DATA MINING

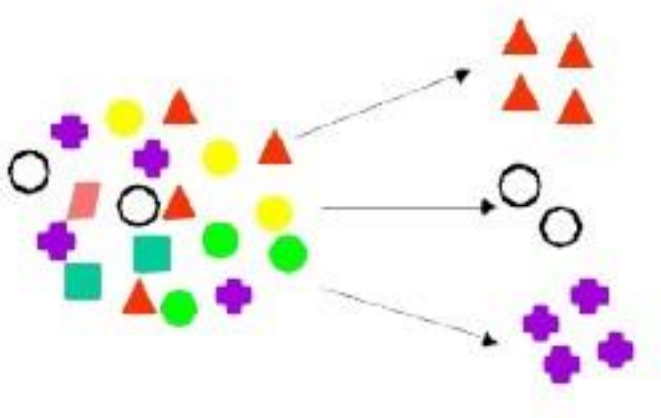
3 SKS | Semester 6 | S1 Sistem Informasi

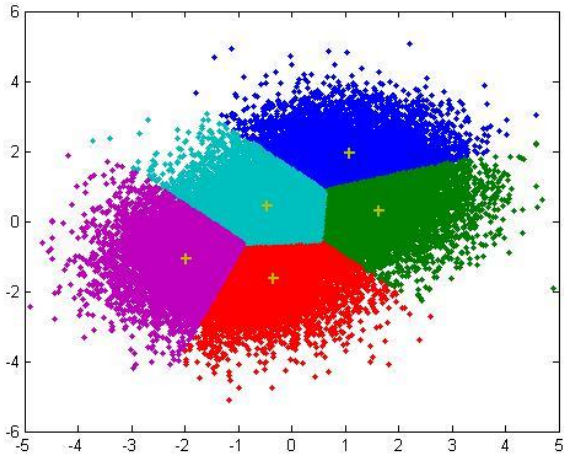
TEKNIK CLUSTERING - K-MEANS

Nizar Rabbi Radliya
nizar@email.unikom.ac.id

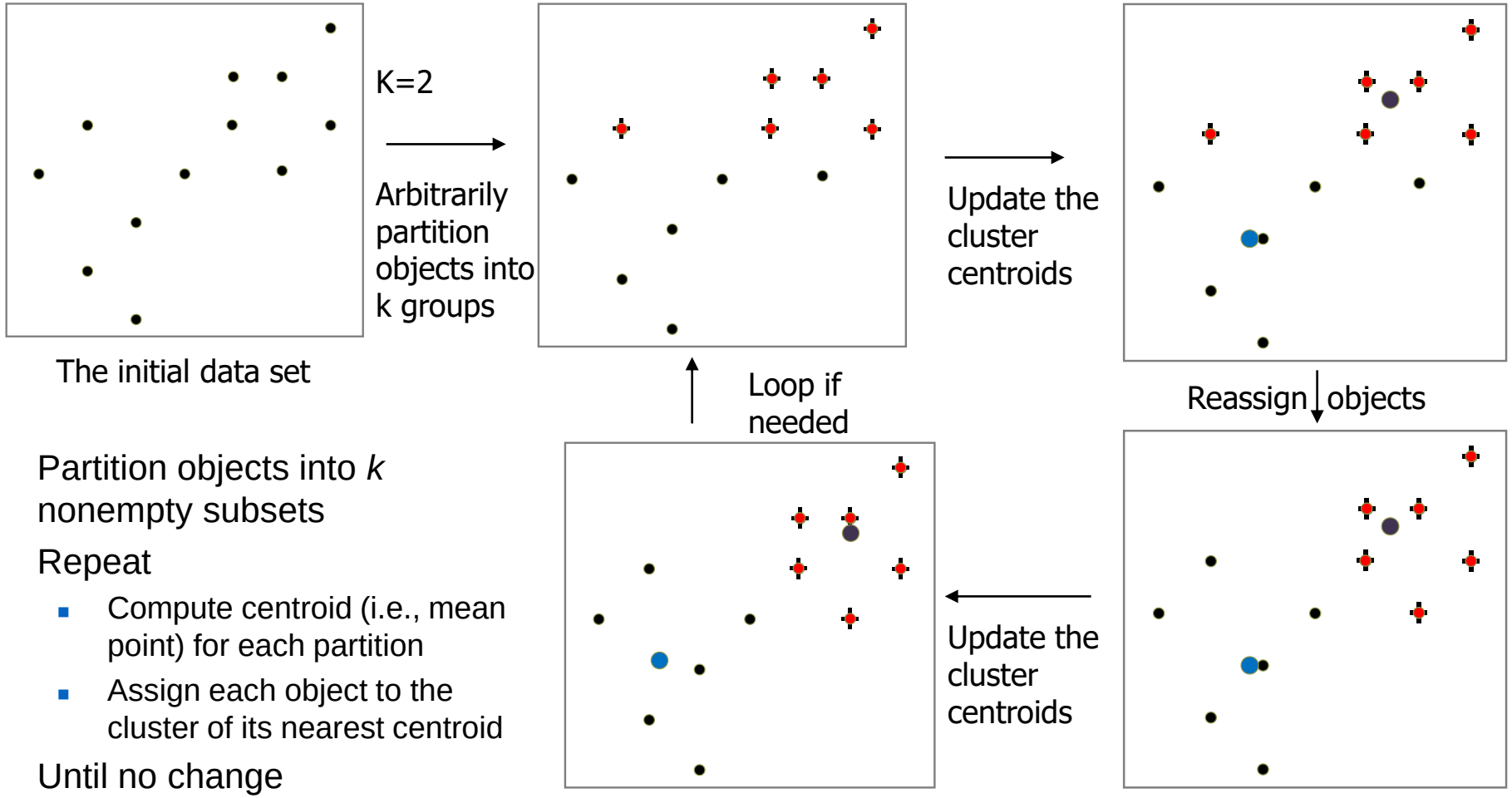
TEKNIK CLUSTERING

Penklusteran (clustering) digunakan untuk melakukan pengelompokan data-data ke dalam sejumlah kelompok (cluster) berdasarkan karakteristik masing-masing data pada kelompok-kelompok yang ada.





TEKNIK CLUSTERING



- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

ALGORITMA K-MEANS

- Metode pengelompokan data partitioning (non hierarki).
- Data berkarakteristik sama dimasukkan ke dalam satu kelompok.
- Meminimalkan variasi dalam satu kelompok dan memaksimalkan variasi antar kelompok.

ALGORITMA K-MEANS : TAHAPAN PENGELOMPOKAN

1. Tentukan jumlah kelompok (K) dan nilai ambang batas atau threshold (T)
2. Alokasikan data ke dalam kelompok secara acak
3. Hitung pusat kelompok atau centroid (C) untuk setiap kelompok
4. Alokasikan semua data ke centroid terdekat
5. Kembali ke langkah 3 (iterasi), apabila masih ada:
   - data yang berpindah kelompok,
   - atau ada perubahan nilai centroid di atas nilai ambang yang ditentukan,
   - atau ada perubahan nilai pada fungsi objektif yang digunakan (di atas nilai ambang yang ditentukan).

ALGORITMA K-MEANS : FORMULA YANG DIGUNAKAN

Formula untuk menghitung centroid:

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j$$

Formula untuk menghitung jarak data dengan centroid:

D(x,y) $= \sqrt{\sum_{j=1}^{n} |x-y|^2}$ ... *Jarak Euclidean*

D(x,y) $= \sum_{j=1}^{n} |x-y|$ ... *Jarak Manhattan/City Block*

D(x,y) = max (|x - y|) ... *Jarak Chebyshev*

Formula untuk menghitung fungsi objektif:

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

CONTOH KASUS

***Data training* (data latih): Pengelompokan data dua dimensi**

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

Langkah 1: menentukan jumlah kelompok (K), K = 3

Ambang batas atau *threshold* (T) yang akan digunakan untuk perubahan fungsi objektif adalah 0.1

Fungsi Objektif (J) awal = 0

CONTOH KASUS

Langkah 2: Alokasikan data ke dalam kelompok secara acak

Data ke-i	Fitur x	Fitur y	Kelompok 1	Kelompok 2	Kelompok 3
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		*	

CONTOH KASUS

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 1:

Data ke-i	Fitur x	Fitur y
1	1	1
M (Jumlah Data) 1	Jumlah x 1	Jumlah y 1
Rata-rata	1	1

Centroid untuk kelompok 2:

Data ke-i	Fitur x	Fitur y
3	6	1
4	1	2
6	5	3
7	2	5
10	3	8
M (Jumlah Data) 5	Jumlah x 17	Jumlah y 19
Rata-rata	3.400	3.800

CONTOH KASUS

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 3:

Data ke-i	Fitur x	Fitur y
2	4	1
5	2	3
8	3	5
9	2	6
M (Jumlah Data)	Jumlah x	Jumlah y
4	11	15
Rata-rata	2.750	3.750

Centroid pertama untuk setiap kelompok:

Kelompok	Fitur x	Fitur y
1	1	1
2	3.400	3.800
3	2.750	3.750

CONTOH KASUS **Menghitung fungsi objektif baru**

Jarak data ke centroid pertama (menggunakan jarak *euclidean*):

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	K
1	1	1	0			1
2	4	1			3.021	3
3	6	1		3.821		2
4	1	2		3.000		2
5	2	3			1.061	3
6	5	3		1.789		2
7	2	5		1.844		2
8	3	5			1.275	3
9	2	6			2.372	3
10	3	8		4.219		2
			0	14.673	7.728	22.401
			Perubahan Fungsi Objektif			22.401

CONTOH KASUS

Menghitung perubahan fungsi objektif

Fungsi Objektif (J) lama/sebelumnya = 0

Fungsi Objektif (J) baru/sekarang = 0 + 14.673 + 7.728 = 22.401

Perubahan Fungsi Objektif = | J baru – J lama |
= | 22.401 – 0 | = 22.401

Perubahan masih di atas ambang batas atau threshold (T); >0.1, artinya pencarian centroid masih terus dilakukan.

CONTOH KASUS

Langkah 4: Alokasikan semua data ke centroid terdekat

(menggunakan jarak *euclidean*)

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	Terdekat	K Baru	K Lama
1	1	1	0	3.688	3.260	0	1	1
2	4	1	3	2.864	3.021	2.864	2	3
3	6	1	5	3.821	4.257	3.821	2	2
4	1	2	1	3	2.475	1	1	2
5	2	3	2.236	1.612	1.061	1.061	3	3
6	5	3	4.472	1.789	2.372	1.789	2	2
7	2	5	4.123	1.844	1.458	1.458	3	2
8	3	5	4.472	1.265	1.275	1.265	2	3
9	2	6	5.099	2.608	2.372	2.372	3	3
10	3	8	7.280	4.219	4.257	4.219	2	2

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

CONTOH KASUS

Langkah 4: Alokasikan semua data ke centroid terdekat

Data ke-i	Fitur x	Fitur y	Kelompok 1	Kelompok 2	Kelompok 3
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		*	

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

CONTOH KASUS

Iterasi – 1

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 1:

Data ke-i	Fitur x	Fitur y
1	1	1
4	1	2
M (Jumlah Data)	Jumlah x	Jumlah y
2	2	3
Rata-rata	1	1.500

Centroid untuk kelompok 2:

Data ke-i	Fitur x	Fitur y
2	4	1
3	6	1
6	5	3
8	3	5
10	3	8
M (Jumlah Data)	Jumlah x	Jumlah y
5	21	18
Rata-rata	4.200	3.600

CONTOH KASUS

Iterasi – 1

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 3:

Data ke-i	Fitur x	Fitur y
5	2	3
7	2	5
9	2	6
M (Jumlah Data) 3	Jumlah x 6	Jumlah y 14
Rata-rata	2	4.667

Centroid kedua untuk setiap kelompok (hasil iterasi – 1):

Kelompok	Fitur x	Fitur y
1	1	1.500
2	4.200	3.600
3	2	4.667

CONTOH KASUS

Iterasi – 1

Menghitung fungsi objektif baru

Jarak data ke centroid kedua (menggunakan jarak *euclidean*):

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	K
1	1	1	0.500			1
2	4	1		2.608		2
3	6	1		3.162		2
4	1	2	0.500			1
5	2	3			1.667	3
6	5	3		1.000		2
7	2	5			0.333	3
8	3	5		1.844		2
9	2	6			1.333	3
10	3	8		4.561		2
			1.000	13.175	3.333	17.508
			Perubahan Fungsi Objektif			4.893

CONTOH KASUS

Iterasi – 1

Menghitung perubahan fungsi objektif

Fungsi Objektif (J) lama/sebelumnya = 22.401

Fungsi Objektif (J) baru/sekarang = 1 + 13.175 + 3.333 = 17.508

Perubahan Fungsi Objektif = | J baru – J lama |
= | 17.508 – 22.401 | = 4.893

Perubahan masih di atas ambang batas atau threshold (T); >0.1, artinya pencarian centroid masih terus dilakukan.

CONTOH KASUS

Iterasi – 1

Langkah 4: Alokasikan semua data ke centroid terdekat

(menggunakan jarak *euclidean*)

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	Terdekat	K Baru	K Lama
1	1	1	0.500	4.123	3.801	0.500	1	1
2	4	1	3.041	2.608	4.177	2.608	2	2
3	6	1	5.025	3.162	5.426	3.162	2	2
4	1	2	0.500	3.578	2.848	0.500	1	1
5	2	3	1.803	2.280	1.667	1.667	3	3
6	5	3	4.272	1	3.432	1	2	2
7	2	5	3.640	2.608	0.333	0.333	3	3
8	3	5	4.031	1.844	1.054	1.054	3	2
9	2	6	4.610	3.256	1.333	1.333	3	3
10	3	8	6.801	4.561	3.480	3.480	3	2

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

CONTOH KASUS

Iterasi – 1

Langkah 4: Alokasikan semua data ke centroid terdekat

Data ke-i	Fitur x	Fitur y	Kelompok 1	Kelompok 2	Kelompok 3
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			*

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

CONTOH KASUS

Iterasi – 2

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 1:

Data ke-i	Fitur x	Fitur y
1	1	1
4	1	2
M (Jumlah Data)	Jumlah x	Jumlah y
2	2	3
Rata-rata	1	1.500

Centroid untuk kelompok 2:

Data ke-i	Fitur x	Fitur y
2	4	1
3	6	1
6	5	3
M (Jumlah Data)	Jumlah x	Jumlah y
3	15	5
Rata-rata	5	1.667

CONTOH KASUS

Iterasi – 2

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 3:

Data ke-i	Fitur x	Fitur y
5	2	3
7	2	5
8	3	5
9	2	6
10	3	8
M (Jumlah Data) 5	Jumlah x 12	Jumlah y 27
Rata-rata	2.400	5.400

Centroid ketiga untuk setiap kelompok (hasil iterasi – 2):

Kelompok	Fitur x	Fitur y
1	1	1.500
2	5	1.667
3	2.400	5.400

CONTOH KASUS

Iterasi – 2

Menghitung fungsi objektif baru

Jarak data ke centroid ketiga (menggunakan jarak *euclidean*):

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	K
1	1	1	0.500			1
2	4	1		1.202		2
3	6	1		1.202		2
4	1	2	0.500			1
5	2	3			2.433	3
6	5	3		1.333		2
7	2	5			0.566	3
8	3	5			0.721	3
9	2	6			0.721	3
10	3	8			2.668	3
			1.000	3.737	7.109	11.846
			Perubahan Fungsi Objektif			5.662

CONTOH KASUS

Iterasi – 2

Menghitung perubahan fungsi objektif

Fungsi Objektif (J) lama/sebelumnya = 17.508

Fungsi Objektif (J) baru/sekarang = 1 + 3.737 + 7.109 = 11.846

Perubahan Fungsi Objektif = | J baru – J lama |
= | 11.846 – 17.508 | = 5.662

Perubahan masih di atas ambang batas atau threshold (T); >0.1, artinya pencarian centroid masih terus dilakukan.

CONTOH KASUS

Iterasi – 2

Langkah 4: Alokasikan semua data ke centroid terdekat

(menggunakan jarak *euclidean*)

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	Terdekat	K Baru	K Lama
1	1	1	0.500	4.055	4.617	0.500	1	1
2	4	1	3.041	1.202	4.682	1.202	2	2
3	6	1	5.025	1.202	5.685	1.202	2	2
4	1	2	0.500	4.014	3.677	0.500	1	1
5	2	3	1.803	3.283	2.433	1.803	1	3
6	5	3	4.272	1.333	3.538	1.333	2	2
7	2	5	3.640	4.484	0.566	0.566	3	3
8	3	5	4.031	3.887	0.721	0.721	3	3
9	2	6	4.610	5.270	0.721	0.721	3	3
10	3	8	6.801	6.641	2.668	2.668	3	3

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

CONTOH KASUS

Iterasi – 2

Langkah 4: Alokasikan semua data ke centroid terdekat

Data ke-i	Fitur x	Fitur y	Kelompok 1	Kelompok 2	Kelompok 3
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			*

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

CONTOH KASUS

Iterasi – 3

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 1:

Data ke-i	Fitur x	Fitur y
1	1	1
4	1	2
5	2	3
M (Jumlah Data) 3	Jumlah x 4	Jumlah y 6
Rata-rata	1.333	2

Centroid untuk kelompok 2:

Data ke-i	Fitur x	Fitur y
2	4	1
3	6	1
6	5	3
M (Jumlah Data) 3	Jumlah x 15	Jumlah y 5
Rata-rata	5	1.667

CONTOH KASUS

Iterasi – 3

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 3:

Data ke-i	Fitur x	Fitur y
7	2	5
8	3	5
9	2	6
10	3	8
M (Jumlah Data)	Jumlah x	Jumlah y
4	10	24
Rata-rata	2.500	6

Centroid keempat untuk setiap kelompok (hasil iterasi – 3):

Kelompok	Fitur x	Fitur y
1	1.333	2
2	5	1.667
3	2.500	6

CONTOH KASUS

Iterasi – 3

Menghitung fungsi objektif baru

Jarak data ke centroid ketiga (menggunakan jarak *euclidean*):

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	K
1	1	1	1.054			1
2	4	1		1.202		2
3	6	1		1.202		2
4	1	2	0.333			1
5	2	3	1.202			1
6	5	3		1.333		2
7	2	5			1.118	3
8	3	5			1.118	3
9	2	6			0.500	3
10	3	8			2.062	3
			2.589	3.737	4.798	11.124
			Perubahan Fungsi Objektif			0.732

CONTOH KASUS

Iterasi – 3

Menghitung perubahan fungsi objektif

Fungsi Objektif (J) lama/sebelumnya = 11.846

Fungsi Objektif (J) baru/sekarang = 2.589 + 3.737 + 4.798 = 11.124

Perubahan Fungsi Objektif = | J baru – J lama |
= | 11.124 – 11.846 | = 0.732

Perubahan masih di atas ambang batas atau threshold (T); >0.1, artinya pencarian centroid masih terus dilakukan.

CONTOH KASUS

Iterasi – 3

Langkah 4: Alokasikan semua data ke centroid terdekat

(menggunakan jarak *euclidean*)

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	Terdekat	K Baru	K Lama
1	1	1	1.054	4.055	5.220	1.054	1	1
2	4	1	2.848	1.202	5.220	1.202	2	2
3	6	1	4.773	1.202	6.103	1.202	2	2
4	1	2	0.333	4.014	4.272	0.333	1	1
5	2	3	1.202	3.283	3.041	1.202	1	1
6	5	3	3.801	1.333	3.905	1.333	2	2
7	2	5	3.073	4.484	1.118	1.118	3	3
8	3	5	3.432	3.887	1.118	1.118	3	3
9	2	6	4.055	5.270	0.500	0.500	3	3
10	3	8	6.227	6.641	2.062	2.062	3	3

Sudah tidak ada data yang berpindah kelompok, artinya pencarian centroid sudah bisa dihentikan.

Tetapi iterasi juga masih bisa dilakukan (kembali ke langkah 3), karena perubahan fungsi objektif masih di atas ambang batas atau threshold (T); 0.732 > 0.1

CONTOH KASUS

Iterasi – 4

Menghitung fungsi objektif baru

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	K
1	1	1	1.054			1
2	4	1		1.202		2
3	6	1		1.202		2
4	1	2	0.333			1
5	2	3	1.202			1
6	5	3		1.333		2
7	2	5			1.118	3
8	3	5			1.118	3
9	2	6			0.500	3
10	3	8			2.062	3
			2.589	3.737	4.798	11.124
			Perubahan Fungsi Objektif			0.000

CONTOH KASUS

Iterasi – 4

Menghitung perubahan fungsi objektif

Fungsi Objektif (J) lama/sebelumnya = 11.124

Fungsi Objektif (J) baru/sekarang = 2.589 + 3.737 + 4.798 = 11.124

Perubahan Fungsi Objektif = | J baru – J lama |
= | 11.846 – 11.124 | = 0

Perubahan sudah di bawah ambang batas atau threshold (T); 0<0.1, artinya centroid sudah bisa digunakan.

CONTOH KASUS

Nilai akhir centroid untuk setiap kelompok:

Kelompok	Fitur x	Fitur y
1	1.333	2
2	5	1.667
3	2.500	6

Hasil pengalokasian semua data latih ke centroid terdekat

Data ke-i	Fitur x	Fitur y	Kelompok 1	Kelompok 2	Kelompok 3
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			*

LATIHAN

Terapkan teknik clustering menggunakan algoritma k-means terhadap data latih di bawah ini:

Jumlah kelompok (K) = 2

No	Atribut 1	Atribut 2
1	1	3
2	4	2
3	5	7
4	8	6
5	9	11
6	12	10

NEXT

UAS