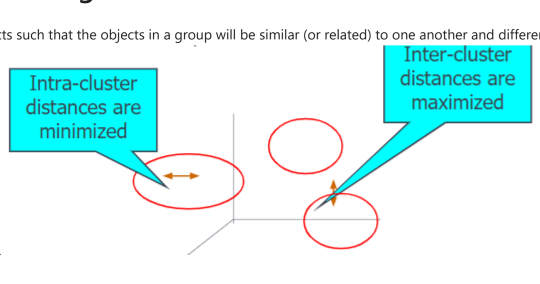
What is Clustering?

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

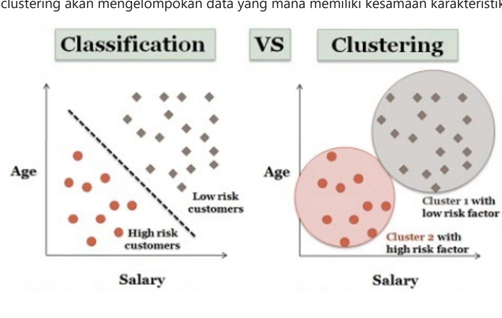


Different Between Classification and Clustering

Perbedaan antara clustering dengan klasifikasi yaitu dalam klasifikasi telah terdapat input dan output untuk memprediksi sesuatu, sedangkan pada clustering hanya terdapat input yang mana model akan mengelompokan data-data input berdasarkan kesamaan karakteristik.

Misalnya seperti contoh kasus di bawah ini: Kasus apakah customer jika diberi pinjaman masuk ke dalam kategori berisiko rendah atau tinggi. Jika di dalam klasifikasi telah dibuat classifier berdasarkan data historis yang mampu menentukan jika umur sekian dan gaji sekian masuk ke kategori yang mana (risiko rendah atau tinggi).

Sedangkan Clustering hanya memiliki input berupa umur dan gaji, belum memiliki output yang berupa classifier atau garis pemisah yang menentukan risiko rendah atau tinggi. Maka model clustering akan mengelompokan data yang mana memiliki kesamaan karakteristik, dalam kasus ini memiliki kesamaan gaji dan umur.



Clustering Algorithm

1. K-Means Clustering
2. K-Medoids Clustering
3. Hierarchical Clustering
4. Density-based Clustering
5. Fuzzy Clustering
6. Biclustering

etc

> Clustering yang umum digunakan di industri adalah K-Means Clustering dan Hierarchical Clusting.

Apllication

1. Customer Segmentation: Mengelompokan karakteristik dari customer menggunakan clustering. Salah satu tujuannya untuk memaksimalkan promosi-promosi yang diberikan. Agar promosi yang diberikan sesuai dengan karakteristik customer.
2. Fraud/Criminal Activity Identification

etc

K-Means Clustering

K-Means Clustering memiki konsep yang sama seperti KNN (Supervised). Pebedaannya terletak pada jika di KNN, K nya berupa jumlah tetangga terdekat. Sedangkan K-Means, K nya merupakan jumlah cluster(kelompok).

Jadi sebelum melakukan clustering, dengan metode K-Means kita harus menentukan terlebih dahulu berapa jumlah K atau kelompok yang ingin kita buat. Misalnya dengan kasus Customer Segmentation maka K nya kita tentukan terlebih dahulu, contohnya K=3 maka segmentasi pelanggan akan dibagi menjadi 3 kelompok.

- Number of clusters, K, must be specified
- Each cluster is associated with a centroid (center point). Misalnya kita ingin membagi kelompok customer loyal dengan customer tidak loyal. Maka titik tengah antara keduanya disebut centroid berdasarkan rata-rata (mean).
- Each point is assigned to the cluster with the closest centroid
- The basic algorithm is very simple

Step by step:

```
1: Select K points as the initial centroids.
2: repeat
3:    Form K clusters by assigning all points to the closest centroid.
4:    Recompute the centroid of each cluster.
5: until The centroids don't change
```

Jarak

Dalam menghitung jarak antara data input dengan centroid, sama seperti KNN.

- Untuk data numerik maka menggunakan Euclidean, Manhattan, Minkowski. Sedangkand data kategorik menggunakan Hamming Distance.

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

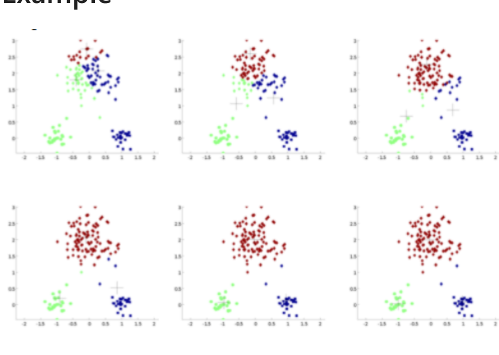
Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Hamming Distance

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$
$$x \neq y \Rightarrow D = 1$$

Example



Kondisi saat Centroidnya tidak berubah atau data yang masuk kedalam suatu kelompok tidak mengalami perubahan maka disebut konvergen yang mana clustering dianggap selesai.

Contoh k-means

Misalkan ada dua variabel X1 dan X2 yang tiap objeknya diberi nama A, B, C dan D. Datanya sebagai berikut:

Objek	Pengamatan	
	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

1. Dikelompokkan ke dalam 2 kelompok.
2. Centroid dipilih secara acak : c1 = (2, 2) dan c2 = (-1, -2).
3. Jarak yang digunakan jarak Euclidian. Memasukkan objek ke gerombol berpatokan pada jarak terdekat Diperoleh matriks jarak sbb :

	c_1	c_2
A	$(5-2)^2 + (3-2)^2 = 10$	$(5+1)^2 + (3+2)^2 = 61$
B	$(-1-2)^2 + (1-2)^2 = 10$	$(-1+1)^2 + (1+2)^2 = 9$
C	$(1-2)^2 + (-2-2)^2 = 17$	$(1+1)^2 + (-2+2)^2 = 4$
D	$(-3-2)^2 + (-2-2)^2 = 41$	$(-3+1)^2 + (-2+2)^2 = 4$

- objek A jaraknya lebih dekat dengan centroid1 maka akan dikelompokan ke centroid1
- objek B jaraknya lebih dekat dengan centroid2 maka akan dikelompokan ke centroid2
- objek c jaraknya lebih dekat dengan centroid2 maka akan dikelompokan ke centroid2
- objek d jaraknya lebih dekat dengan centroid2 maka akan dikelompokan ke centroid2

1. Hitung centroid baru, rataan dari vektor masing-masing unsur.

c1 = (5, 3)

c2 = [(-1, 1) + (1, -2) + (-3, -2)]/3 = (-1, -1)

Diperoleh matriks yang sbb :

	c_1	c_2
A	$(5-5)^2 + (3-3)^2 = 0$	$(5+1)^2 + (3+1)^2 = 52$
B	$(-1-5)^2 + (1-3)^2 = 40$	$(-1+1)^2 + (1+1)^2 = 4$
C	$(1-5)^2 + (-2-3)^2 = 41$	$(1+1)^2 + (-2+1)^2 = 5$
D	$(-3-5)^2 + (-2-3)^2 = 89$	$(-3+1)^2 + (-2+1)^2 = 5$

- objek A jaraknya lebih dekat dengan centroid1 maka akan dikelompokan ke centroid1
- objek B jaraknya lebih dekat dengan centroid2 maka akan dikelompokan ke centroid2
- objek c jaraknya lebih dekat dengan centroid2 maka akan dikelompokan ke centroid2
- objek d jaraknya lebih dekat dengan centroid2 maka akan dikelompokan ke centroid2

Karena tidak ada diperubahan maka 2 cluster telah ditetapkan (konvergen) : C1 = {A} dan C2 = {B, C, D}.

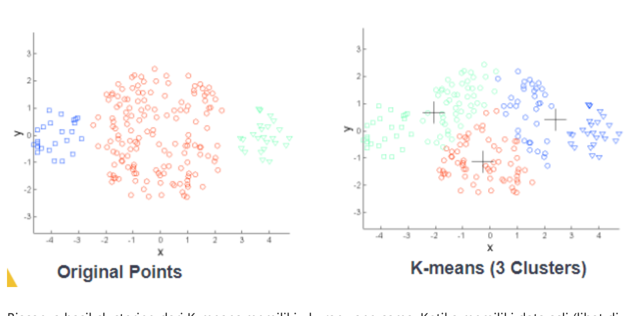
Limitations of K-means

1. K-means has problems when clusters are of differing

> Sizes: K-means tidak bisa mengolah ukuran kelompok yang berbeda
>
> Densities: K-means tidak bisa mengolah kepadatan yang berbeda
>
> Non-globular shapes: bentuknya tidak lingkaran

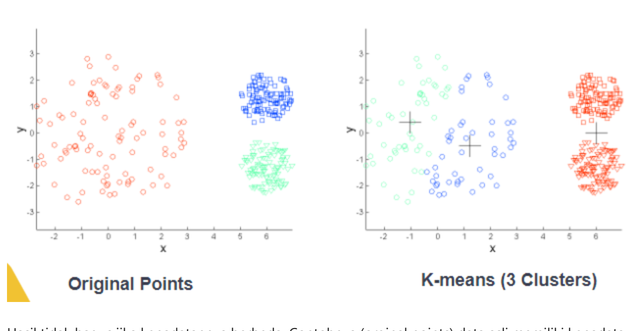
1. K-means has problems when the data contains outliers.

Limitations of K-means: Different Size



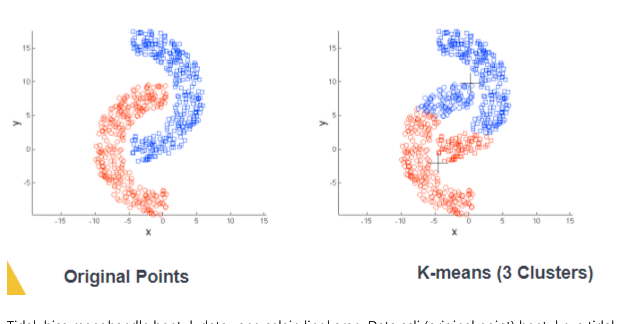
Biasanya hasil clustering dari K-means memiliki ukuran yang sama. Ketika memiliki data asli (lihat di original points), ketika menggunakan k-means maka sizenya akan sama. Sehingga tidak bisa mengelompokan dengan benar.

Limitations of K-means: Differing Density



Hasil tidak bagus jika kepadatannya berbeda. Contohnya (orginal points) data asli memiliki kepadatan yang berbeda. Data kelompok merah memiliki kepadatan yang renggang. Pada saat di olah dengan k-means akan mengclustering ke kepadatan yang sama atau kelompok yang berbeda. Sehingga tidak dapat mengelompokan dengan benar.

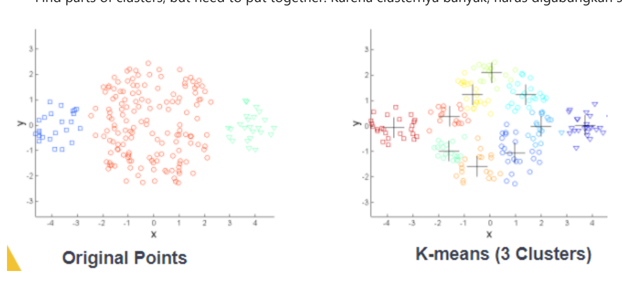
Limitations of K-means: Non-globular Shapes



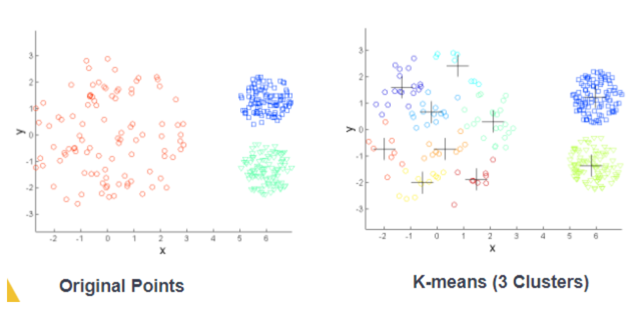
Tidak bisa menghandle bentuk data yang selain lingkaran. Data asli (original point) bentuknya tidak lingkaran sehingga pada saat di oleh dengan K-means akan memprediksi ke kelompok yang tidak sesuai dengan data aslinya.

Overcoming K-means Limitations

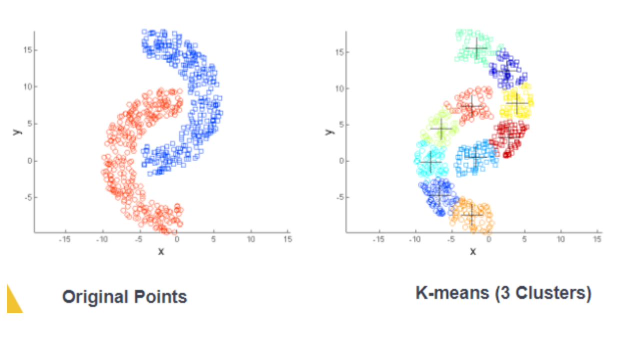
- One solution is to use many clusters. Misalnya aslinya terdapat 3 cluster, maka harus dibuat lebih banyak lagi.
- Find parts of clusters, but need to put together. Karena clusternya banyak, harus digabungkan sesuai dengan keiingan kita



Overcoming K-means Limitations



Overcoming K-means Limitations



Solutions to Initial Centroids Problem

Jika salah menentukan initial centroid maka menyebabkan model tidak bagus.

Solusinya:

1. Multiple runs dan dibandingkan mana yang menghasilkan initial centroid terbaik

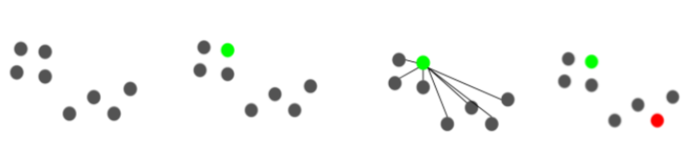
> Helps, but probability is not on your side

1. Select more than k initial centroids and then select among these initial centroids

> Select most widely separated

K-means ++

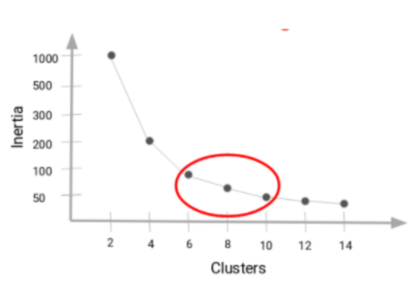
Cara kedua untuk mengatasi kesalaha initial centroid.



- Paling kiri, data input yang belum memiliki kelompok. Kemudian menentukan jumlah K, dalam kasus K nya 2 sehingga centroidnya 2
- Di K-means ++ dalam memilih initial centroid tidak melalui random, tetapi dipilih satu-persatu. Pemilihan berdasarkan data. Data mana yang akan menjadi centroid pertama.
- Untuk menentukan centroid kedua gimana?, pertama kita menghitung jarak centroid pertama dengan masing-masing data. Data terjauh dari centroid pertama akan dijadikan centroid kedua.

How to Choose the Right Number of Clusters

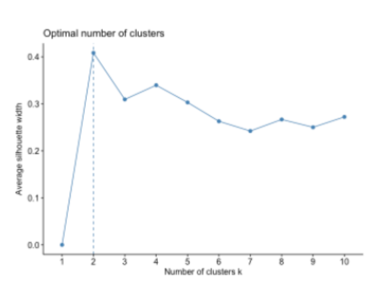
Inertia yang semakin kecil, semakin bagus



Cara terbaik dalam menentukan jumlah K atau cluster adalah dengan tunning hyperparameter. Jika K nya 2 Inertianya berapa, jika K nya 3 Inertianya berapa. dan seterusnya. Cara memilihnya, pilih K yang ketika turun tidak drastis. Selain itu juga berdasarkan intuitif.

Cara kedua

Menggunakan siluet. Semakin tinggi semakin bagus



Pros

1. Relatively simple to implement
2. Scales to large data sets
3. Guarantees convergence
4. Can warm-start the positions of centroids

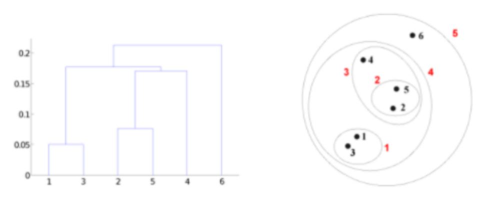
Cons

1. Choosing k manually
2. Being dependent on initial values
3. Clustering data of varying sizes and density
4. Clustering outliers
5. Scaling with number of dimensions

Hierarchical Clustering

1. Produces a set of nested clusters organized as a hierarchical tree
2. Can be visualized as a dendrogram

> A tree like diagram that records the sequences of merges or splits



Strengths of Hierarchical Clustering

1. Do not have to assume any particular number of clusters (tidak perlu menentukan jumlah cluster)

> Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level

1. They may correspond to meaningful taxonomies (mudah di interpretasi)

> Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, ...)

Hierarchical Clustering

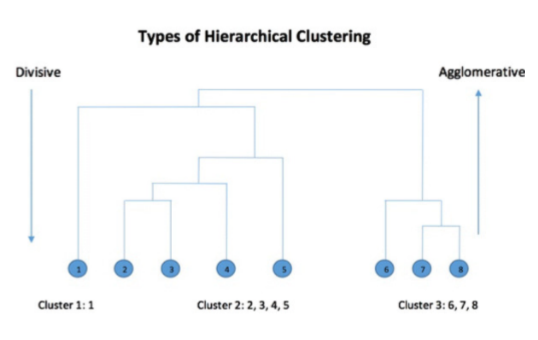
Two main types of hierarchical clustering

1. Agglomerative:

> Start with the points as individual clusters. Awalnya masing-masing data membentuk cluster sendiri. Misalnya kita mempunyai 10 data. Data 1 cluster 1, data 2 cluster 2 dll.
>
> At each step, merge the closest pair of clusters until only one cluster (or k clusters) left. Dari masing-masing cluster digabung-gabungkan yang pada akhirnya menjadi 1 cluster yang sama.

1. Divisive: (kebalikan dari Agglomerative)

> Start with one, all-inclusive cluster
>
> At each step, split a cluster until each cluster contains an individual point (or there are k clusters)



Agglomerative lebih banyak digunakanan

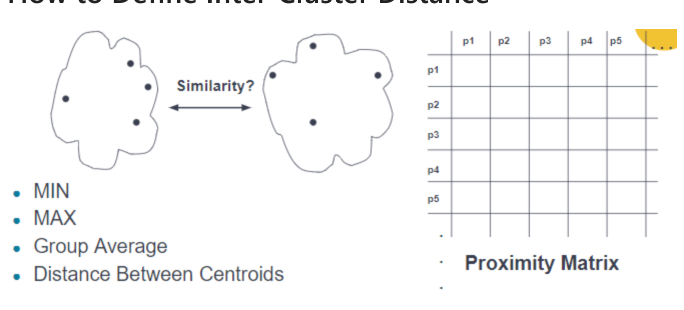
Agglomerative Clustering Algorithm

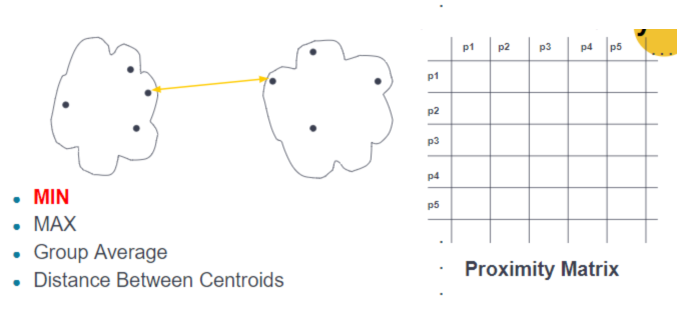
- Most popular hierarchical clustering technique
- Basic algorithm is straightforward

1. Compute the proximity matrix. Menghitung jarak antar data.
2. Let each data point be a cluster. Kemudian digabungkan cluster/data yang memiliki jarak terdekat.
3. Repeat. Di hitung ulang seperti step 1 dan 2. Sampai terbentuk satu cluster yang sama.
4. Merge the two closest clusters
5. Update the proximity matrix
6. Until only a single cluster remains

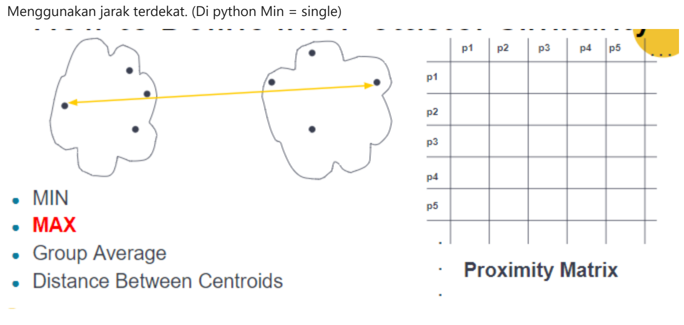
- Key operation is the computation of the proximity of two clusters

How to Define Inter-Cluster Distance

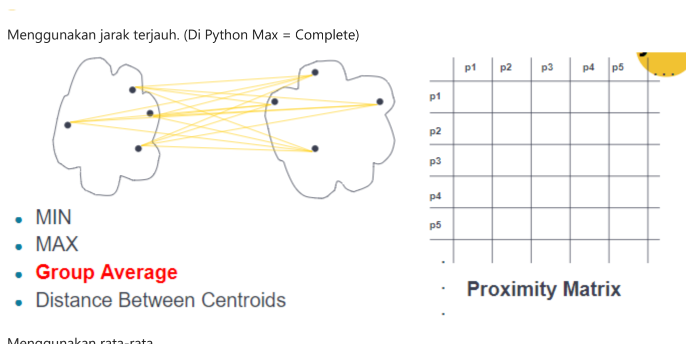




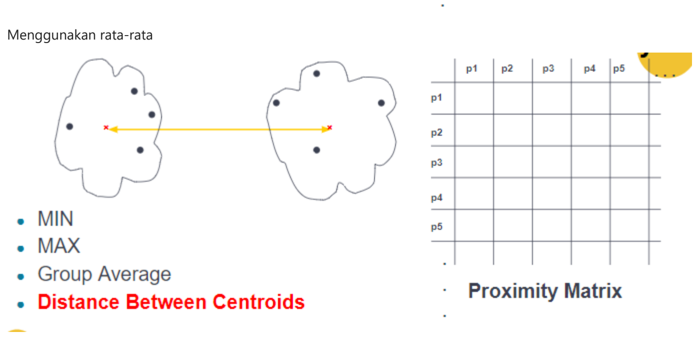
Menggunakan jarak terdekat. (Di python Min = single)



Menggunakan jarak terjauh. (Di Python Max = Complete)



Menggunakan rata-rata



Sama seperti K-means menggunakan centroid

Contoh Hierarchical Clustering

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

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

$D_{man}(Data1, Data1) = |1-1| + |1-1| = 0$
$D_{man}(Data1, Data2) = |1-4| + |1-1| = 3$
$D_{man}(Data1, Data3) = |1-1| + |1-2| = 1$
$D_{man}(Data1, Data4) = |1-3| + |1-4| = 5$
$D_{man}(Data1, Data5) = |1-5| + |1-4| = 7$
$D_{man}(Data2, Data3) = |4-1| + |1-2| = 4$
$D_{man}(Data2, Data4) = |4-3| + |1-4| = 4$
$D_{man}(Data2, Data5) = |4-5| + |1-4| = 4$
$D_{man}(Data3, Data4) = |1-3| + |2-4| = 4$
$D_{man}(Data3, Data5) = |1-5| + |2-4| = 6$
$D_{man}(Data4, Data5) = |3-5| + |4-4| = 2$

Contoh

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

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

Terpilih kelompok 1 dan 3, sehingga kedua kelompok ini digabungkan.

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

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

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

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

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

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

Contoh

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)

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

Contoh

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).

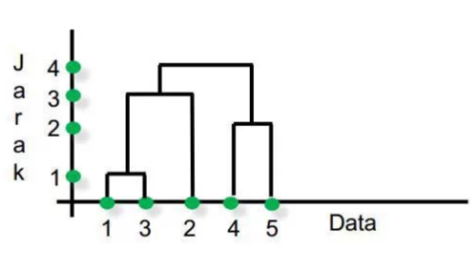
Menghitung jarak antar kelompok ((13) dan 2) dengan kelompok lain yang tersisa, yaitu (45).

$d_{(13 2)(45)} = \min\{d_{14}, d_{15}, d_{34}, d_{35}, d_{24}, d_{25}\} = \min\{5,7,4,6,4,4\} = 4$

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

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

Contoh



Dalam menentukan berapa jumlah cluster sesuai kehendak kita. Misalnya 2 cluster maka cluster 1 berisikan 1, 3, 2. Sedangkan cluster 2 berisikan 4, 5.

Untuk menentukan jumlah cluster optimal, dengan melihat tinggi dari data ke jarak. Maka paling baik 3 cluster.

Pros

1. No assumption of a particular number of clusters (i.e. k-means)
2. May correspond to meaningful taxonomies

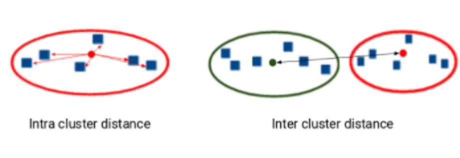
Cons

1. Once a decision is made to combine two clusters, it can't be undone
2. Too slow for large data sets

Evaluation Matrix

1. Inertia : inertia actually calculates the sum of distances of all the points within a cluster from the centroid of that cluster.

$$\text{Dunn Index} = \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$$



2. Dunn Index

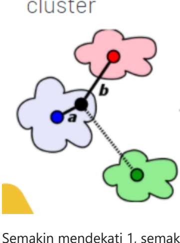
- Intra cluster distance: jarak antara masing-masing data dengan centroidnya.
- Inter cluster distance: jarak antar cluster yang dihitunga dari centroidnya.

> Pada clustering, untuk menentukan bagus atau tidak. Jarak data di dalam intra cluster dekat, dan jarak antar clusternya jauh.
>
> Jika nilai Inertianya kecil maka jarak antar data di dalam cluster dekat. Begitu sebaliknya. Semakin kecil inertia semakin bagus.

Evaluation Matrix

- Silhouette score

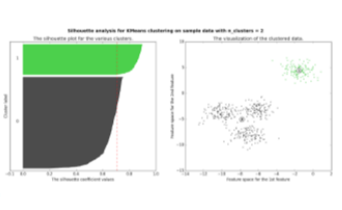
Score = [-1,1] . +1 means a cluster is far away from its neighbor cluster



$$SSI_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

a= average intra-cluster distance

b= average inter-cluster distance



Semakin mendekati 1, semakin bagus. Semakin mendekati -1 maka semakin jelek (tumpang tindih antar cluster.