UNIVERSITAS AMIKOM YOGYAKARTA
Creative Economy Park

# Clustering – Konsep Dasar & Penerapan Algoritma K Means

## Data Minig

*Sumarni Adi, S.Kom., M.Cs*

# Unsupervised Learning

## Basic Concept Clustering

- Disebut juga klasterisasi (clustering) yang mampu mengelompokkan himpunan data secara otomatis.
- Jika pada supervised learning (Klasifikasi) membutuhkan label kelas maka pada Unsupervised learning tidak memerlukan label kelas.

# Contoh

| Handphone | Baterai | Kamera | Harga | Layak direkomendasikan |
|---|---|---|---|---|
| H1 | 26 | 8 | 1,2 | ? |
| H2 | 27 | 13 | 15 | ? |
| H3 | 28 | 5 | 6 | ? |
| H4 | 25 | 2 | 5 | ? |
| H5 | 23 | 10 | 1 | ? |

- **Bagaimana cara mengelompokkan data tersebut ke dalam 2 klaster sehingga Anda bisa menjawab handphone mana yang layak dan mana yang tidak layak direkomendasikan?**

# Contoh

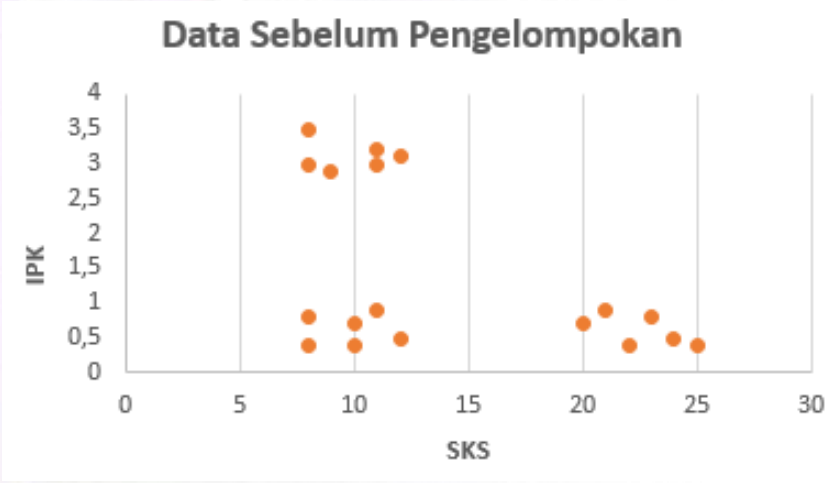
## Data disebuah perusahaan telekomunikasi

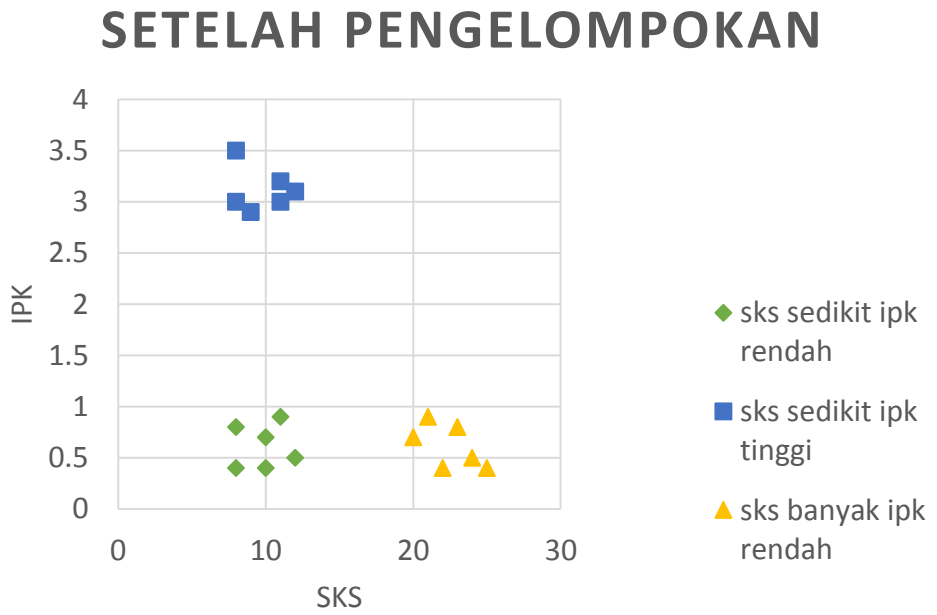
| Panggilan | Blok | Layak Dapat Bonus |
|---|---|---|
| 30 | 50 | ? |
| 40 | 140 | ? |
| 50 | 220 | ? |
| 60 | 300 | ? |

- Bagaimana cara mengelompokkan data tersebut ke dalam 2 klaster sehingga Anda bisa menjawab pelanggan mana yang layak dan mana yang tidak layak mendapatkan bonus?

# Contoh

- Catatan akademik pada sejumlah mahasiswa diketahui data jumlah sks dan nilai IPK mahasiswa.



# Classification vs Clustering

## Basic Concept Clustering

1. Classification bertujuan untuk memetakan satu titik data ke dalam satu kelas yang telah ditentukan sebelumnya
2. Classification dilakukan secara supervised, artinya algoritma pemelajaran untuk melakukan klasifikasi diberikan contoh titik data dan kelas apa seharusnya titik data tersebut dipetakan.

1. Clustering bertujuan untuk mengelompokan titik-titik data yang berdekatan dan mimisahkannya dengan kelompok-kelompok lain yang berjauhan dalam suatu ruang.
2. Clustering dilakukan secara unsupervised, artinya tidak ada contoh bagaimana seharusnya mengelompokan titik-titik tersebut.

## Basic Concept Clustering

**Clustering atau Klasterisasi** adalah proses pengelompokan himpunan data ke dalam beberapa group atau klaster sedemikian hingga objek-objek yang ada di dalam kluster memiliki kemiripan yang tinggi,

namun sangat berbeda (memiliki ketidakmiripan yang tinggi) dengan objek –objek di klaster-klaster lainnya (J Han et Al, 2012)

# Clustering Implementation



- Riset pasar : segmentasi profiling pelanggan untuk merancang startegi produk, harga, tempat, promosi dll
- Recomender System : Jual beli online pendekatan collaborative filtering, business intelligence
- Pencarian Informasi : Mengelompokkan hasil halaman yang diberikan mesin pencari
- ...dll

# Clustering Methods

- Partitioning methods (metode partisi)
- Hierarchical methods (metode hirarki)
- Density-based methods (metode kepadatan)
- Grid-based methods (metode berbasis kisi)

# Partitioning methods (metode partisi)

## Partitioning Methods

- Metode ini bekerja dengan cara membagi/mempartisi data kedalam sejumlah kelompok.

- Misalnya sejumlah himpunan data D berisi *n* objek. *n* objek dimasukkan kedalam *k* kluster $C_1, C_2 \ldots C_k$ tanpa ada objek yang saling tumpah tindih sehingga $C_1 \in D$ dan $C_i \cap C_j = 0$

Algoritma yang digunakan :

1. K-Means
2. K-Harmonic means
3. K-Modes
4. Fuzzy C-Means

# K-Means



- K-means merupakan algoritma klasterisasi yang paling tua dan paling banyak digunakan diberbagai aplikasi kecil dan menengah.
- Peneliti yang berpengaruh adalah Lloyd (1982), Friedman dan Rubin (1967). McQueen (1967)
- Ide dasar algoritma ini adalah meminimalkan Sum of Squared Error (SSE) antara objek-objek data dengan sejumlah *k centroid.*

# Cara Kerja K-Means



1. Dari himpunan data yang akan diklaster, tentukan jumlah *k kluster* dan pilih secara acak sebagai centroid awal sejumlah *k kluster*
2. Setiap objek yang bukan centroid dimasukkan ke dalam kluster terdekat berdasarkan ukuran jarak tertentu.
3. Setiap centroid diperbarui berdasarkan rata-rata dari objek yang ada didalam setiap kluster.
4. Langkah ke-2 dan ke-3 diulang-ulang sampai semua centroid stabil atau konvergen. Artinya semua centroid yang dihasilkan dalam iterasi saat ini sama dengan centroid sebelumnya

# k-Means Clustering

Algoritma 4.1 *k-means clustering*

*k-means*(*D*, *k*)

Pilih sejumlah *k* objek secara acak dari himpunan data *D* sebagai *centroid* awal — **Langkah 1**

**repeat**

  **for** semua objek di dalam *D* — **Langkah 2**

    Masukkan setiap objek yang bukan *centroid* ke klaster yang paling dekat di antara *k* klaster yang ada

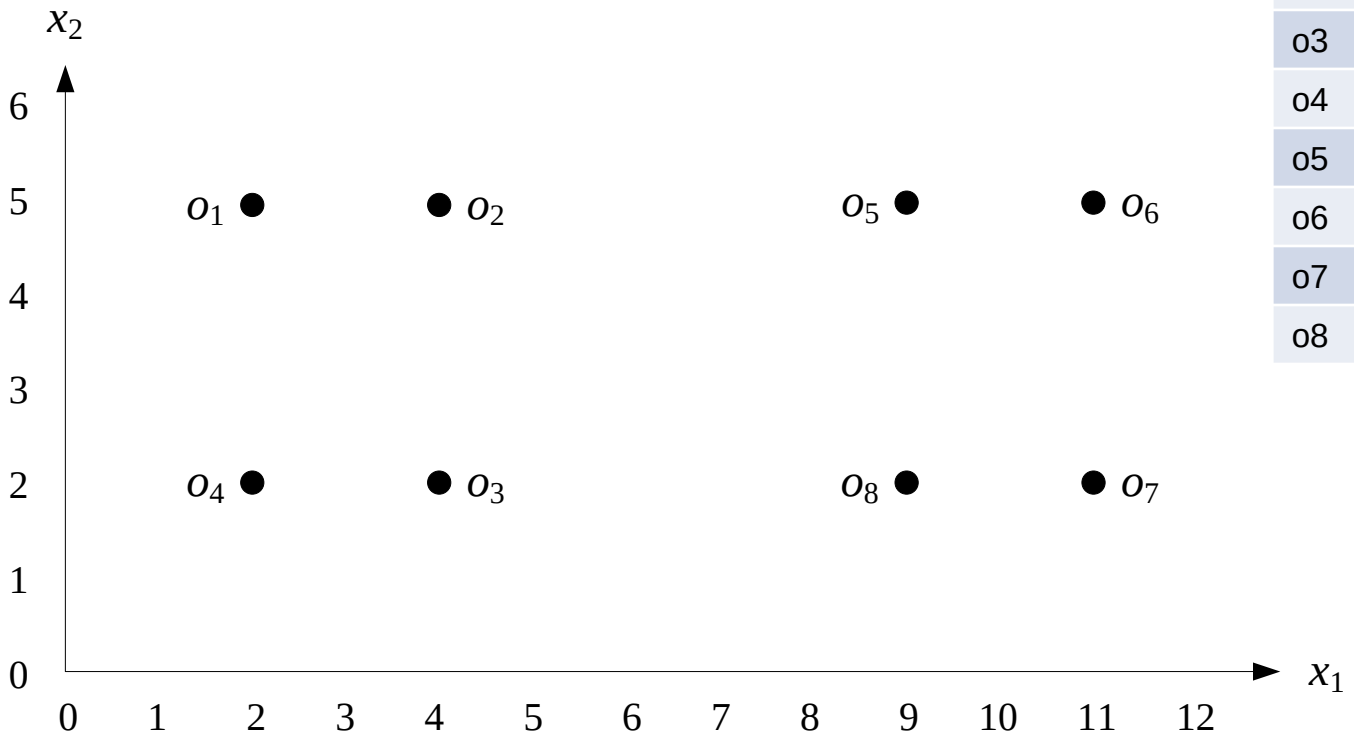
  **end** — **Langkah 3**

  Perbarui setiap *centroid* dengan menghitung rata-rata dari semua objek yang berada di dalam klaster tersebut

**until** tidak ada perubahan *centroid*

# k-Means Clustering

| Data | x1 | x2 |
|---|---|---|
| o1 | 2 | 5 |
| o2 | 4 | 5 |
| o3 | 4 | 2 |
| o4 | 2 | 2 |
| o5 | 9 | 5 |
| o6 | 11 | 5 |
| o7 | 11 | 2 |
| o8 | 9 | 2 |



# Langkah 1

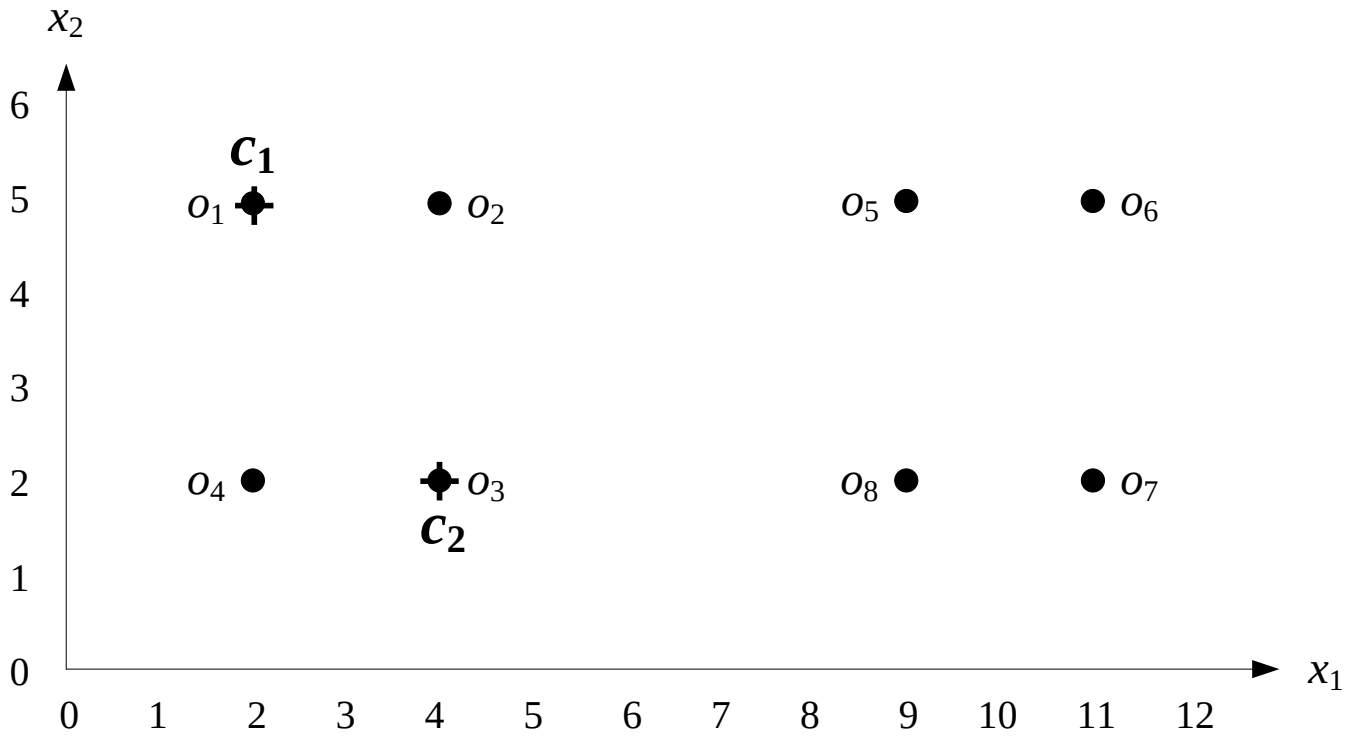
- Menentukan Centroid awal secara acak k=2 yaitu :

$C_1$ yang berada di objek o1

$C_2$ berada di objek o3

# *k-Means Clustering*

Misal *k* = 2. Pilih dua *centroid* secara acak dari 8 objek data (titik)



# Langkah 2

K-means

- Menentukan anggota kluster dengan menghitung jarak objek ke posisi centroid terdekat

$$d\,(x_1, c_1) = \sqrt{\sum_{i=1}^{r} (x_{1i} - c_{1i})^{\,2}}$$

| Data | Jarak ke Centroid | | Cluster yang diikuti |
|---|---|---|---|
| | C1 | C2 | |
| o1 | 0 | 2 | C1 |
| o2 | 2 | 3 | C1 |
| o3 | 3,6 | 0 | C2 |
| o4 | 3 | 2 | C2 |
| o5 | 7 | 5,8 | C2 |
| o6 | 9 | 7,6 | C2 |
| o7 | 9,8 | 7 | C2 |
| o8 | 7,6 | 5 | C2 |

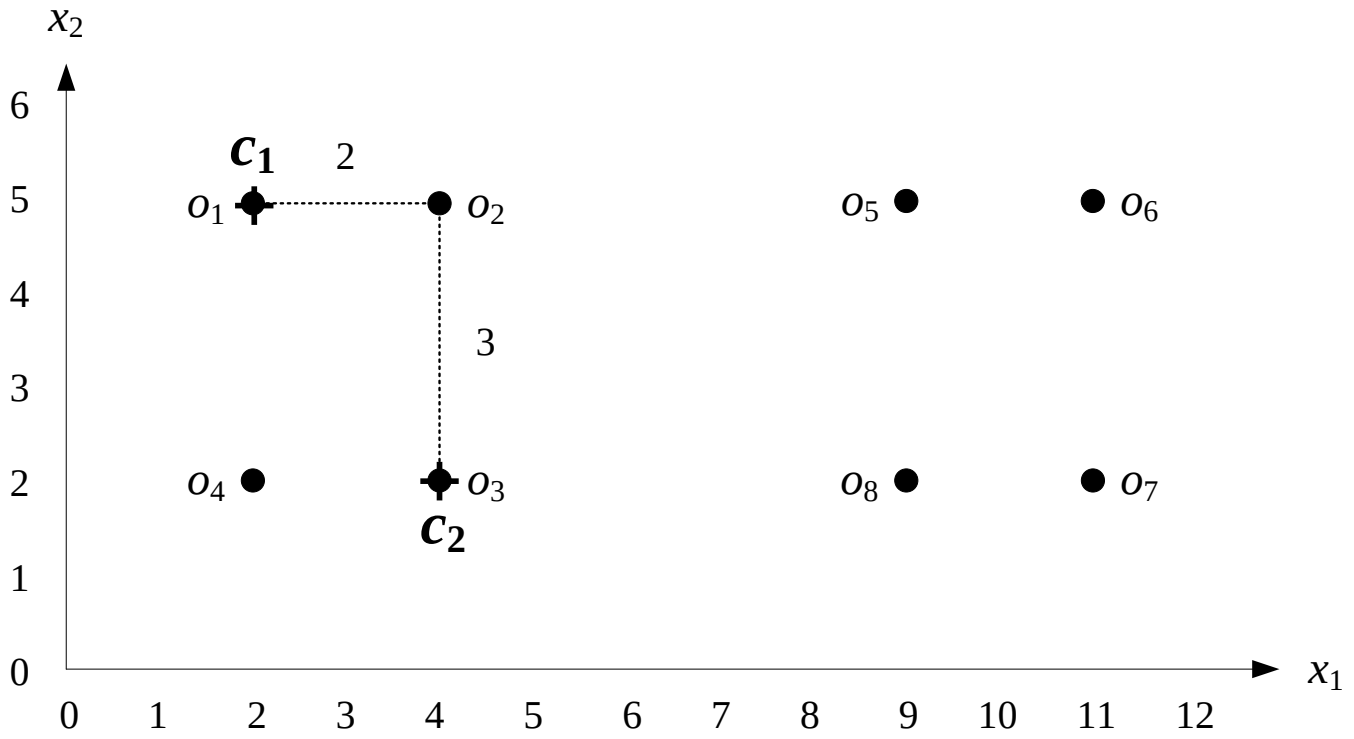
Contoh Perhitungan jarak dari Objek o2 ke centroid C1 dan C2

$$d\,(o_2, c_1) = \sqrt{(4-2)^2 + (5-5)^2} = 2$$

$$d\,(o_2, c_2) = \sqrt{(4-4)^2 + (5-2)^2} = 3$$

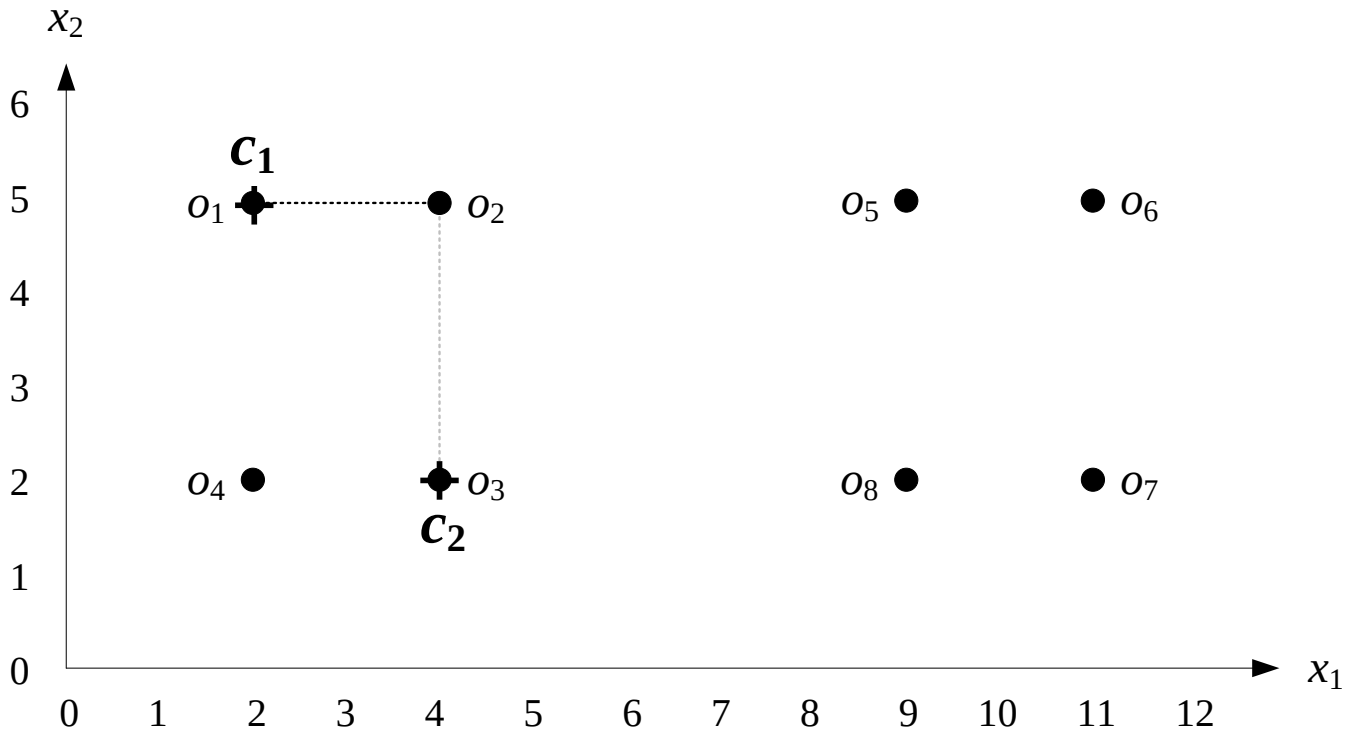
# *k-Means Clustering*

Tentukan anggota setiap klaster dengan memiilih *centroid* terdekat



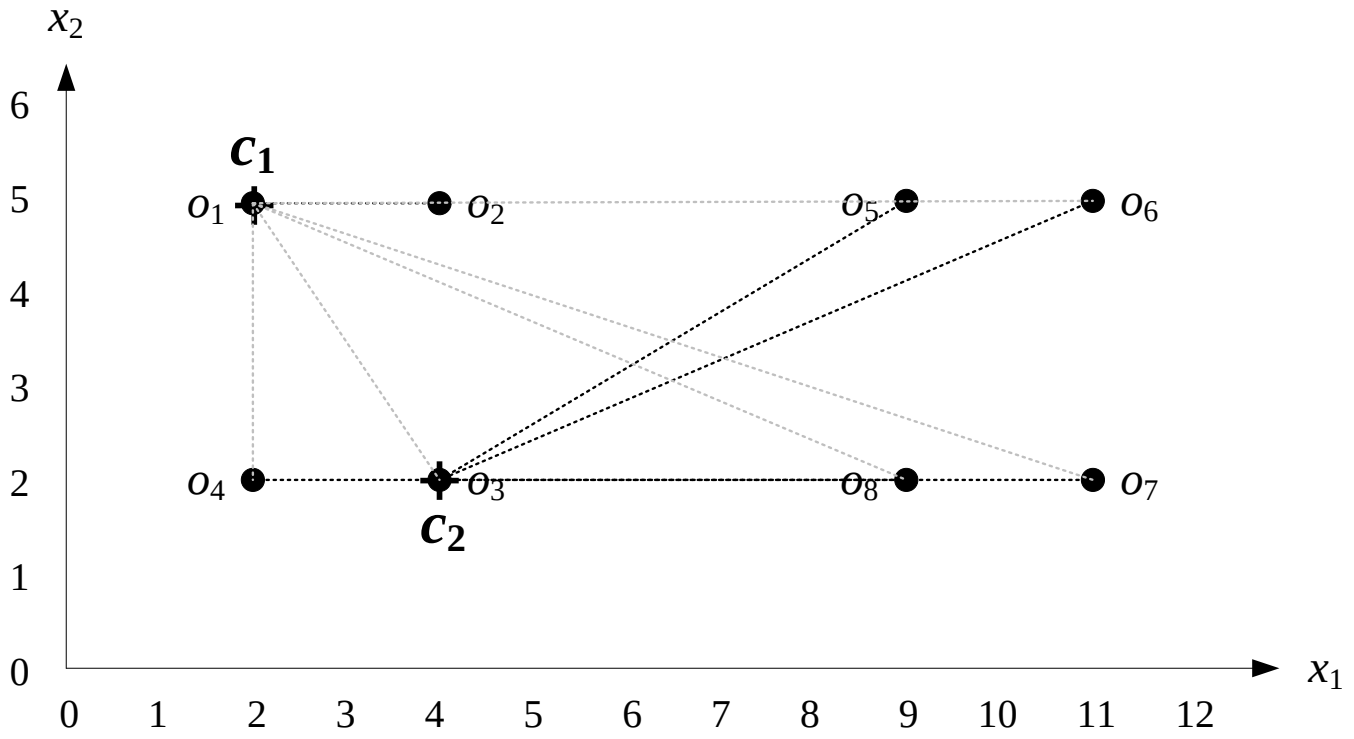
# *k-Means Clustering*

Tentukan anggota setiap klaster dengan memiilih *centroid* terdekat



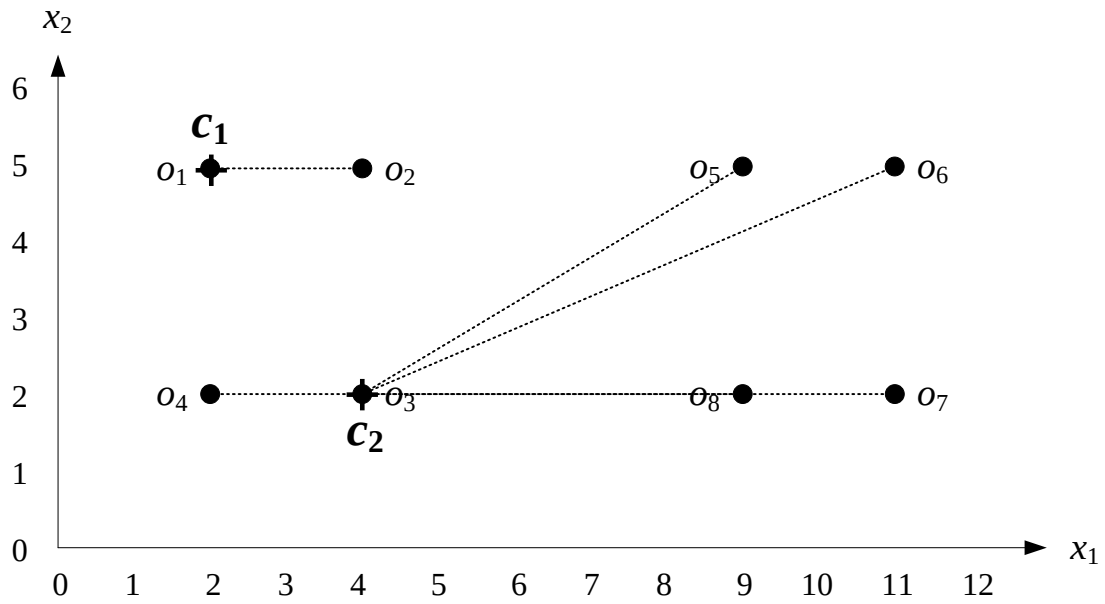
# *k-Means Clustering*

Tentukan anggota setiap klaster dengan memiilih *centroid* terdekat



# *k-Means Clustering*

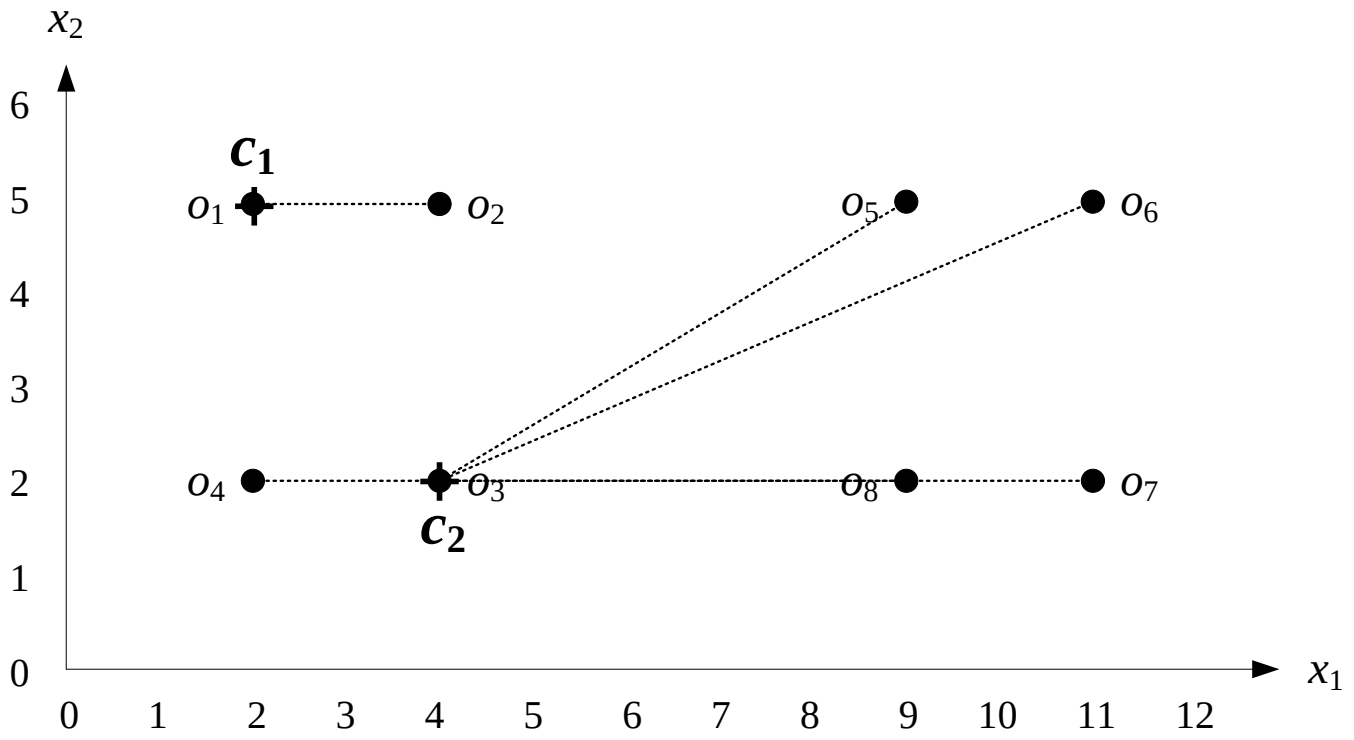
Tentukan anggota setiap klaster dengan memiilih *centroid* terdekat




| Data | Jarak ke Centroid | | Cluster yang diikuti |
|---|---|---|---|
| | C1 | C2 | |
| o1 | 0 | 2 | C1 |
| o2 | 2 | 3 | C1 |
| o3 | 3,6 | 0 | C2 |
| o4 | 3 | 2 | C2 |
| o5 | 7 | 5,8 | C2 |
| o6 | 9 | 7,6 | C2 |
| o7 | 9,8 | 7 | C2 |
| o8 | 7,6 | 5 | C2 |

# *k-Means Clustering*

Hitung rata-rata titik di setiap klaster untuk mendapatkan *centroid* baru



# Langkah 3

- Hitung rata – rata titik masing-masing anggota kluster untuk menentukan titik centroid baru

C1 (x1,x2) = (2+4)/2,(5+5)/2

= 3,5

C2 (x1,x2) =(4+2+9+11+11+9)/6,

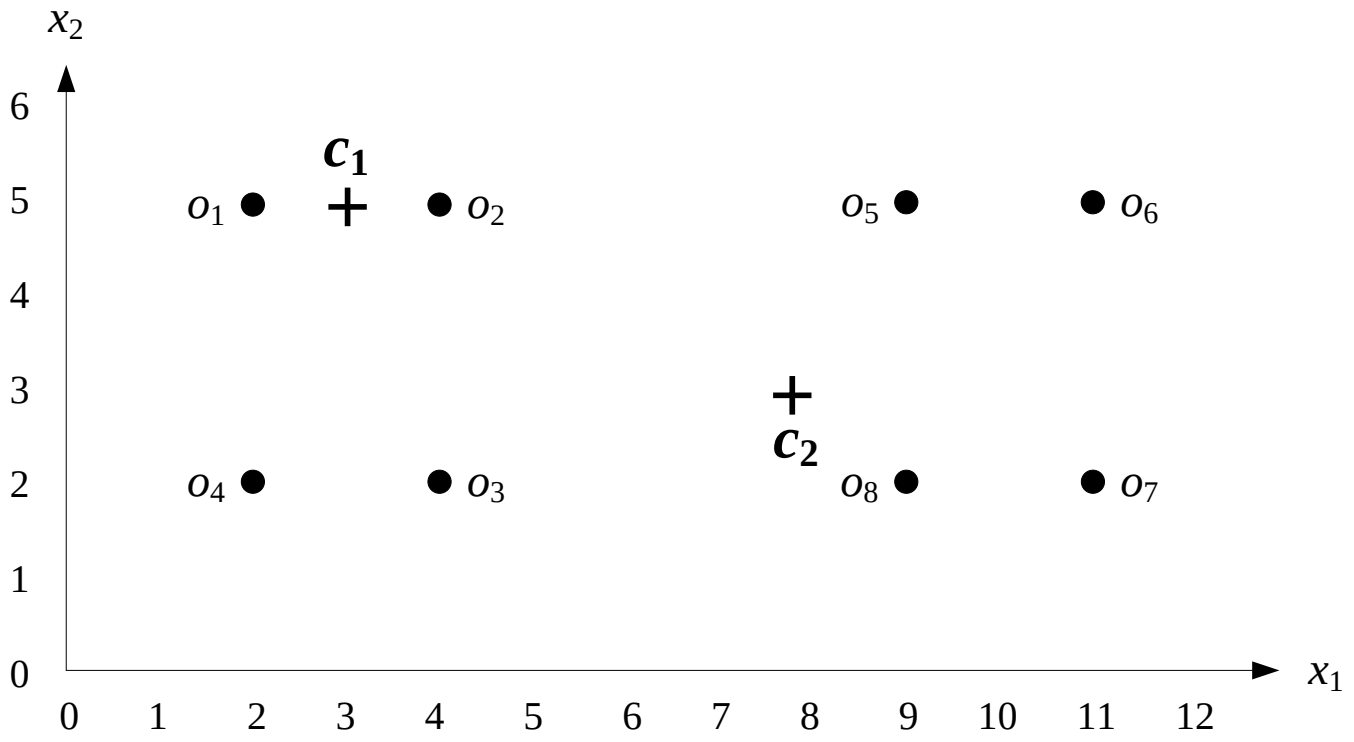
(2+2+5+5+2+2)/6

=7.6 , 3

| Data | x1 | x2 | Centroid |
|---|---|---|---|
| o1 | 2 | 5 | C1 |
| o2 | 4 | 5 | C1 |
| o3 | 4 | 2 | C2 |
| o4 | 2 | 2 | C2 |
| o5 | 9 | 5 | C2 |
| o6 | 11 | 5 | C2 |
| o7 | 11 | 2 | C2 |
| o8 | 9 | 2 | C2 |

# k-Means Clustering

Centroid baru menjadi seperti ini :



# Fungsi Objektif

- **Fungsi objektif berdasarkan jarak dan nilai keanggotaan data dalam cluster**

$$J = \sum_{i=1}^{N}\sum_{l=1}^{K} a_{ic} D(x_i, C_l)^2$$

  - Dimana N adalah jumlah data, K adalah jumlah cluster, $a_{il}$ adalah nilai keanggotaan titik data $x_i$ ke pusat cluster $C_l$, $C_l$ adalah pusat cluster ke-*l*, $D(x_i,C_l)$ adalah jarak titik $x_i$ ke cluster $C_l$ yang diikuti.
  - Untuk *a* mempunyai nilai 0 atau 1. Apabila suatu data merupakan anggota suatu kelompok maka nilai $a_{il}$ =1 , jika tidak, akan maka nilai $a_{il}$ =0

$$J = \sum_{i=1}^{N} \sum_{l=1}^{K} a_{ic} D(x_i, C_l)^2$$

# Langkah 4

K-means

- Apakah Centroid Konvergen ...? Hitung dengan fungsi Objektif (J)

Contoh Fungsi Objektif $O_1$ :

$D(X_1,C_1)^2 = (2-3)^2+(5-5)^2 = 1$

Perubahan Fungsi Objektif

=|j baru- J lama |

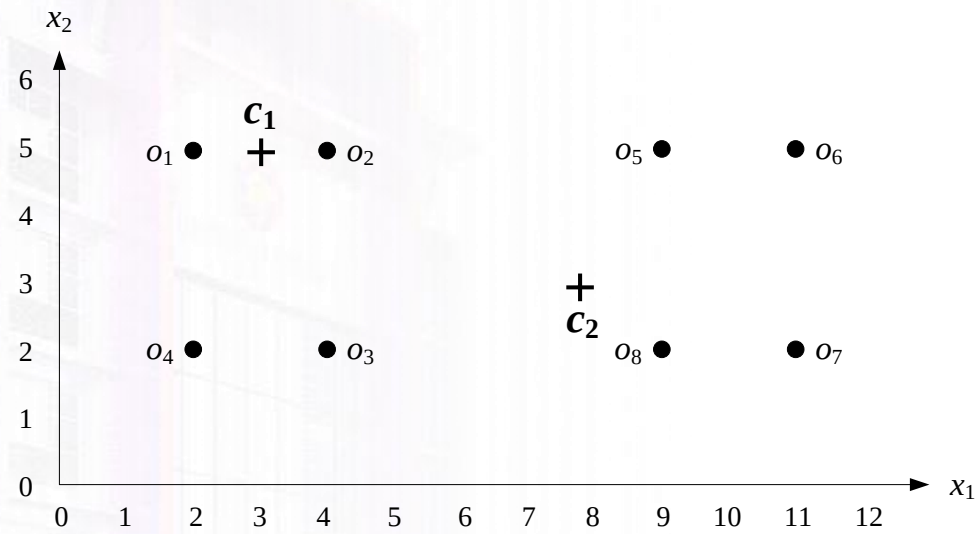
=|114,16-0= 114,16

Perubahan masih di atas ambang batas threshold (T)>0,1, artinya pencarian centroid masih terus dilakukan

| Data | x1 | x2 | C1 | C2 | Centroid |
|---|---|---|---|---|---|
| o1 | 2 | 5 | 1 | | C1 |
| o2 | 4 | 5 | 1 | | C1 |
| o3 | 4 | 2 | | 13,96 | C2 |
| o4 | 2 | 2 | | 32,36 | C2 |
| o5 | 9 | 5 | | 10,76 | C2 |
| o6 | 11 | 5 | | 25,16 | C2 |
| o7 | 11 | 2 | | 22,16 | C2 |
| o8 | 9 | 2 | | 7,76 | C2 |
| | | | 2 | 112,16 | |
| | | | | Fungsi Objektif | 114,16 |

# Ulangi Langkah 2 dengan centroid baru



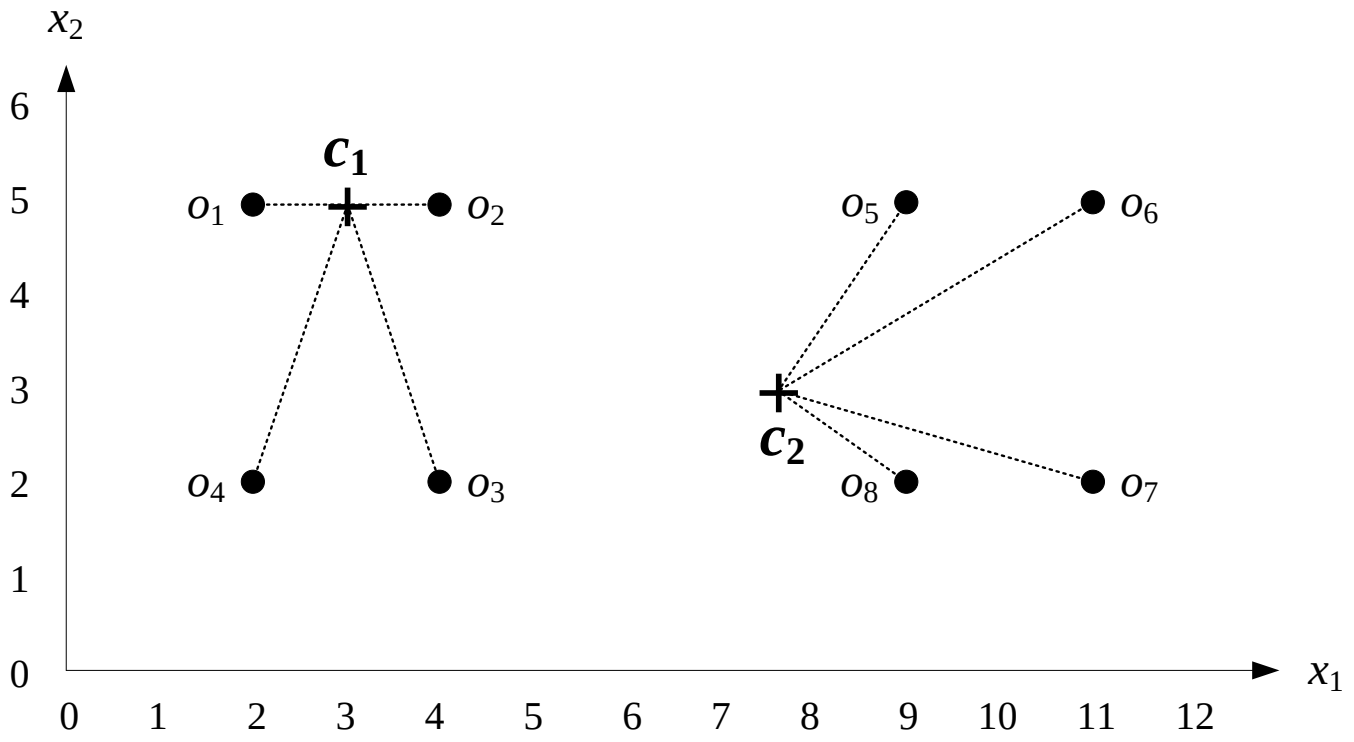



| Data | Jarak ke Centroid | | Cluster Lama | Cluster Baru |
|---|---|---|---|---|
| | C1 | C2 | | |
| o1 | ? | ? | C1 | C1 |
| o2 | ? | ? | C1 | C1 |
| o3 | 3,16 | 3,6 | C2 | C1 |
| o4 | ? | ? | C2 | C1 |
| o5 | ? | ? | C2 | C2 |
| o6 | ? | ? | C2 | C2 |
| o7 | ? | ? | C2 | C2 |
| o8 | ? | ? | C2 | C2 |

**Perbarui anggota setiap klaster dengan memilih *centroid* terdekat**

# k-Means Clustering

Anggota kluster yang baru (ke-2)



# Ulangi Langkah 3

- Hitung rata – rata titik masing-masing anggota kluster untuk menentukan titik centroid baru

C1 (x1,x2) = 12/4, 14/4

= 3, 3.5

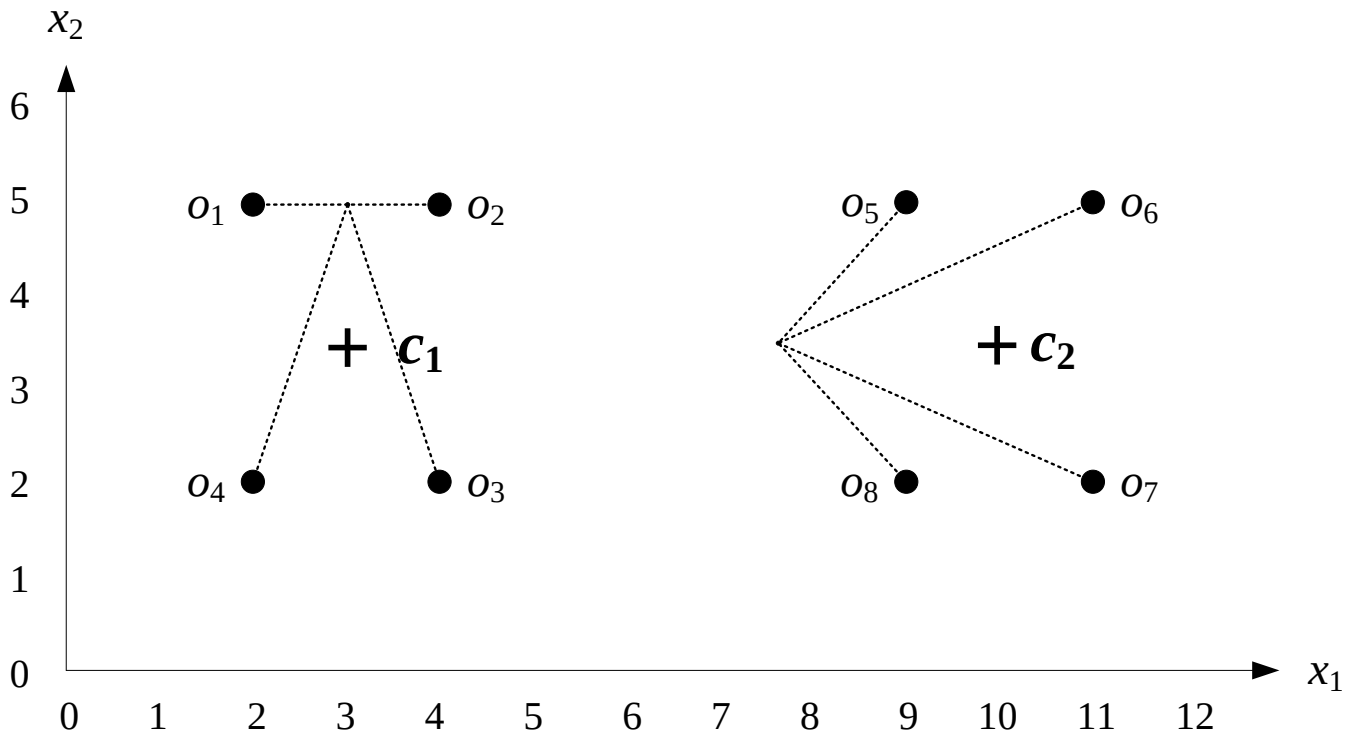
C2 (x1,x2) = 40/4, 14/4

=10, 3.5

| Data | x1 | x2 | Centroid baru |
|---|---|---|---|
| o1 | 2 | 5 | C1 |
| o2 | 4 | 5 | C1 |
| o3 | 4 | 2 | C1 |
| o4 | 2 | 2 | C1 |
| o5 | 9 | 5 | C2 |
| o6 | 11 | 5 | C2 |
| o7 | 11 | 2 | C2 |
| o8 | 9 | 2 | C2 |

# *k-Means Clustering*

Posisi Centroid Baru



# Ulangi Langkah 4



- Apakah Centroid Konvergen ...? Hitung dengan fungsi Objektif (J)

Perubahan Fungsi Objektif

=|j baru- J lama |

=|25,5-114,16 |

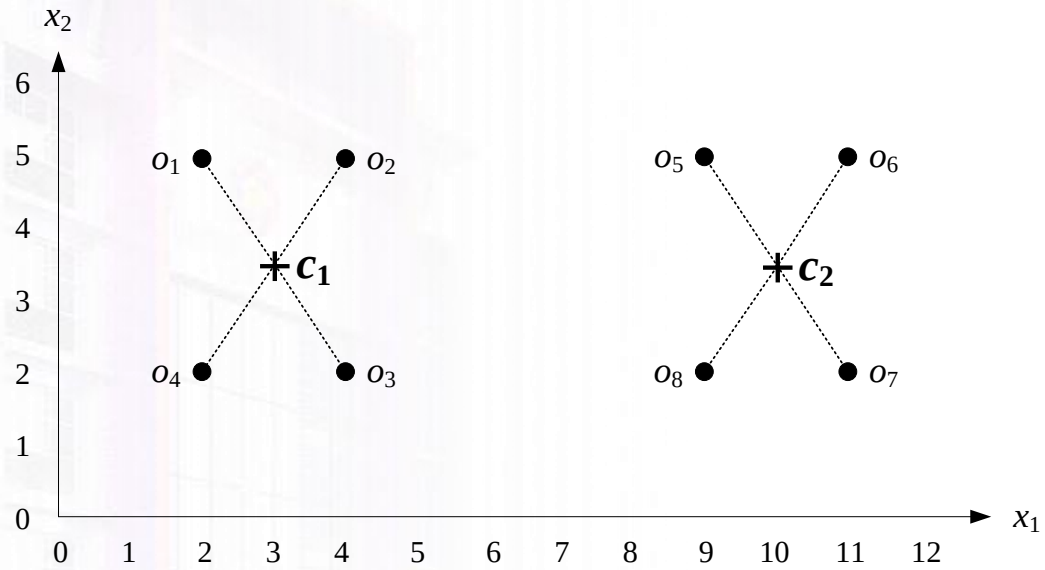
=88,66

Perubahan masih di atas ambang batas threshold (T)>0,1, artinya pencarian centroid masih terus dilakukan

| Data | x1 | x2 | C1 | C2 | Centroid |
|---|---|---|---|---|---|
| o1 | 2 | 5 | 1 | | C1 |
| o2 | 4 | 5 | 1 | | C1 |
| o3 | 4 | 2 | 3,25 | | C1 |
| o4 | 2 | 2 | 3,25 | | C1 |
| o5 | 9 | 5 | | 3,25 | C2 |
| o6 | 11 | 5 | | 7,25 | C2 |
| o7 | 11 | 2 | | 3,25 | C2 |
| o8 | 9 | 2 | | 3,25 | C2 |
| | | | 8,5 | 17 | |
| | | | | Fungsi Objektif | 25,5 |

# Ulangi langkah 2 dengan centroid baru





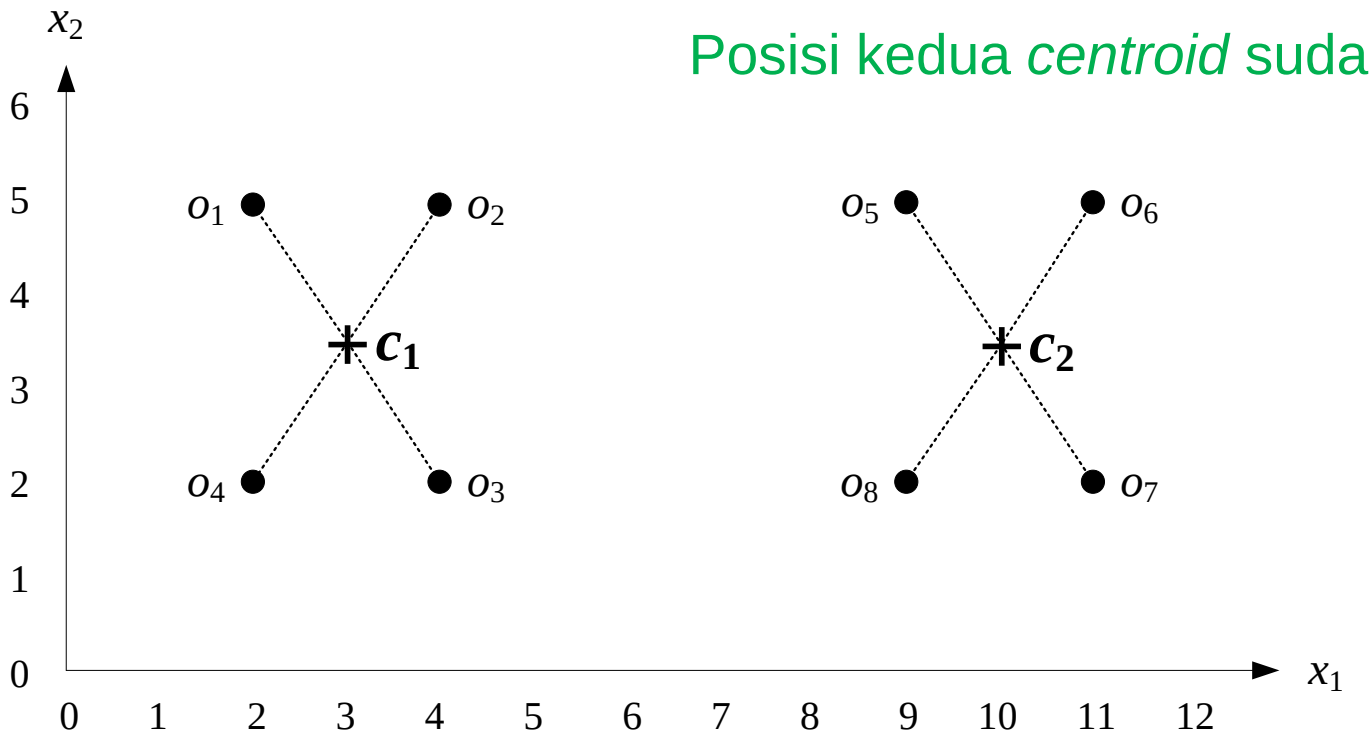

| Data | Jarak ke Centroid | | Cluster Lama | Cluster Baru |
|---|---|---|---|---|
| | C1 | C2 | | |
| o1 | 1,80 | 8,13 | C1 | C1 |
| o2 | 1,80 | 6,18 | C1 | C1 |
| o3 | 1,80 | 6,18 | C1 | C1 |
| o4 | 1,80 | 8,13 | C1 | C1 |
| o5 | 6,18 | 1,80 | C2 | C2 |
| o6 | 8,13 | 1,80 | C2 | C2 |
| o7 | 8,13 | 1,80 | C2 | C2 |
| o8 | 6,18 | 1,80 | C2 | C2 |

Tidak ada perubahan cluster lama dengan cluster baru, pencarian centroid berakhir

# *k-Means Clustering*

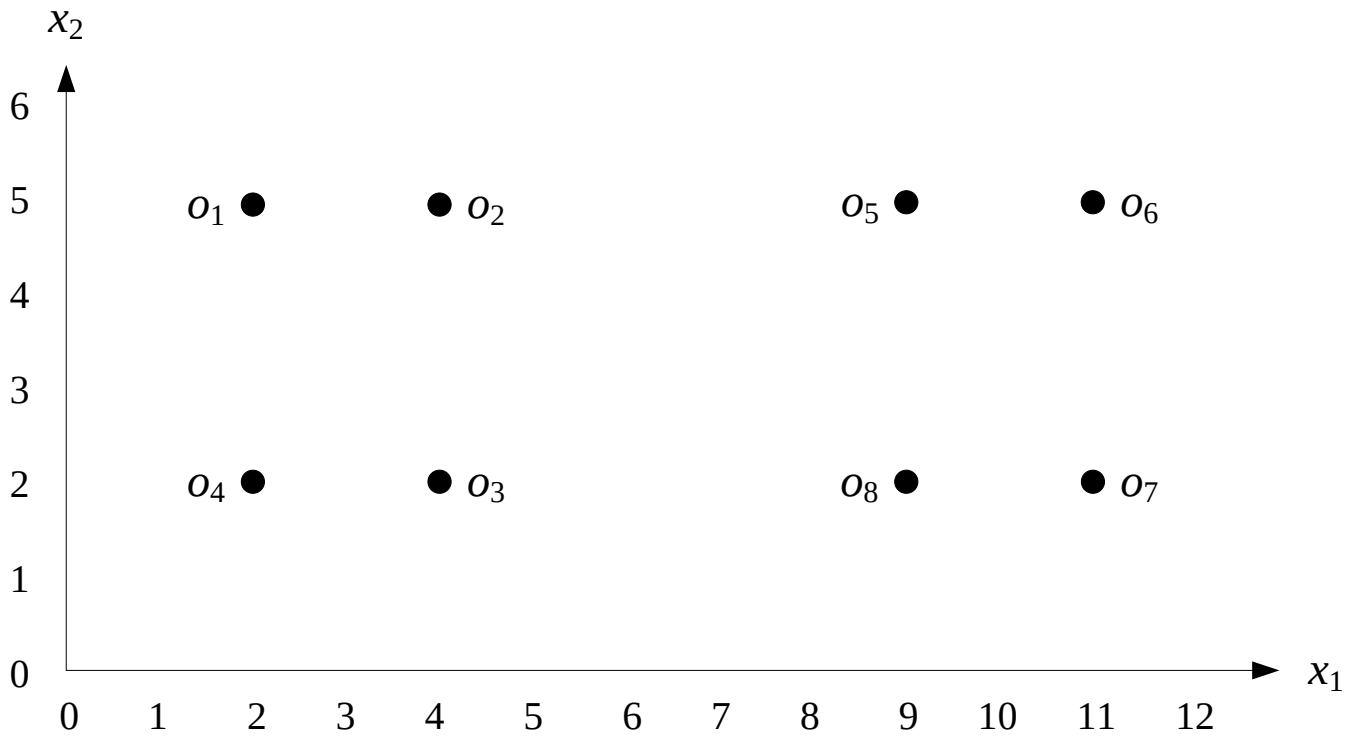
Perbarui anggota setiap klaster dengan memiilih *centroid* terdekat.

Posisi kedua *centroid* sudah **stabil**



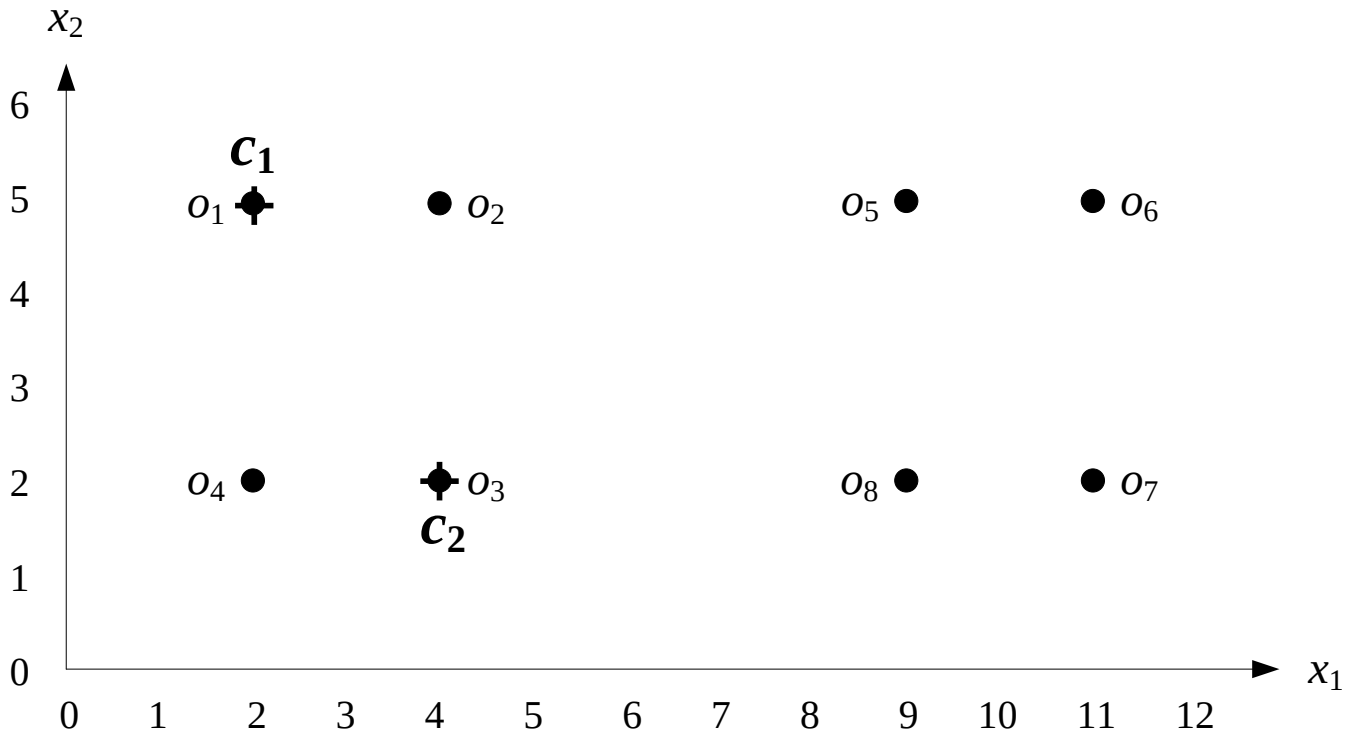
# *k-Means Clustering*

Ilustrasi *k-Means* dengan klaster berbentuk lingkaran



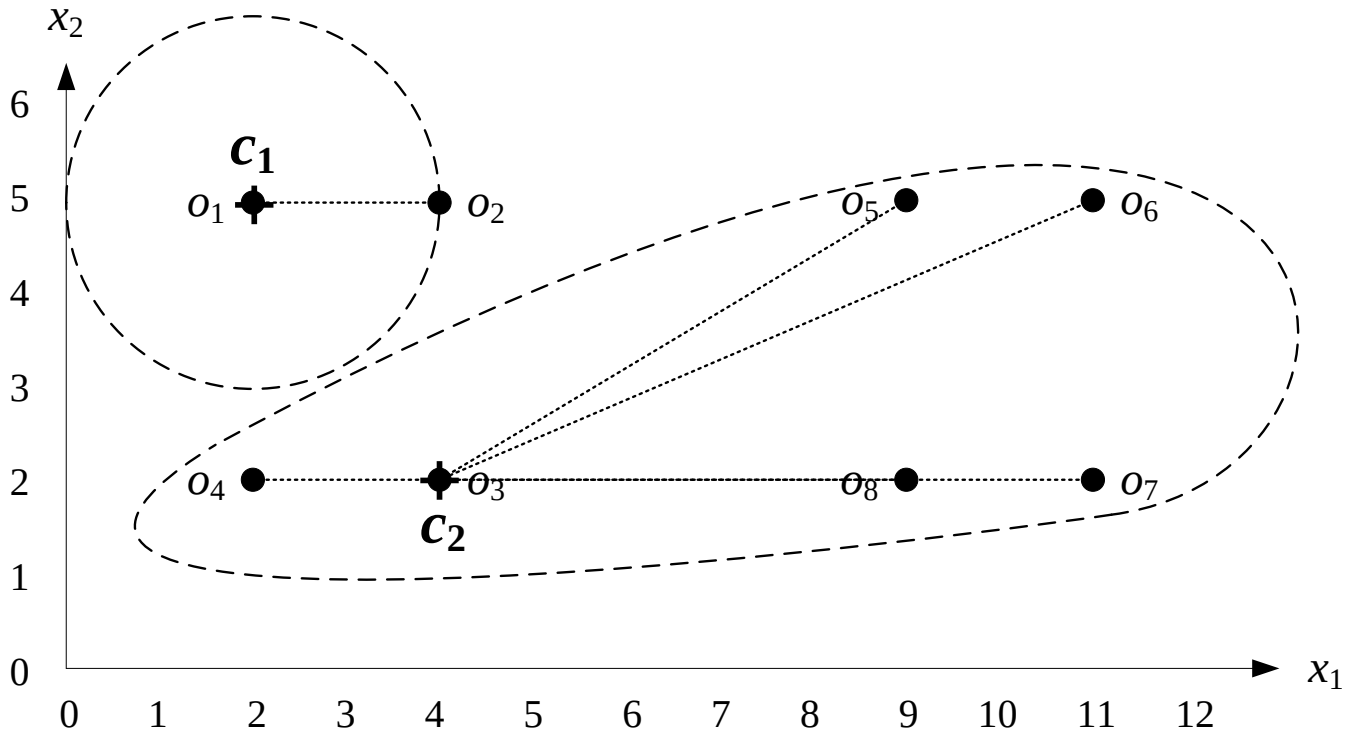
# *k-Means Clustering*

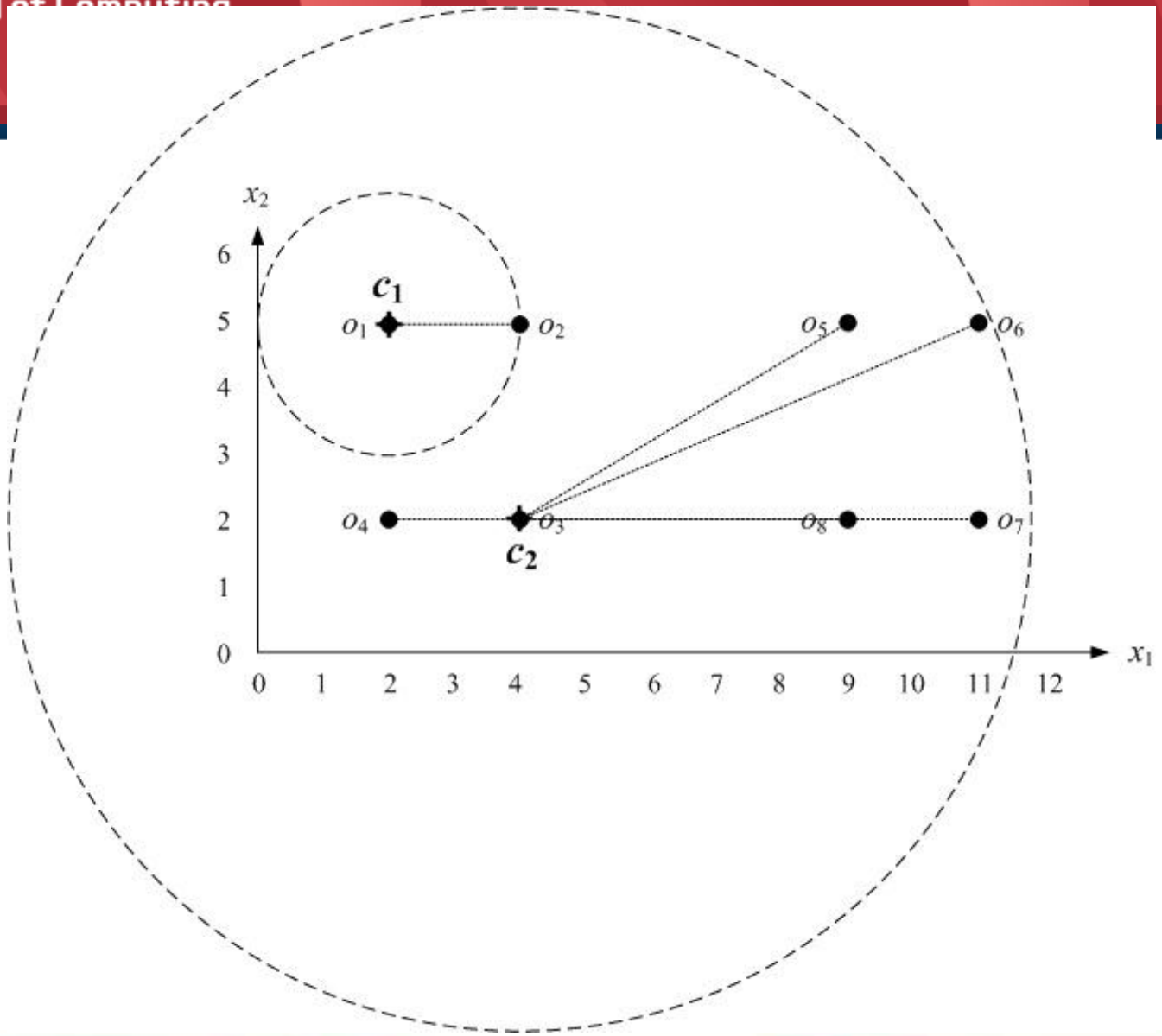
Misal *k* = 2. Pilih dua *centroid* secara acak dari 8 objek data (titik)



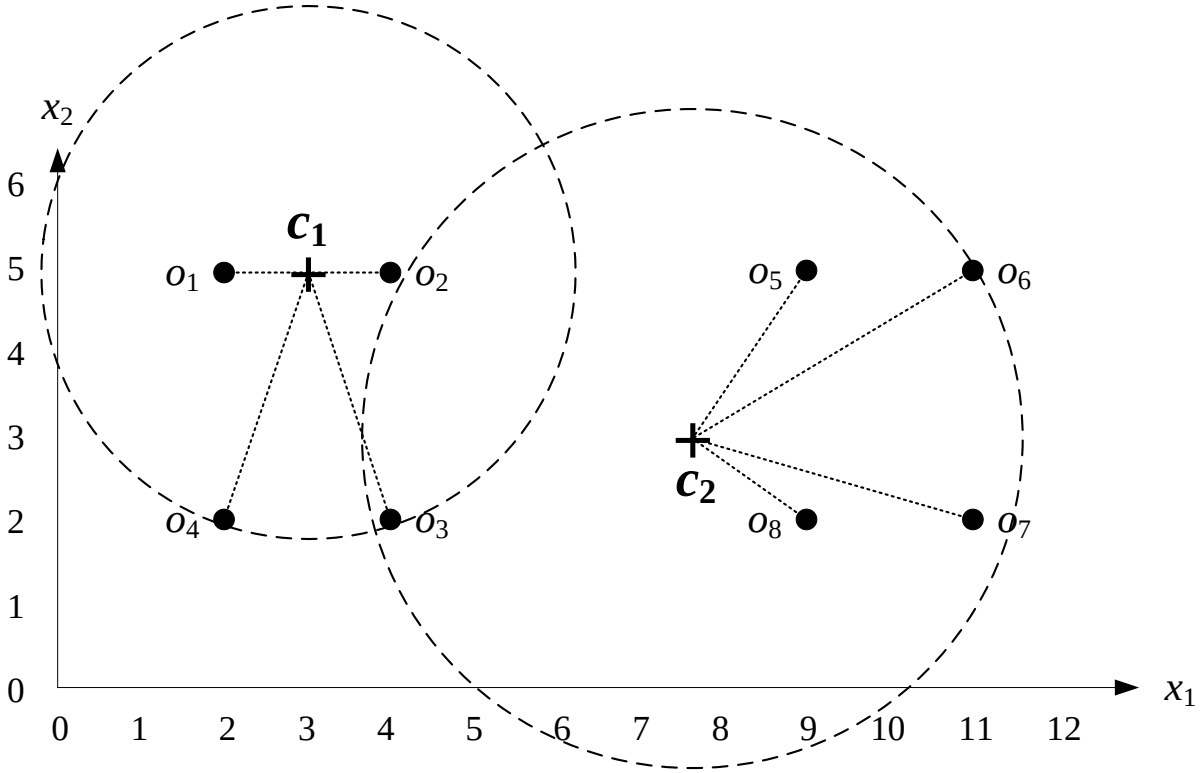
# *k-Means Clustering*

Tentukan anggota setiap klaster dengan memiilih *centroid* terdekat

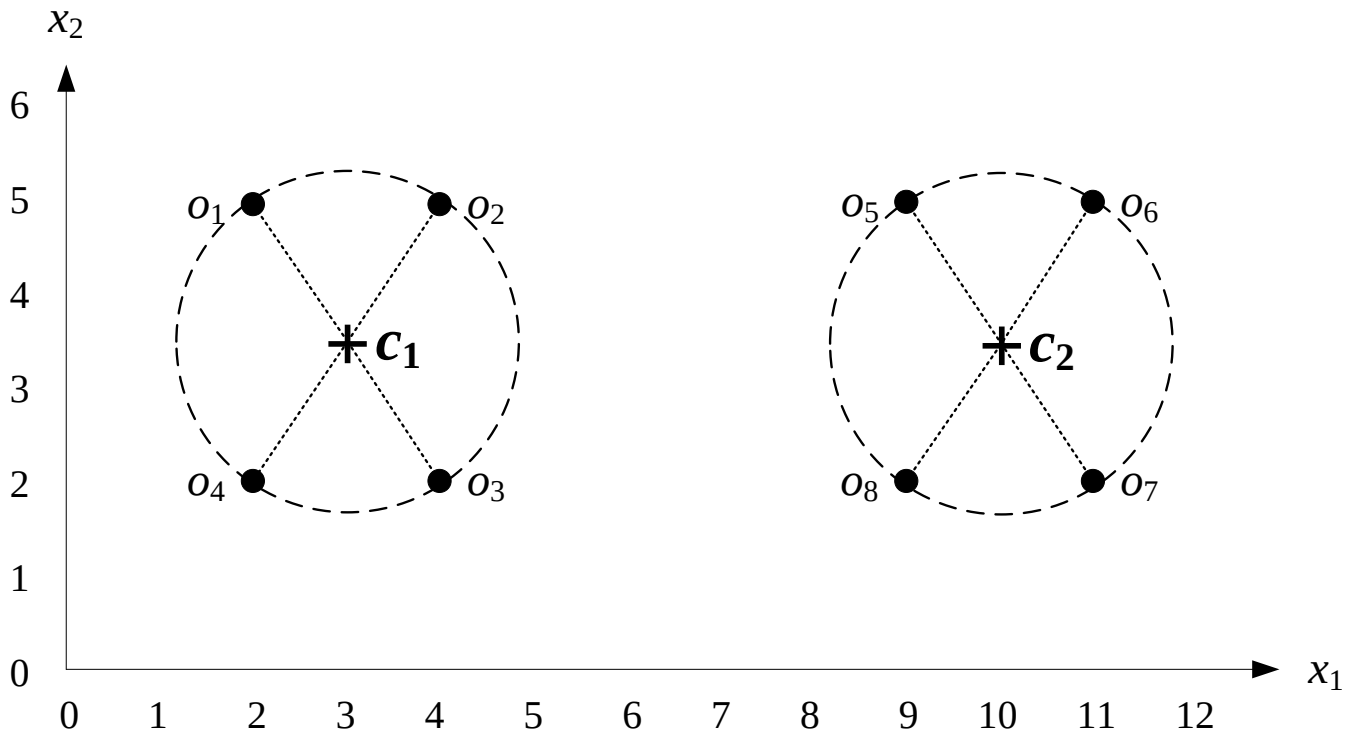



$x_2$
6
5
4
3
2
1
0
$c_1$
$o_1$
$o_2$
$o_5$
$o_6$
$o_4$
$o_3$
$o_8$
$o_7$
$c_2$
0 1 2 3 4 5 6 7 8 9 10 11 12
$x_1$

# k-Means Clustering



# k-Means Clustering



# latihan



- Ada 10 data pada data set.
- Dimensi data ada 2 fitur (agar mudah dalam visualisasi koordinat kartesius).
- Fitur yang digunakan dalam pengelompokan adalah x dan y
- Jarak yang digunakan adalah Euclidean distance.
- Jumlah cluster (K) adalah 3.
- Threshold (T) yang digunakan untuk perubahan fungsi objektif adalah 0.1.

| Data ke-i | Fitur x | Fitur y |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 4 | 1 |
| 3 | 6 | 1 |
| 4 | 1 | 2 |
| 5 | 2 | 3 |
| 6 | 5 | 3 |
| 7 | 2 | 5 |
| 8 | 3 | 5 |
| 9 | 2 | 6 |
| 10 | 3 | 8 |

Misalnya : Centroid Awal

| Cluster | Fitur x | Fitur y |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 3.4 | 3.8 |
| 3 | 2.75 | 3.75 |