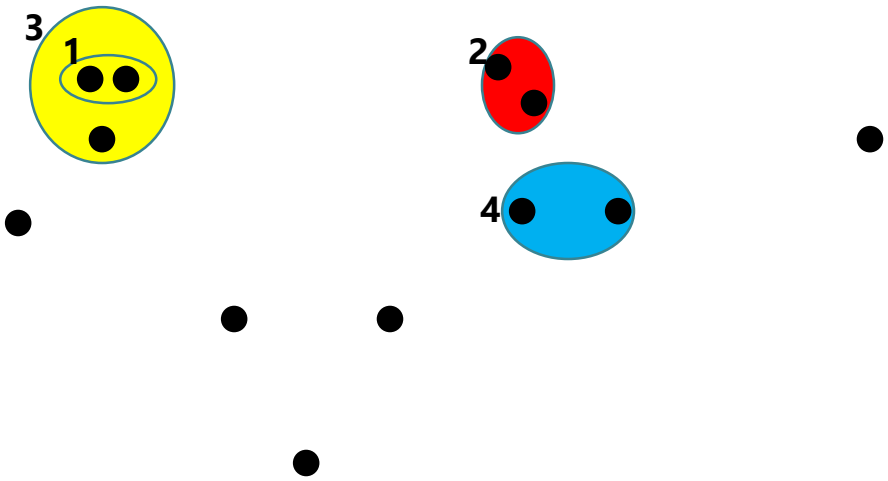


12

Hierarchical clustering

Agglomerative (Bottom up)

4th iteration

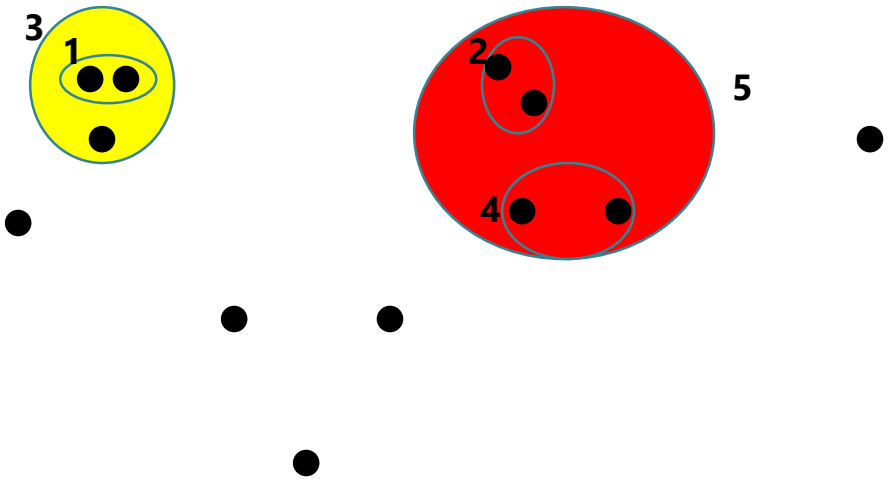


13

Hierarchical clustering

Agglomerative (Bottom up)

5th iteration

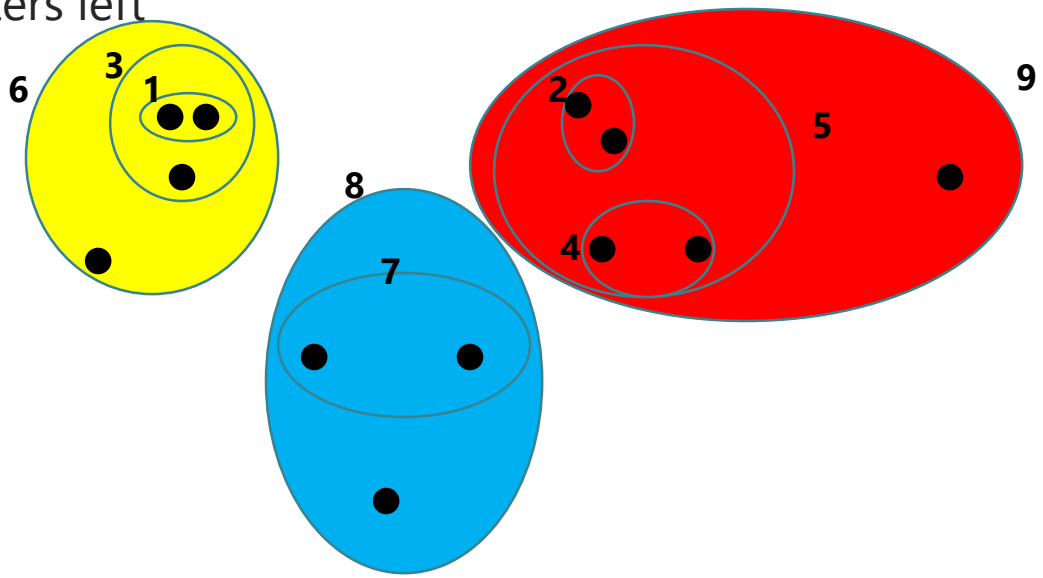


14

Hierarchical clustering

Agglomerative (Bottom up)

Finally k clusters left



15

Hierarchical clustering

Divisive (Top-down)

Start at the top with all patterns in one cluster

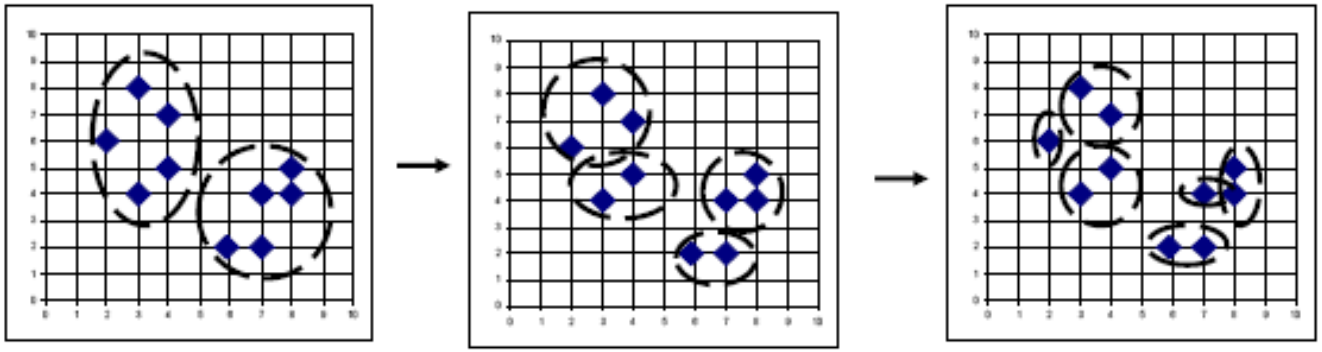
The cluster is split using a flat clustering algorithm

This procedure is applied recursively until each pattern is in its own singleton cluster

16

Hierarchical clustering

Divisive (Top-down)



17

K Means Clustering

Choose k initial centroids (center points).

Each cluster is associated with a centroid.

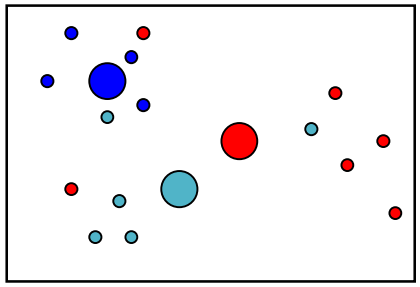
Each data object is assigned to closet centroid.

The centroid of each cluster is then updated based on the data objects assignment to the cluster.

Repeat the assignment and update steps until convergence.

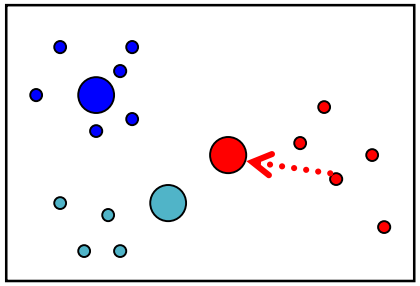
18

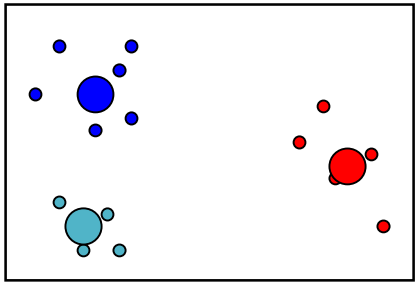
K-means: Example, k = 3



Step 1: Make random assignments and compute centroids (big dots)

Step 2: Assign points to nearest centroids

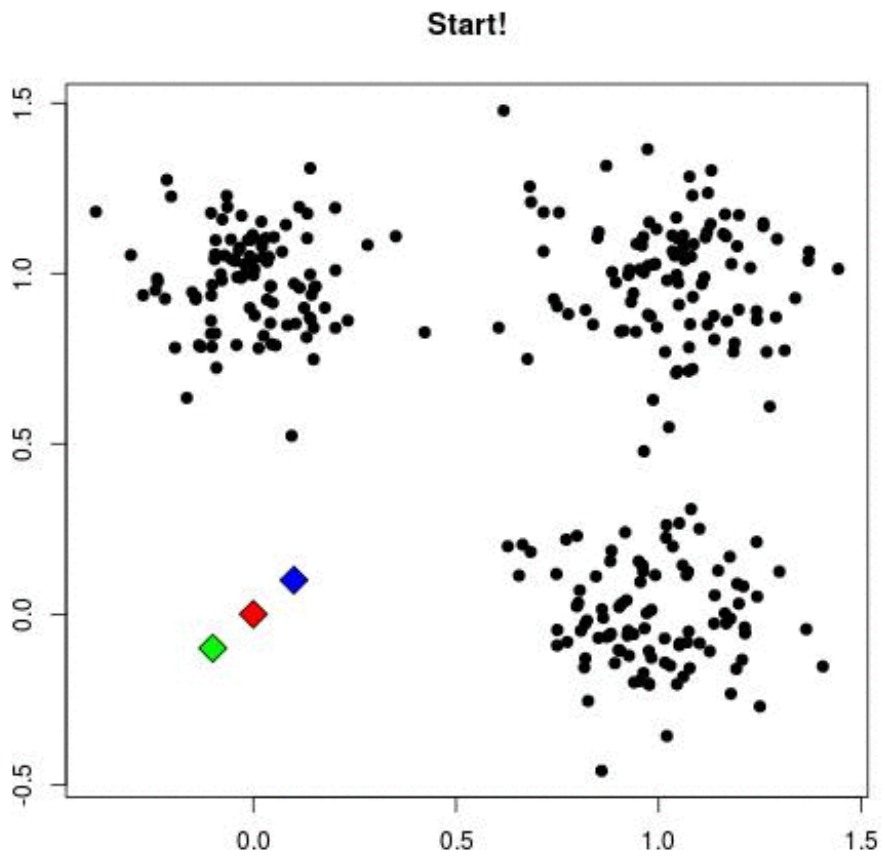




Step 3: Re-compute centroids (in this example, solution is now stable)

19

K Means Clustering



Fuzzy c-means

An extension of k-means

Hierarchical, k-means generates partitions

each data point can only be assigned in one cluster

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

It is frequently used in pattern recognition.

21

K-means Algorithm

For a given cluster assignment C of the data points, compute the cluster means m_k:

$$m_k = \frac{\sum_{i:C(i)=k} x_i}{N_k},\ k = 1, \ldots, K.$$

For a current set of cluster means, assign each observation as:

$$C(i) = \arg\min_{1 \le k \le K} \left\| x_i - m_k \right\|^2,\ i = 1, \ldots, N$$

Iterate above two steps until convergence

22

K Means - Example

Implementation of K-means algorithm (k=2)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

24

Step 1 Initialization

Randomly we choose 2 centroids (k=2) for two clusters.

In this case the two centroids are m1=(1.0,1.0) and m2=(5.0,7.0)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

25

Step 2/1 Finding the nearest centroid for every element

For every element calculating its distance from center (Euclidian Distance)

Individual	(v1,v2)	C1	C2
1	(1.0,1.0)	0	7.12
2	(1.5,2.0)	1.12	6.10
3	(3.0,4.0)	3.61	3.61
4	(5.0,7.0)	7.21	0
5	(3.5,5.0)	4.72	2.5
6	(4.5,5.0)	5.31	2.06
7	(3.5,4.5)	4.30	2.92

26

Step 2/2 Assigning elements to any of the clusters

Thus we obtain 2 clusters containing {1,2,3} and {4,5,6,7}

New centroids are

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12, 5.38)$$

Individual	(v1,v2)	C1	C2
1	(1.0,1.0)	0	7.12
2	(1.5,2.0)	1.12	6.10
3	(3.0,4.0)	3.61	3.61
4	(5.0,7.0)	7.21	0
5	(3.5,5.0)	4.72	2.5
6	(4.5,5.0)	5.31	2.06
7	(3.5,4.5)	4.30	2.92

27

Step 3 Assigning elements to new clusters according to distance

Now using these centroids m1=(1.83,2.33) & m2=(4.12,5.38) we compute the Euclidean distance of each object, as shown in table.

Individual	(v1,v2)	C1	C2
1	(1.0,1.0)	1.57	5.38
2	(1.5,2.0)	0.47	4.28
3	(3.0,4.0)	2.04	1.78
4	(5.0,7.0)	5.64	1.84
5	(3.5,5.0)	3.15	0.73
6	(4.5,5.0)	3.78	0.54
7	(3.5,4.5)	2.74	1.08

28

Step 3 Assigning elements to new clusters according to distance

Therefore, new clusters are:

{1,2} and {**3**,4,5,6,7}

Next centroids are: m1=(1.25,1.5) and m2 = (3.9,5.1)

Individual	(v1,v2)	C1	C2
1	(1.0,1.0)	1.57	5.38
2	(1.5,2.0)	0.47	4.28
3	(3.0,4.0)	2.04	1.78
4	(5.0,7.0)	5.64	1.84
5	(3.5,5.0)	3.15	0.73
6	(4.5,5.0)	3.78	0.54
7	(3.5,4.5)	2.74	1.08

29

Step 4

Now using these centroids m1=(1.25,1.5) & m2 = (3.9,5.1) we compute the Euclidean distance of each object, as shown in table.

Therefore, new clusters are:

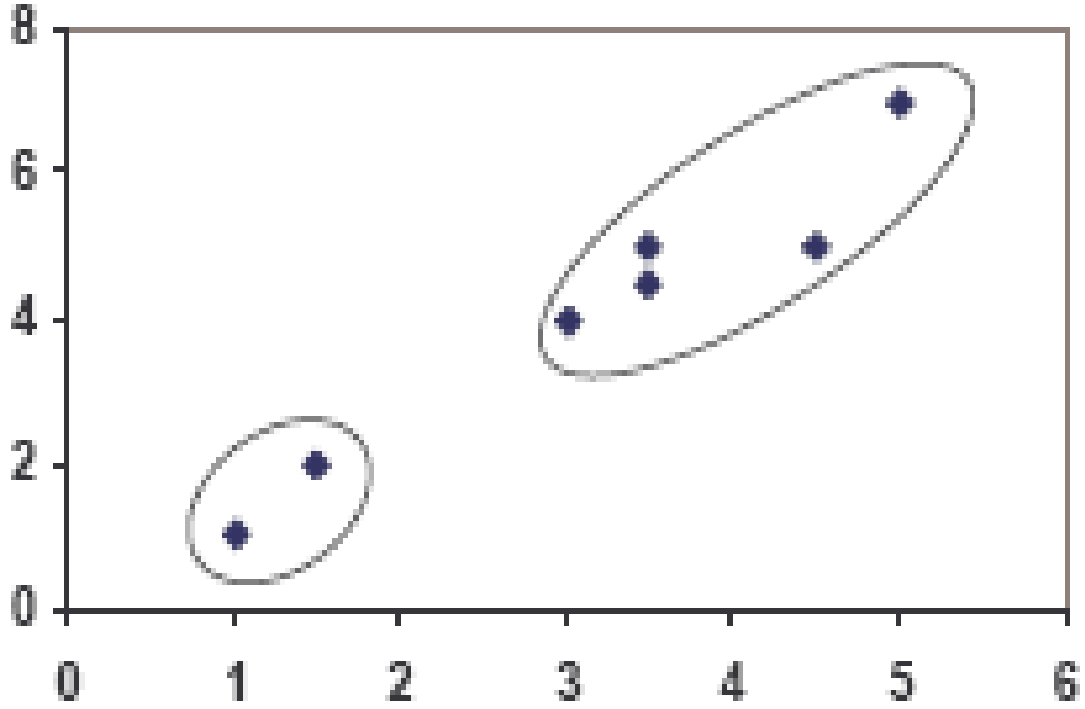
{1,2} and {**3**,4,5,6,7}

Therefore, there is no change in the cluster.

Individual	(v1,v2)	C1	C2
1	(1.0,1.0)	0.56	5.02
2	(1.5,2.0)	0.56	3.92
3	(3.0,4.0)	3.05	1.42
4	(5.0,7.0)	6.66	2.20
5	(3.5,5.0)	4.16	0.41
6	(4.5,5.0)	4.78	0.61
7	(3.5,4.5)	3.75	0.72

Thus, algorithm comes to a halt here and final result consist of 2 clusters {1,2} and {3,4,5,6,7}.

30



31

How to choose K?

Use another clustering method, like EM.

Run algorithm on data with several different values of K.

Use the prior knowledge about the characteristics of the problem.

32

Applications of K-Mean Clustering

It is relatively efficient and fast. It computes result at **O(tkn),** where n is number of objects or points, k is number of clusters and t is number of iterations.

k-means clustering can be applied to machine learning or data mining

Used on acoustic data in speech understanding to convert waveforms into one of k categories (known as Vector Quantization or Image Segmentation).

Also used for choosing color palettes on old fashioned graphical display devices and Image Quantization.

33

CONCLUSION

K-means algorithm is useful for undirected knowledge discovery and is relatively simple.

K-means has found wide spread usage in lot of fields, ranging from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligence, image processing, machine vision, and many others.

34

Applications of Clustering

Data Mining

Pattern recognition

Image analysis

Bioinformatics

Voice mining

Image processing

Text mining

Web cluster engines

Weather report analysis

35

Summary

This module covered the following topics:

Clustering Analysis

Hierarchical Clustering

K Means Clustering

Fuzzy C Means Clustering

Applications of Clustering

Thank you