Clustering

Sistem Informasi | Fakultas Ilmu Komputer | UDINUS
2019

Clustering

- **Clustering** is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters (Bramer, 2007)

- Cluster analysis is the best-known descriptive data mining method. Given a data matrix composed of *n* observations (rows) and *p* variables(columns), the objective of cluster analysis is to cluster the observations into groups that are internally homogeneous (internal cohesion) and heterogeneous from group to group (external separation). (Guidici, 2009)

- By ***clustering*** we mean the method to divide a set of data (records/tuples/vectors/instances/objects/sample) into several groups (clusters), based on certain predetermined similarities. (Goronescu,2011)

Clustering

- **Cluster** : kumpulan data yang mirip antara satu dengan yang lain, dan memiliki perbedaan bila dibandingkan dengan data dari klaster lain

- Perbedaan utama algoritma klastering dengan klasifikasi adalah **klastering tidak memiliki target/class/label**, jadi termasuk *unsupervised learning*

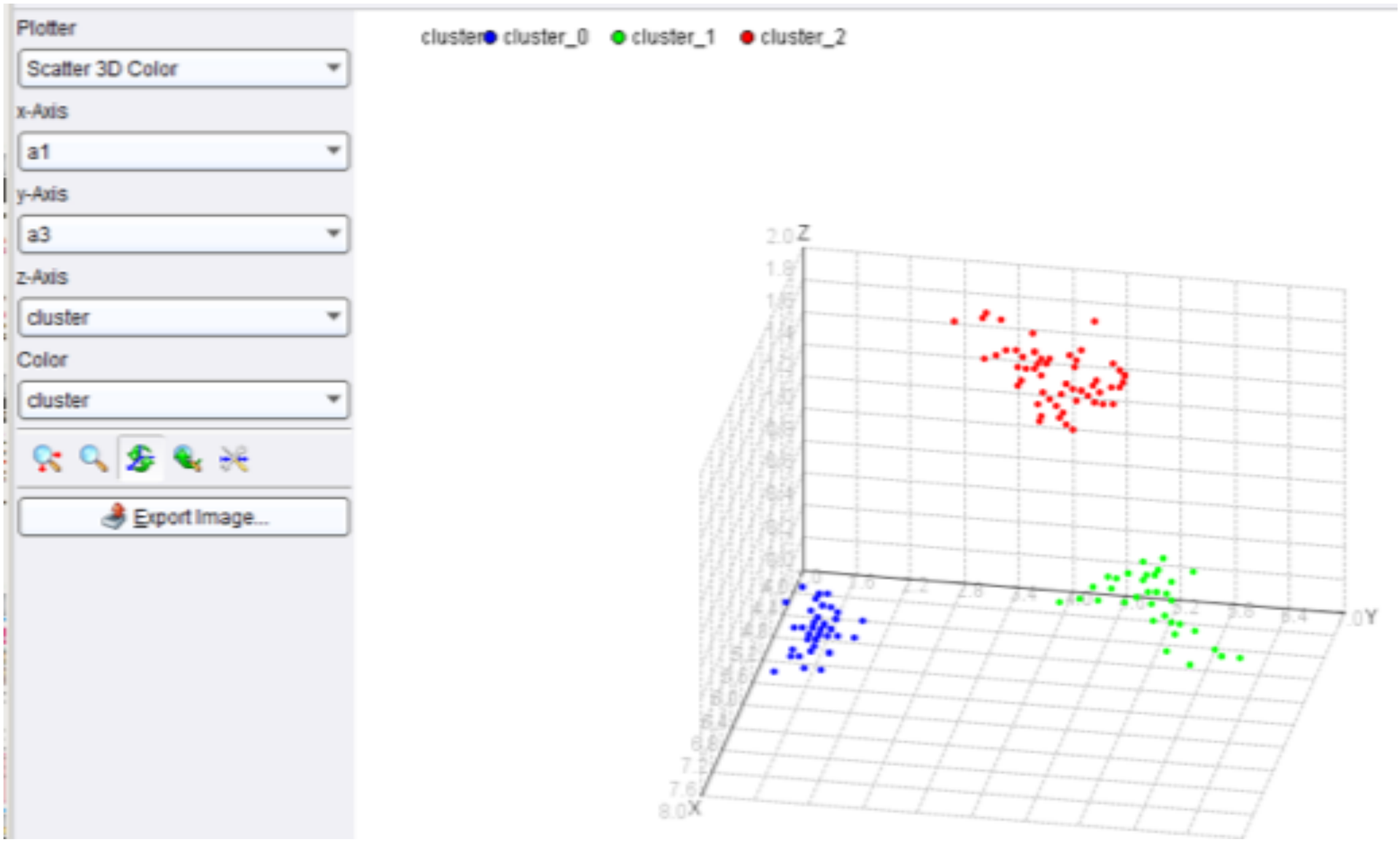
- Klastering sering digunakan sebagai tahap awal dalam proses data mining, dengan hasil klaster yang terbentuk akan menjadi input dari algoritma berikutnya yang digunakan

Contoh: Klastering Bunga Iris

ExampleSet (150 examples, 2 special attributes, 4 regular attributes)

Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	4.800	3	1.400	0.100
14	id_14	Iris-setosa	4.300	3	1.100	0.100
15	id_15	Iris-setosa	5.800	4	1.200	0.200
16	id_16	Iris-setosa	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	5.100	3.800	1.500	0.300
21	id_21	Iris-setosa	5.400	3.400	1.700	0.200
22	id_22	Iris-setosa	5.100	3.700	1.500	0.400
23	id_23	Iris-setosa	4.600	3.600	1	0.200
24	id_24	Iris-setosa	5.100	3.300	1.700	0.500

Contoh: Klastering Bunga Iris



Contoh: Klastering Bunga Iris

ExampleSet (150 examples, 3 special attributes, 4 regular attributes) View

Row No.	id	label	cluster	a1	a2	a3	a4
1	id_1	Iris-setosa	cluster_0	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	cluster_0	4.900	3	1.400	0.200
3	id_3	Iris-setosa	cluster_0	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	cluster_0	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	cluster_0	5	3.600	1.400	0.200
6	id_6	Iris-setosa	cluster_0	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	cluster_0	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	cluster_0	5	3.400	1.500	0.200
9	id_9	Iris-setosa	cluster_0	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	cluster_0	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	cluster_0	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	cluster_0	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	cluster_0	4.800	3	1.400	0.100
14	id_14	Iris-setosa	cluster_0	4.300	3	1.100	0.100
15	id_15	Iris-setosa	cluster_0	5.800	4	1.200	0.200
16	id_16	Iris-setosa	cluster_0	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	cluster_0	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	cluster_0	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	cluster_0	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	cluster_0	5.100	3.800	1.500	0.300

Cluster Model

```
Cluster 0: 50 items
Cluster 1: 39 items
Cluster 2: 61 items
Total number of items: 150
```

K-MEANS ALGORITHm

- Algoritma Clustering meletakkan nilai yang serupa dalam satu segmen, dan meletakkan nilai yang berbeda dalam segmen yang berbeda (Wu & Kumar, 2009).

- Algoritma *K-means diterapkan pada objek yang diwakili dalam bentuk* **titik** didalam ruangan **vektor** berdimensi-d. *K-means mengcluster semua data* didalam setiap dimensi dimana titik dalam segmentasi yang sama diberi custer ID.

- Nilai dari ***k*** adalah masukan dasar dari algoritma yang menentukan **jumlah segmentasi** yang ingin dibentuk. Partisi akan dibentuk dari sekumpulan objek ***n*** ke dalam cluster k sehingga terbentuk kesamaan objek dalam setiap segmentasi k.

Step by Step K-Means

1. Menentukan banyaknya cluster (***k***)
2. Menentukan ***titik pusat*** tiap cluster
3. Hitung jarak antara titik pusat dengan objek menggunakan ***euclidean distance***
4. Mengelompokkan objek berdasarkan jarak yang terdekat.
5. Apakah titik pusat berubah?
    a. Jika berubah, hitung jarak objek ke titik pusat
    b. Jika tidak berubah, selesai.

EUCLIDEAN DISTANCE

- **Euclidean Distance** adalah metrika yang paling sering digunakan untuk menghitung kesamaan 2 vektor.

- Euclidean distance menghitung akar dari kuadrat perbedaan 2 vektor (root of square differences between 2 vectors).

- Rumus dari Euclidean Distance:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

EUCLIDEAN DISTANCE

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

- Contoh
  Terdapat 2 vektor ciri berikut :

$$A = [0, 3, 4, 5]$$
$$B = [7, 6, 3, -1]$$

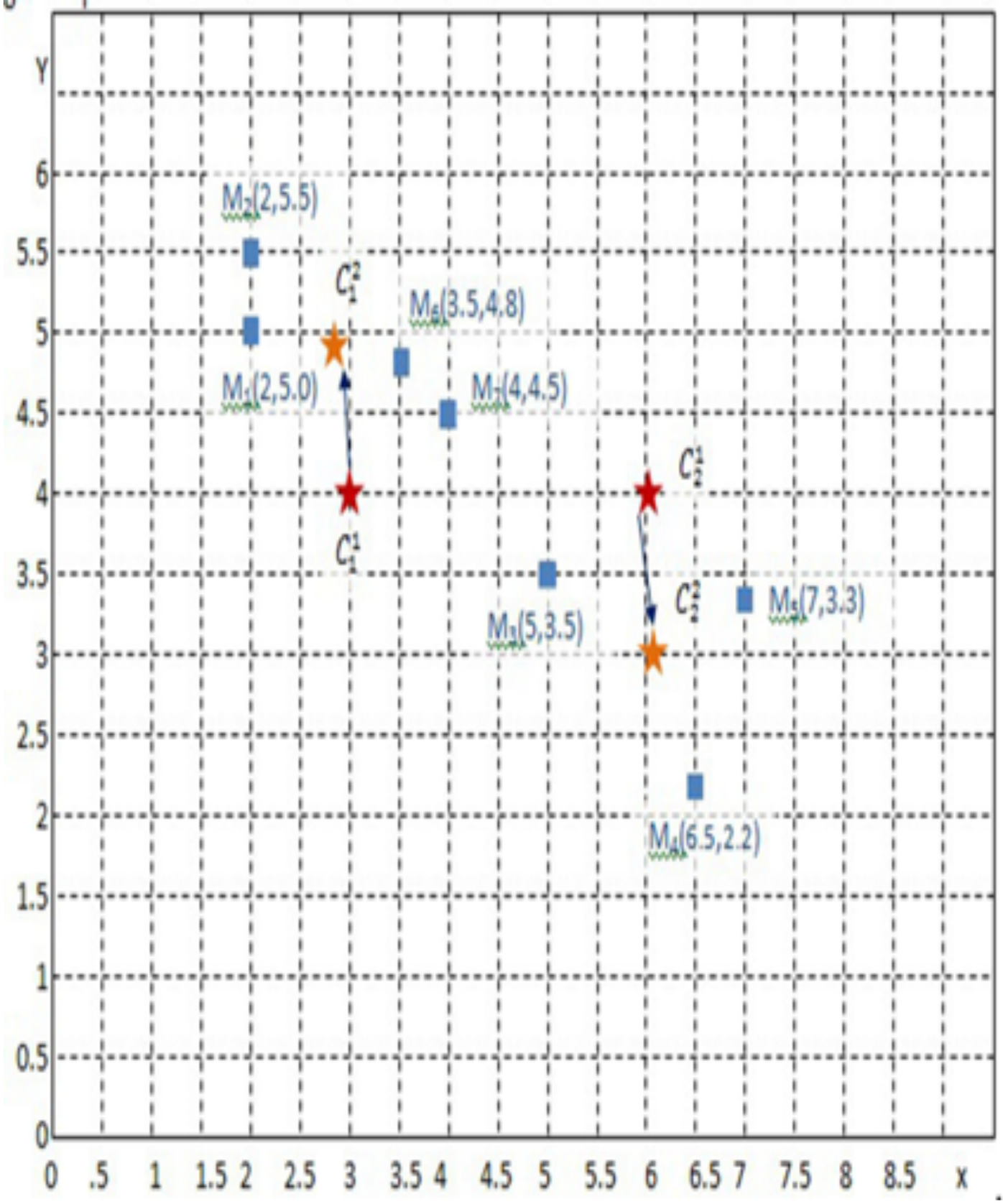
- Euclidean Distance dari vektor A dan B

$$d_{AB} = \sqrt{(0-7)^2 + (3-6)^2 + (4-3)^2 + (5-(-1))^2}$$
$$= \sqrt{49+9+1+36} = 9.747$$

Contoh Kasus

- Dengan menggunakan algoritma k-means, temukan cluster yang terbentuk dari objek data 2D berikut ini.

  M1 = (2, 5.0),
  M2 = (2, 5.5),
  M3 = (5, 3.5),
  M4 = (6.5, 2.2),
  M5 = (7, 3.3),
  M6 = (3.5, 4.8),
  M7 = (4, 4.5)



Kondisi awal :

Semua data akan dikelompokkan ke dalam 2 cluster

Titik pusat dari cluster : $C_1(3,4)$, $C_2(6,4)$

Iterasi 1

- Menghitung *Euclidean distance* dari semua data ke tiap titik pusat pertama ***C1***

M1 = (2, 5.0),
M2 = (2, 5.5),
M3 = (5, 3.5),
M4 = (6.5, 2.2),
M5 = (7, 3.3),
M6 = (3.5, 4.8),
M7 = (4, 4.5)

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

- Titik pusat untuk kedua cluster yaitu **C_1(3,4), C_2(6,4)**
- **Hitung D1 dan D2**

Iterasi 1-1

- Sehingga didapatkan :

  D11 = 1.41,
  D12 = 1.80,
  D13 = 2.06,
  D14 = 3.94,
  D15 = 4.06,
  D16 = 0.94,
  D17 = 1.12,

$$D_{11} = \sqrt{(M_{1x} - C_{1x})^2 + (M_{1y} - C_{1y})^2} = \sqrt{(2-3)^2 + (5-4)^2} = \sqrt{2} = 1.41$$

$$D_{12} = \sqrt{(M_{2x} - C_{1x})^2 + (M_{2y} - C_{1y})^2} = \sqrt{(2-3)^2 + (5.5-4)^2} = \sqrt{3.25} = 1.80$$

$$D_{13} = \sqrt{(M_{3x} - C_{1x})^2 + (M_{3y} - C_{1y})^2} = \sqrt{(5-3)^2 + (3.5-4)^2} = \sqrt{4.25} = 2.06$$

$$D_{14} = \sqrt{(M_{4x} - C_{1x})^2 + (M_{1y} - C_{1y})^2} = \sqrt{(6.5-3)^2 + (2.2-4)^2} = \sqrt{2} = 3.94$$

$$D_{15} = \sqrt{(M_{5x} - C_{1x})^2 + (M_{5y} - C_{1y})^2} = \sqrt{(7-3)^2 + (3.3-4)^2} = \sqrt{2} = 4.06$$

$$D_{16} = \sqrt{(M_{6x} - C_{1x})^2 + (M_{6y} - C_{1y})^2} = \sqrt{(3.5-3)^2 + (4.8-4)^2} = \sqrt{2} = 0.94$$

$$D_{17} = \sqrt{(M_{7x} - C_{1x})^2 + (M_{7y} - C_{1y})^2} = \sqrt{(4-3)^2 + (4.5-4)^2} = \sqrt{2} = 1.12$$

- Dengan cara yang sama hitung jarak tiap titik ke titik pusat kedua, dan kita akan mendapatkan :

  D21 = 4.12, D22 = 4.27, D23 = 1.18, D24= 1.86,
  D25 =1.22, D26 = 2.62, D27 = 2.06

Iterasi 1-2

- Dari penghitungan *Euclidean distance*, kita dapat membandingkan :

		M1	M2	M3	M4	M5	M6	M7
Jarak ke	C1	**1.41**	**1.80**	2.06	3.94	4.06	**0.94**	**1.12**
	C2	4.12	4.27	**1.18**	**1.86**	**1.22**	2.62	2.06

Anggota C1 : {M1, M2, M6, M7}

Anggota C2 : {M3, M4, M5}

Iterasi 1-3

M1 = (2, 5.0),
M2 = (2, 5.5),
M3 = (5, 3.5),
M4 = (6.5, 2.2),
M5 = (7, 3.3),
M6 = (3.5, 4.8),
M7 = (4, 4.5)

Anggota C_1 : {M1, M2, M6, M7} —> 4

Anggota C_2 : {M3, M4, M5} —> 3

Hitung titik pusat baru, dengan menggunakan hasil anggota cluster yang diperoleh **(*yaitu C baru*)**

$$\mathbf{C1} = \left(\frac{M1+M2+M6+M7}{4}\right)$$

$$\left(\frac{2+2+3.5+4}{4}, \frac{5+5.5+4.8+4.5}{4}\right) = \mathbf{(2.85, 4.95)}$$

$$\mathbf{C2} = \left(\frac{M3+M4+M5}{3}\right)$$

$$\left(\frac{5+6.5+7}{3}, \frac{3.5+2.2+3.3}{3}\right) = \mathbf{(6.17, 3)}$$

Iterasi 2

- Hitung Euclidean distance (D11, D12, dst..) dari tiap data ke titik pusat yang baru, Dengan cara yang sama seperti iterasi 1

M1 = (2, 5.0),
M2 = (2, 5.5),
M3 = (5, 3.5),
M4 = (6.5, 2.2),
M5 = (7, 3.3),
M6 = (3.5, 4.8),
M7 = (4, 4.5)

Gunakan :

- Titik pusat dari cluster : C_1(2.85, 4.95), C_2(6.17, 3)
- Hitung D1 dan D2 (iterasi 2)

Iterasi 2-1

- kita akan mendapatkan perbandingan sebagai berikut :

		M1	M2	M3	M4	M5	M6	M7
Jarak ke	C1	**0.76**	**0.96**	2.65	4.62	4.54	**0.76**	**1.31**
	C2	4.62	4.86	**1.27**	**0.86**	**0.88**	3.22	2.63

Iterasi 2-2

- Dari perbandingan tersebut diketahui bahwa

  C_1 baru = {M_1, M_2, M_6, M_7}

  C_2 baru = {M_3, M_4, M_5}

- Karena anggota kelompok tidak ada yang berubah maka titik pusat pun tidak akan berubah.

- Dan perhitungan berhenti.