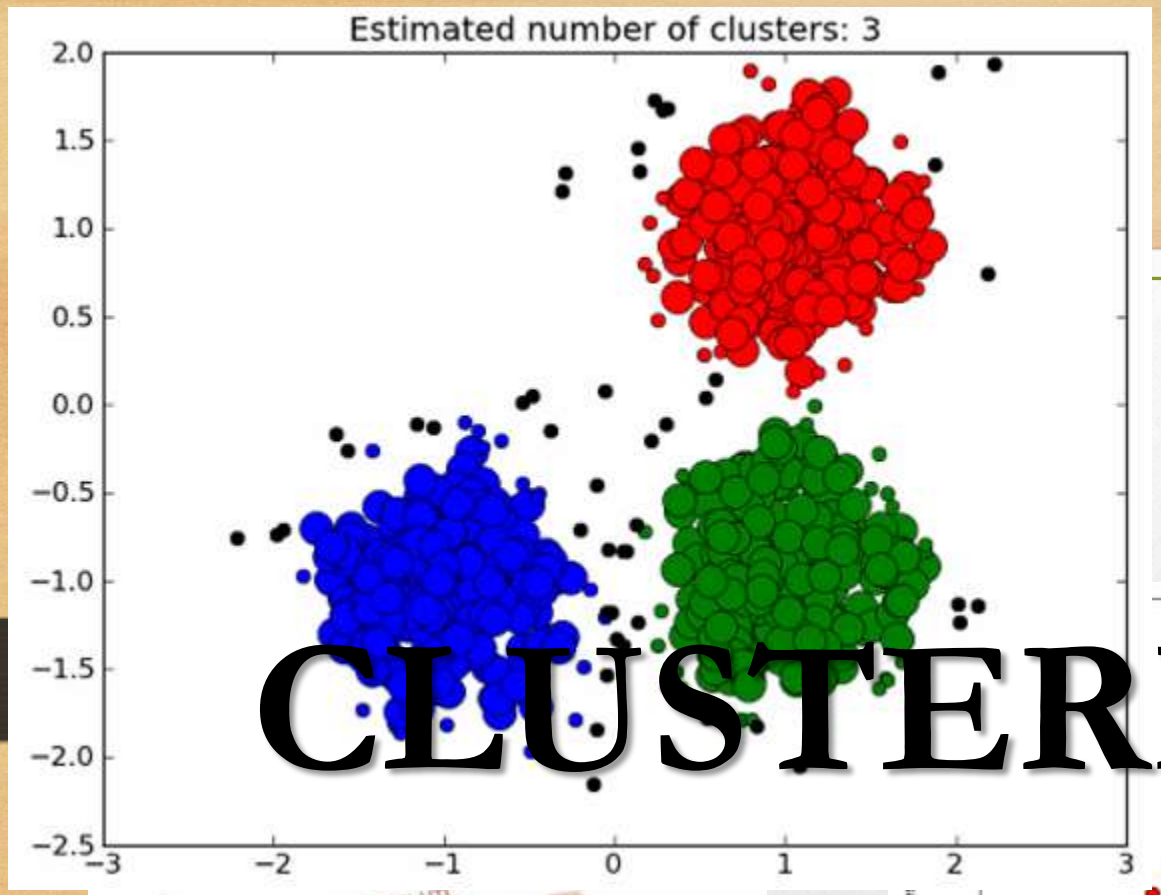
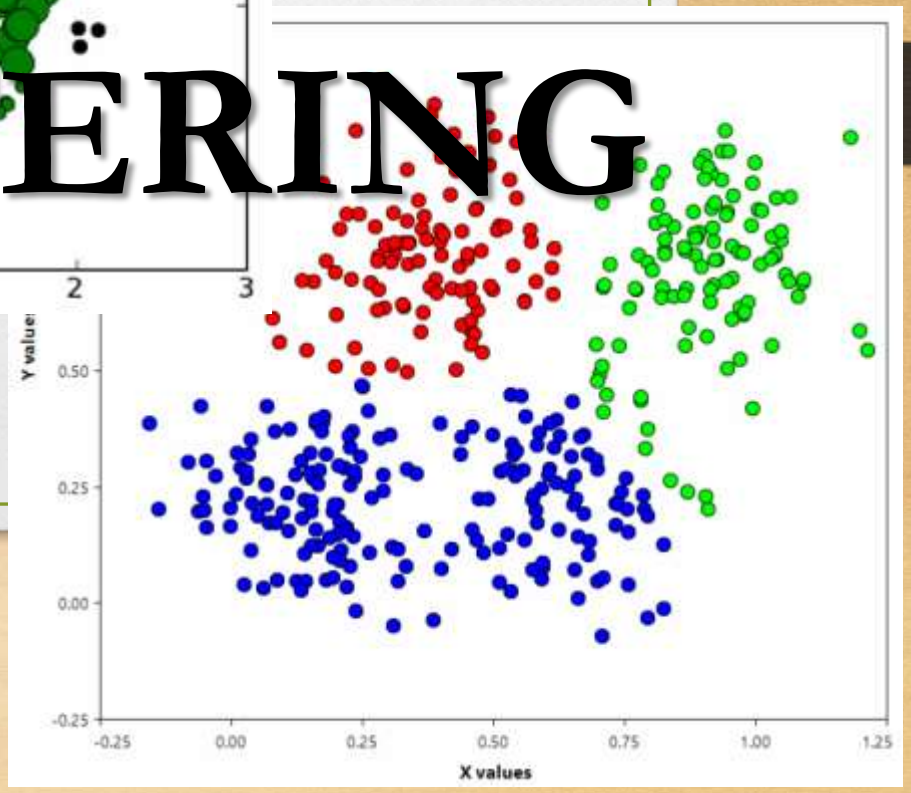

Estimated number of clusters: 3
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3
−2
−1
0
1
2
3

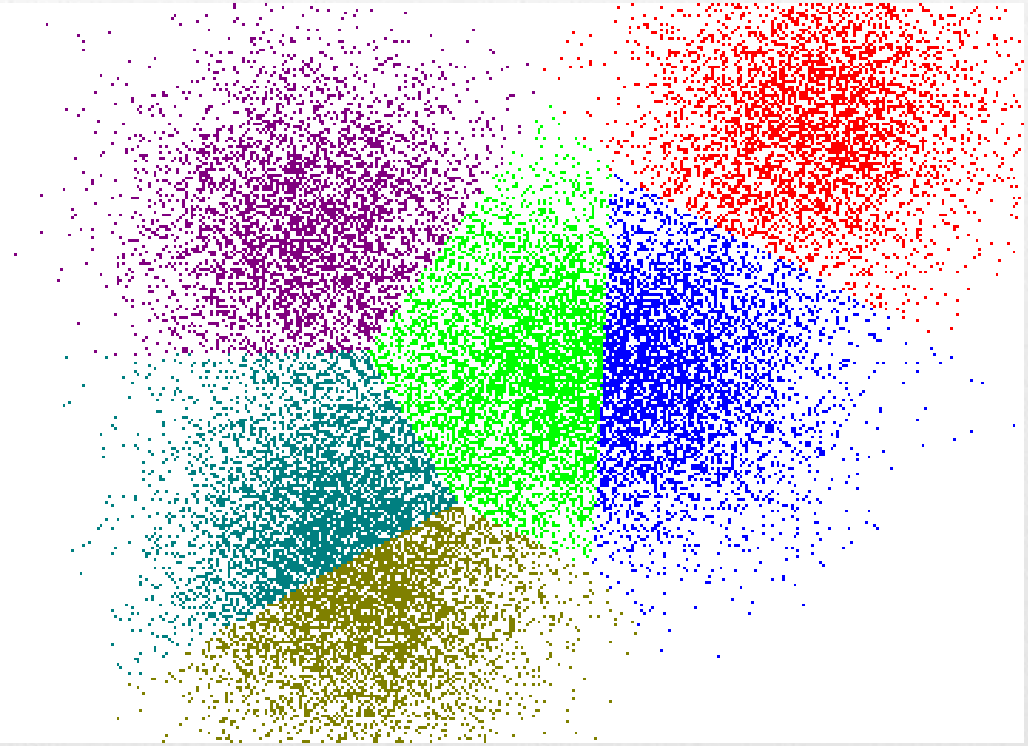
CLUSTERING


CLUSTER
ANALYSIS
NUMBER
SUM
COMPUTING
UNDERLYING
MARKET
OBVERSE


Y values
0.50
0.25
0.00
-0.25
-0.25
0.00
0.25
0.50
0.75
1.00
1.25
X values

KEMAMPUAN AKHIR YANG DIHARAPKAN

Dapat menjelaskan konsep dasar cluster dan penerapannya pada data.

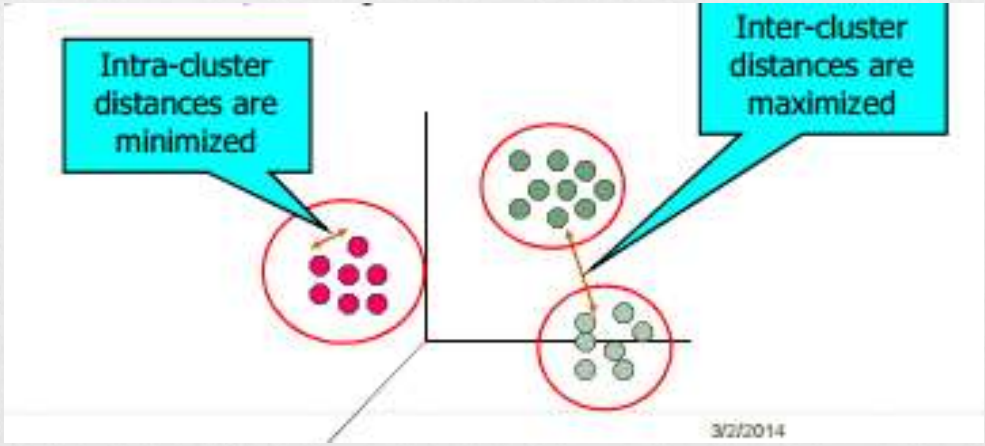


CLUSTERING

- Cluster

  Suatu cluster merupakan sekelompok entitas yang memiliki kesamaan dan memiliki perbedaan dengan entitas dari kelompok lain(Everitt,1980).

  Cluster berguna untuk **mengelompokkan objek-objek data yang memiliki kemiripan ke dalam satu grup dan yang berbeda dikelompokkan ke dalam grup lainnya**
- Semakin besar tingkat kemiripan/*similarity* (atau homogenitas) di dalam satu grup dan semakin besar tingkat perbedaan diantara grup, maka semakin baik (atau lebih berbeda) *clustering* tersebut.



Tujuan Pengelompokan

- Tujuan *clustering (pengelompokan) data dapat dibedakan menjadi dua, yaitu pengelompokan untuk* ***pemahaman dan clustering untuk penggunaan (Prasetyo,2012).***

- Biasanya proses pengelompokan untuk tujuan pemahaman hanya sebagai proses awal untuk kemudian dilanjutkan dengan pekerjaan seperti summarization (rata-rata, standar deviasi), pelabelan kelas untuk setiap kelompok sehingga dapat digunakan sebagai data training dalam klasifikasi supervised.
- Sementara jika untuk penggunaan, tujuan utama clustering biasanya adalah mencari prototipe kelompok yang paling representatif terhadap data, memberikan abstraksi dari setiap obyek data dalam kelompok dimana sebuah data terletak didalamnya

clustering untuk pemahaman

- Contoh tujuan clustering untuk pemahaman diantaranya: dibidang Biologi (pengelompokan berdasarkan karakter tertentu secara hirarkis) , pengelompokan gen yang memiliki fungsi sama.
- Dibidang information retrieval (web search),bidang klimatologi (pengelompokam pola tekanan udara yang berpengaruh pada cuaca), bidang bisnis (pengelompokan konsumen yang berpotensi untuk analisa dan strategi pemasaran).

clustering untuk penggunaan

- Contoh tujuan clustering untuk penggunaan dibidang *summarization, dengan semakin besarnya jumlah data maka ongkos melakukan peringkasan semakin mahal (berat&kompleks), maka perlu diterapkan pengelompokan data untuk membuat prototipe yang dapat mewakili keseluruhan data yang akan digunakan.*
- *Kompresi , data yang terletak dalam satu cluster dapat dikompresi dengan diwakili oleh indeks prototipe yang dikaitkan dengan kelompok , teknik kompresi ini dikenal sebagai quantization vector.*

Aplikasi Teknik Clustering

Clustering telah diterapkan diberbagai bidang seperti di jelaskan sebagai berikut:

- 1. Teknik
- Digunakan dalam bidang *biometric recognition & speech recognition, analisa sinyal radar, Information Compression,dan noise removal.*

- 2. Ilmu Komputer
- Web mining,analisa database spatial,*information retrieval,textual document collection,dan image segmentation.*

Aplikasi Teknik Clustering

- 3. Medis
- Digunakan dalam mendefinisikan taxonomi dalam bidang biologi, identifikasi fungsi protein dan gen, diagnosa penyakit dan penanganannya.

- 4. Astronomy
- Digunakan untuk mengelompokkan bintang dan planet, menginvestigasi formasi tanah, mengelompokkan wilayah /kota, digunakan dalam studi tentang sistem pada sungai dan gunung.

Aplikasi Teknik Clustering

- 5. Sosial
- Digunakan pada analisa pola perilaku,identifikasi hubungan diantara budaya yang berbeda, pembentukan sejarah evolusi bahasa, dan studi psikologi criminal.

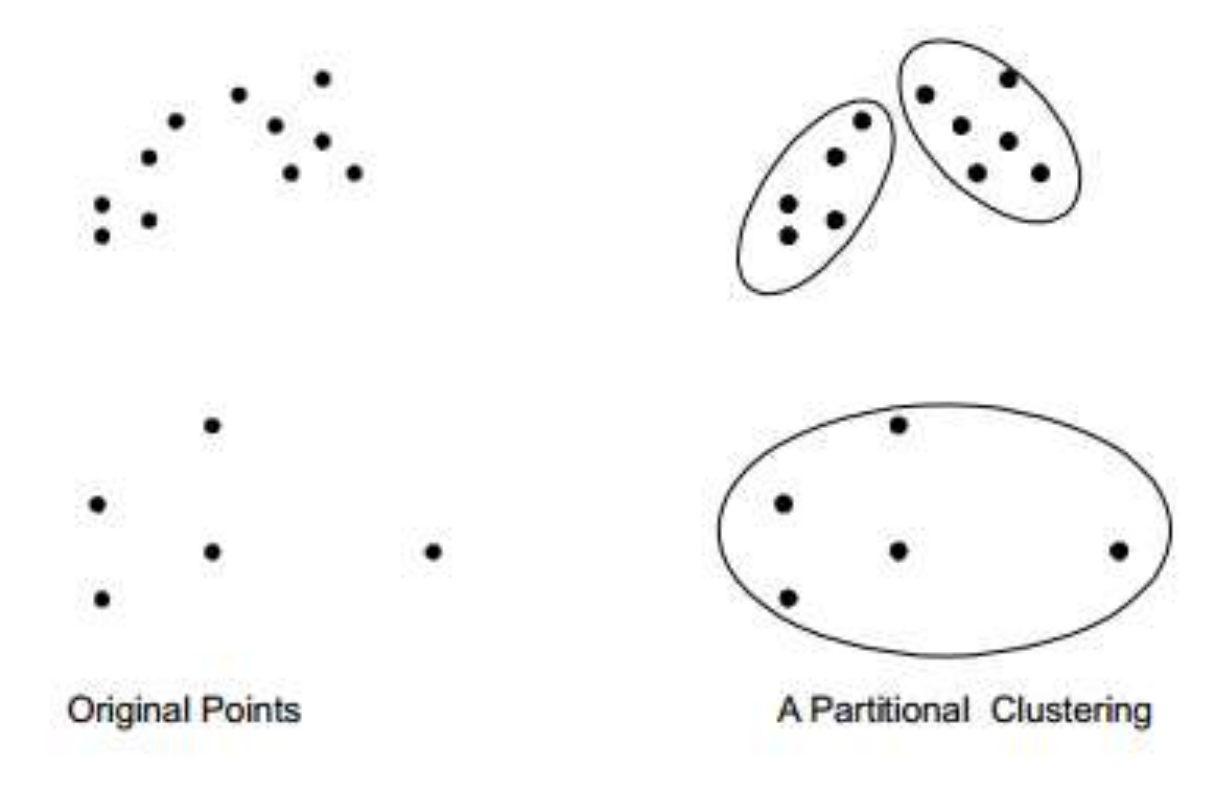
- 6. Ekonomi
- Penerapan pada pengenalan pola pembelian& karakteristik konsumen, pengelompokan perusahaan, analisa trend stok.

TIPE-TIPE *CLUSTERING*

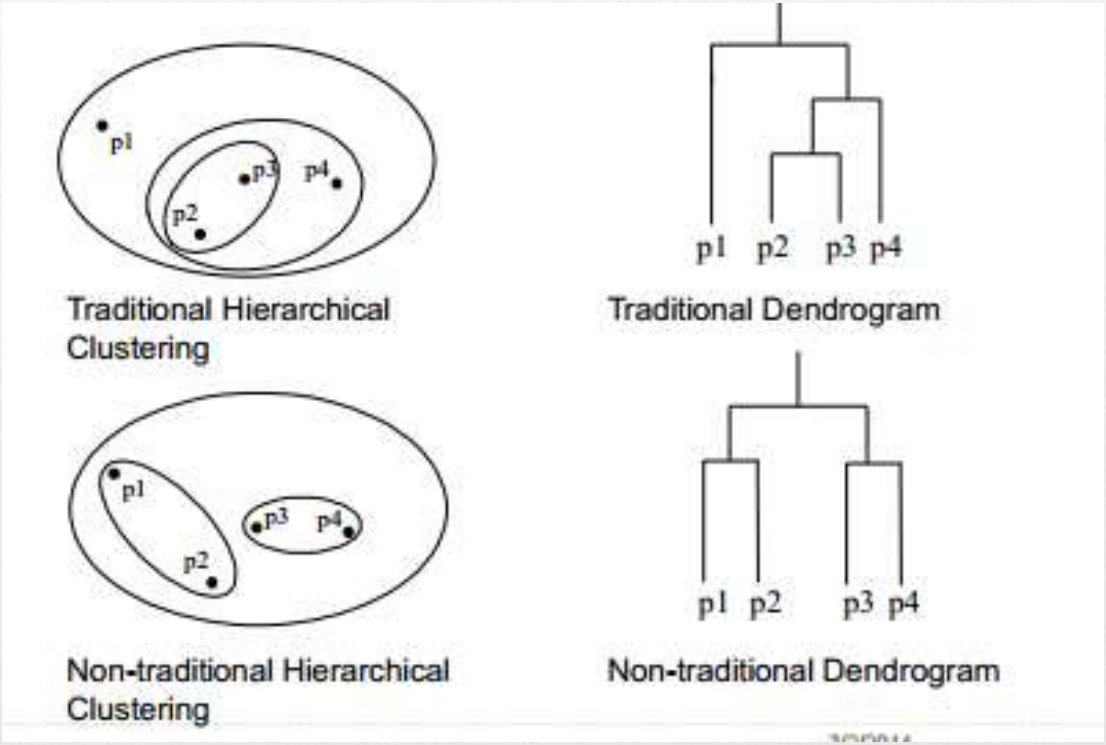
- *Partitional clustering* adalah himpunan obyek data ke dalam sub-himpunan (cluster) yang tidak overlap, sehingga setiap obyek data berada dalam tepat satu cluster.

- *Hierarchical clustering* adalah cluster yang memiliki subcluster. Himpunan cluster besarang yang diatur dalam tree.

PARTITIONAL CLUSTERING



HIERARCHICAL CLUSTERING



ALGORITMA CLUSTERING

- ***K-Means***
- ***K-Medoids***
- ***Hierarchical Clustering***

K-MEANS CLUSTERING

- K-Means clustering adalah metode untuk mengelompokkan item ke dalam kelompok (dimana k adalah jumlah kelompok yang diinginkan).

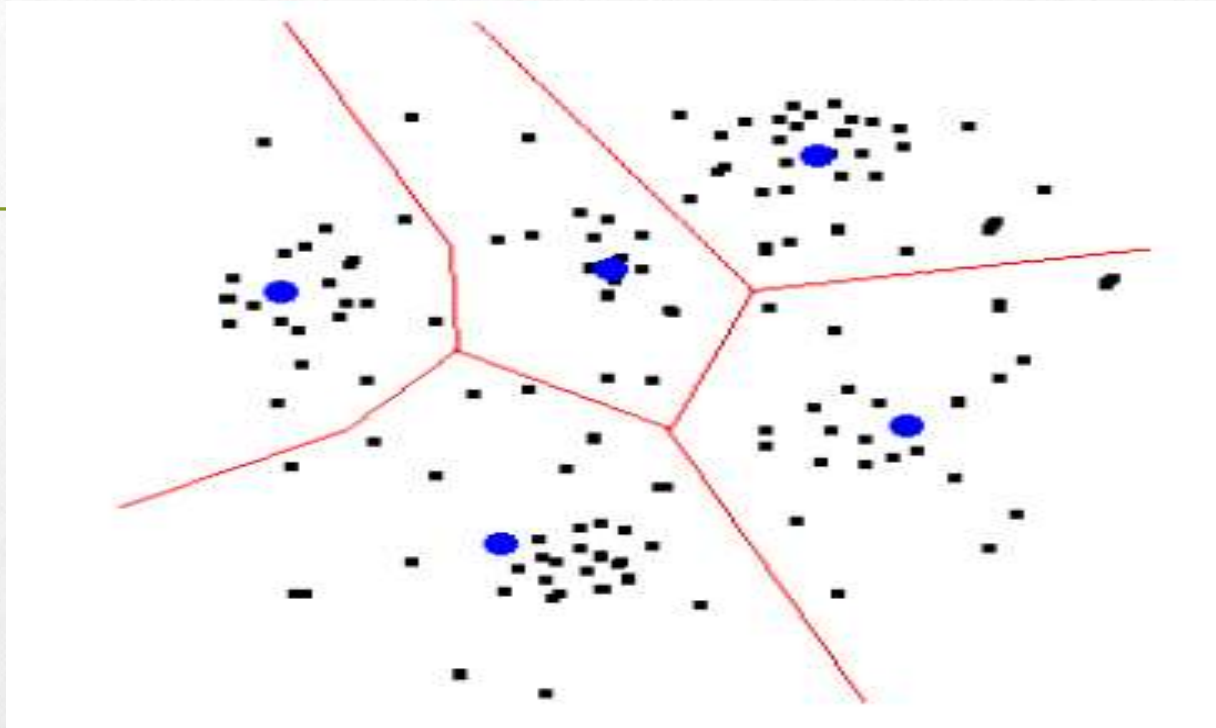
- Kelompok/cluster dibentuk dengan meminimalkan jumlah dari Euclidean distances) diantara data dengan titik pusat (centroid) yang berkorespondensi.

- Centroid adalah titik pusat data, dalam hal ini kita mengasumsikan rata-rata vector sebagai centroid.

K-MEANS CLUSTERING

- Pendekatan *partitional clustering*
- Setiap cluster diasosiasikan dengan sentroid
- Setiap titik di tandai ke *cluster* dengan sentroid terdekat
- K menandakan jumlah cluster yang akan terbentuk
- Algoritma *Clustering:*

1. Menentukan jumlah cluster
2. Menentukan nilai centroid biasanya dilakukan secara random atau biasanya menggunakan rumus rata-rata
3. Menghitung jarak antara titik centroid dengan titik tiap objek. Biasanya menggunakna jarak Euclidean distance.
4. Mengelompokkan objek berdasarkan jarak terdekat
5. Kembali ke tahap ke 2 dan lakukan perulangan hingga nilai centroid yang dihasilkan tetap dan anggota cluster tidak berpindah ke cluster lain.



- Titik hitam menyatakan data. Garis merah menyatakan partisi/pemisah. Titik biru merepresentasikan titik pusat (centroid) yang mendefinisikan suatu partisi

Inisialisasi titik pusat (centroid)

- Inisialisasi centroid dapat dilakukan dengan beberapa cara, contohnya 3 cara berikut:

– **Dipilih secara dinamik:** Metode ini tepat digunakan jika data baru ditambahkan secara cepat dan banyak. Untuk menyederhanakan persoalan, inisial cluster dipilih dari beberapa data baru, misal jika data dikelompokkan menjadi 3 clusters, maka inisial cluster berarti 3 item pertama dari data.

– **Dipilih secara random:** Paling banyak digunakan, dimana inisial cluster dipilih secara random dengan range data antara nilai terendah sampai nilai tertinggi.

– **Memilih dari batasan nilai tinggi dan rendah:** tergantung pada tipe datanya, nilai data tertinggi dan terendah dipilih sebagai inisial cluster. Contoh berikut menggunakan metode ini.

Diketahui : Jumlah Cluster = 3,

jumlah data =12,

jumlah atribut= 2

NO	Kota /Kab	Luas Lahan	Produksi
1	Kab. Ponorogo	66,693.00	402,047.00
2	Kab. Trenggalek	31,136.00	182,848.00
3	Kab. Tulungagung	49,230.00	259,581.00
4	Kab. Blitar	50,577.00	289,494.00
5	Kab. Kediri	51,083.00	281,392.00
6	Kab. Malang	65,597.00	464,498.00
7	Kab. Lumajang	72,552.00	387,168.00
8	Kab. Jember	162,619.00	964,001.00
9	Kab. Banyuwangi	113,609.00	706,419.00
10	Kab. Bondowoso	61,330.00	329,557.00
11	Kab. Situbondo	48,902.00	290,954.00
12	Kab. Probolinggo	59,130.00	311,258.00

CONTOH SOAL

DATA	X	Y
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

C1=(3,4) dan C2=(6,4)

CONTOH SOAL

Iterasi 1

a. Menghitung Euclidean distance dari semua data ke tiap titik pusat pertama

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

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

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

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

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

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

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

Dengan cara yang sama hitung jarak tiap titik ke titik pusat ke dan kita akan mendapatkan D_{21}= 4.12, D_{22}=4.27, D_{23}= 1.18, D_{24}= 1.86, D_{25}=1.22, D_{26}=2.62, D_{27}=2.06

CONTOH SOAL

Iterasi 1

b. Dari perhitungan Euclidean distance, kita dapat membandingkan

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

$\{M_1, M_2, M_6, M_7\}$ anggota C_1 and $\{M_3, M_4, M_5\}$ anggota C_2

CONTOH SOAL

Iterasi 1

c. Hitung titik pusat baru

$$C_1 = \left(\frac{2+2+3+4}{4}, \frac{5+5.5+4.8+4.5}{4}\right) = (2.75, 4.9)$$

$$C_2 = \left(\frac{5+6.5+7}{3}, \frac{3.5+2.2+3.3}{3}\right) = (6.17, 3)$$

Lakukan iterasi ke 2 seperti iterasi ke 1, sampai anggota kelompok C1 dan C2 tidak berubah lagi seperti sebelumnya,

Kesimpulan $\{M_1, M_2, M_6, M_7\}$ anggota C_1 dan $\{M_3, M_4, M_5\}$ anggota C_2

HIERARCHICAL CLUSTERING

Strategi pengelompokkannya umumnya ada dua jenis, yaitu:

- Agglomerative (Bottom-Up)
- Devisive (Top-Down)

Algoritma Agglomerative Hierarchical Clustering :

1. Hitung Matrik Jarak antar data.
2. Ulangi langkah 3 dan 4 higga hanya satu kelompok yang tersisa.
3. Gabungkan dua kelompok terdekat berdasarkan metode pengelompokan (*Single Linkage, Complete Linkage, Average Linkage)*
4. Perbarui Matrik Jarak antar data untuk merepresentasikan kedekatan diantara kelompok baru dan kelompok yang masih tersisa.
5. Selesai

Metode Pengelompokan Hierarki Aglomeratif

Beberapa metode pengelompokan secara hierarki Aglomeratif:

- Single Linkage (Jarak Terdekat)

$$d_{uv} = \min\{d_{uv}\}, d_{uv} \in D$$

- Complete Linkage (Jarak Terjauh)

$$d_{uv} = \max\{d_{uv}\}, d_{uv} \in D$$

- Average Linkage (Jarak rata-rata)

$$d_{uv} = average\{d_{uv}\}, d_{uv} \in D$$

CONTOH STUDI KASUS

Data	Fitur x	Fitur y
1	1	1
2	4	1
3	1	2
4	3	4
5	5	4

Kelompokkan dataset tersebut dengan menggunakan metode AHC (Single Linkage) menggunakan jarak Manhattan!

CONTOH STUDI KASUS

- Menghitung Jarak Pada Semua Pasangan dua data :

Data	Fitur x	Fitur y
1	1	1
2	4	1
3	1	2
4	3	4
5	5	4

$$D_{min}(Data_1, Data_1) = \sum_{j=1}^{2} |x_j - y_j| = |1-1| + |1-1| = 0$$

$$D_{min}(Data_1, Data_2) = |1-4| + |1-1| = 3$$

$$D_{min}(Data_1, Data_3) = |1-1| + |1-2| = 1$$

$$D_{min}(Data_1, Data_4) = |1-3| + |1-4| = 2+3 = 5$$

$$D_{min}(Data_1, Data_5) = |1-5| + |1-4| = 4+3 = 7$$

$$D_{min}(Data_2, Data_3) = |4-1| + |1-2| = 3+1 = 4$$

$$D_{min}(Data_2, Data_4) = |4-3| + |1-4| = 1+3 = 4$$

$$D_{min}(Data_2, Data_5) = |4-5| + |1-4| = 1+3 = 4$$

$$D_{min}(Data_3, Data_4) = |1-3| + |2-4| = 2+2 = 4$$

$$D_{min}(Data_3, Data_5) = |1-5| + |2-4| = 4+2 = 6$$

$$D_{min}(Data_4, Data_5) = |3-5| + |4-4| = 2+0 = 2$$

CONTOH STUDI KASUS

❖

Hasil Matrik Jarak :

Dman	1	2	3	4	5
1	0	3	1	5	7
2	3	0	4	4	4
3	1	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0

- Menggunakan Metode Single Linkage :

  Dengan memperlakukan data sebagai kelompok, selanjutnya kita pilih jarak dua kelompok yang terkecil.

Dman	1	2	3	4	5
1	0	3	1	5	7
2	3	0	4	4	4
3	1	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0

$$\min(D_{man}) = \min(d_{13}) = 1$$

terpilih kelompok 1 dan 3, sehingga kedua kelompok ini digabungkan. (Melanjutkan pengelompokan).

- Menghitung jarak antar kelompok (1 dan 3) dengan kelompok lain yang tersisa, yaitu 2, 4 dan 5.

$$d_{(13)2} = \min\{d_{12}, d_{32}\} = \min\{3,4\} = 3$$

$$d_{(13)4} = \min\{d_{14}, d_{34}\} = \min\{5,4\} = 4$$

$$d_{(13)5} = \min\{d_{15}, d_{35}\} = \min\{7,6\} = 6$$

- Menghitung jarak antar kelompok (1 dan 3) dengan kelompok lain yang tersisa, yaitu 2, 4 dan 5.

$$d_{(13)2} = \min\{d_{12}, d_{32}\} = \min\{3,4\} = 3$$
$$d_{(13)4} = \min\{d_{14}, d_{34}\} = \min\{5,4\} = 4$$
$$d_{(13)5} = \min\{d_{15}, d_{35}\} = \min\{7,6\} = 6$$

- Dengan menghapus baris-baris dan kolom-kolom matrik jarak yang bersesuaian dengan kelompok 1 dan 3, serta menambahkan baris dan kolom untuk kelompok (13)

Dman	1	2	3	4	5
1	0	3	1	5	7
2	3	0	4	4	4
3	1	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0

→

Dman	(13)	2	4	5
(13)	0	3	4	6
2	3	0	4	4
4	4	4	0	2
5	6	4	2	0

Selanjuttnya dipilih jarak dua kelompok yang terkecil.

$$\min(D_{man}) = \min(d_{45}) = 2$$

- Dengan menghapus baris-baris dan kolom-kolom matrik jarak yang bersesuaian dengan kelompok 1 dan 3, serta menambahkan baris dan kolom untuk kelompok (13).

Dman	1	2	3	4	5
1	0	3	1	5	7
2	3	0	4	4	4
3	1	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0

Dman	(13)	2	4	5
(13)	0	3	4	6
2	3	0	4	4
4	4	4	0	2
5	6	4	2	0

- Selanjutnya dipilih jarak dua kelompok yang terkecil.

$$\min(D_{man}) = \min(d_{45}) = 2$$

- Menghitung jarak antar kelompok (4 dan 5) dengan kelompok lain yang tersisa, yaitu (13) dan 2.

$$d_{(45)(13)} = \min\{d_{41}, d_{43}, d_{51}, d_{53}\} = \min\{5,4,7,6\} = 4$$

$$d_{(45)2} = \min\{d_{42}, d_{52}\} = \min\{4,4\} = 4$$

- Menghapus baris dan kolom matrik yang bersesuaian dengan kelompok 4 dan 5, serta menambahkan baris dan kolom untuk kelompok (45)

- Menghitung jarak antar kelompok (4 dan 5) dengan kelompok lain yang tersisa, yaitu (13) dan 2.

$$d_{(45)(13)} = \min\{d_{41}, d_{43}, d_{51}, d_{53}\} = \min\{5,4,7,6\} = 4$$

$$d_{(45)2} = \min\{d_{42}, d_{52}\} = \min\{4,4\} = 4$$

- Menghapus baris dan kolom matrik yang bersesuaian dengan kelompok 4 dan 5, serta menambahkan baris dan kolom untuk kelompok (45)

Dman	(13)	2	4	5
(13)	0	3	4	6
2	3	0	4	4
4	4	4	0	2
5	6	4	2	0

Dman	(45)	(13)	2
(45)	0	4	4
(13)	4	0	3
2	4	3	0

- Selanjutnya dipilih jarak dua kelompok yang terkecil.

$$\min(D_{man}) = \min(d_{(13)2}) = 3$$

terpilih kelompok (13) dan 2, sehingga kedua kelompok ini digabungkan. (Melanjutkan pengelompokan).

- Menghitung jarak antar kelompok ((13) dan 2) dengan kelompok lain yang tersisa, yaitu (45).

- Menghitung jarak antar kelompok ((13) dan 2) dengan kelompok lain yang tersisa, yaitu (45).

$$d_{(132)(45)} = \min\{d_{14}, d_{15}, d_{34}, d_{35}, d_{24}, d_{25}\} = \min\{5,7,4,6,4,4\} = 4$$

- Menghapus baris dan kolom matrik yang bersesuaian dengan kelompok (13) dan 2, serta menambahkan baris dan kolom untuk kelompok (123)

Dman	1	2	3	4	5
1	0	3	1	5	7
2	3	0	4	4	4
3	1	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0

Dman	(13)	2	4	5
(13)	0	3	4	6
2	3	0	4	4
4	4	4	0	2
5	6	4	2	0

Dman	(45)	(13)	2
(45)	0	4	4
(13)	4	0	3
2	4	3	0

Dman	(132)	(45)
(132)	0	4
(45)	4	0

- Jadi kelompok (132) dan (45) digabung untuk menjadi kelompok tunggal dari lima data, yaitu kelompok (13245) dengan jarak terdekat 4.

- Menghapus baris dan kolom matrik yang bersesuaian dengan kelompok (13) dan 2, serta menambahkan baris dan kolom untuk kelompok (132).

Dman	1	2	3	4	5
1	0	3	1	5	7
2	3	0	4	4	4
3	1	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0

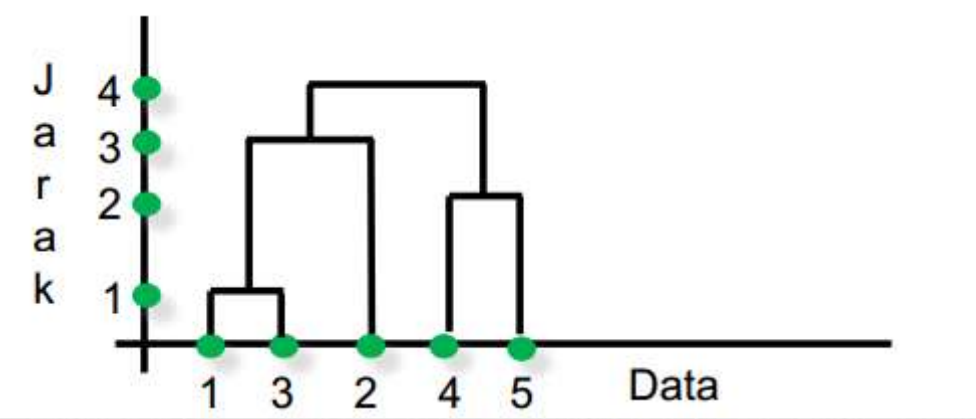
Dman	(13)	2	4	5
(13)	0	3	4	6
2	3	0	4	4
4	4	4	0	2
5	6	4	2	0

Dman	(45)	(13)	2
(45)	0	4	4
(13)	4	0	3
2	4	3	0

Dman	(132)	(45)
(132)	0	4
(45)	4	0

- Jadi kelompok (132) dan (45) digabung untuk menjadi kelompok tunggal dari lima data, yaitu kelompok (13245) dengan jarak terdekat 4. Berikut

  Dendogram Hasil Metode Single Linkage :



Tugas

- Latihan: Gunakan metode k-means untuk mengelompokkan mahasiswa berdasarkan tinggi & berat badan:

Data Mahasiswa

No	Nama	Tinggi Badan (Cm)	Berat Badan (Kg)
1	Agus	170	70
2	Arif	180	75
3	Iwan	168	80
4	Yasinta	160	60
5	Esti	165	65
7	Bayu	172	80
8	Beno	175	70
9	Ramadhan	168	60
10	Indah	160	60

- Clustering yang diharapkan mampu menghasilkan kelompok mahasiswa yang memenuhi sifat berikut :

- Mahasiswa yang memiliki berat dan tinggi badan yang hampir sama dikelompokkan tersendiri

- 1. Mahasiswa yang memiliki berat dan tinggi badan yang hampir sama dikelompokkan akan berada pada kelompok yang sama.

- 2. Mahasiswa yang yang memiliki berat dan tinggi badan yang berbeda akan berada pada kelompok yang berbeda.

TERIMA KASIH