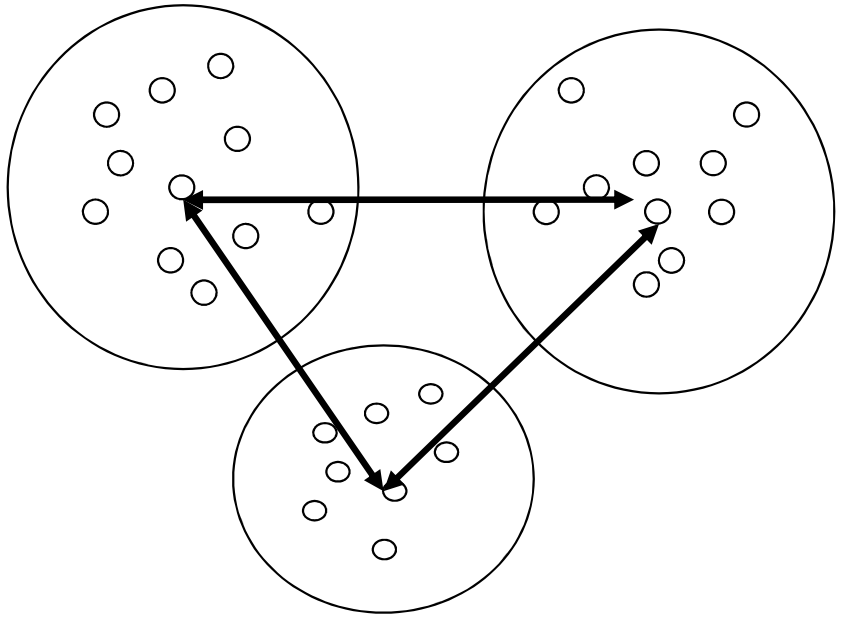
DATA MINING

METODE CLUSTERING

# ALGORITMA K-MEANS



# Pendahuluan

- *Clustering* merupakan suatu teknik data mining yang membagi-bagikan data ke dalam beberapa kelompok (grup atau *cluster* atau segmen) yang tiap *cluster* dapat ditempati beberapa anggota bersama-sama.
- *Clustering* adalah suatu metode pengelompokan berdasarkan ukuran kedekatan (kemiripan).
- *Clustering* tidak mensyaratkan pengetahuan sebelumnya dari kelompok yang dibentuk, juga dari para anggota yang harus mengikutinya.

# Pendahuluan

- Algoritma K-Means diperkenalkan oleh J.B. MacQueen pada tahun 1976, salah satu algoritma *clustering* sangat umum yang mengelompokkan data sesuai dengan karakteristik atau ciri-ciri bersama yang serupa.
- Kelompok data ini dinamakan sebagai klaster *(cluster)*.
- Data di dalam suatu klaster mempunyai ciri-ciri (atau fitur, karakteristik, atribut, properti) serupa dan tidak serupa dengan data pada klaster lain.
- Clustering K-Means hanya dapat digunakan pada data *numerical*.

# Manfaat

- Identifikasi obyek (*Recognition*):
  - Dalam bidang *image processing*, *computer vision* atau *robot vision*.
- *Decision Support System* dan data mining:
  - Segmentasi pasar, pemetaan wilayah, manajemen marketing dll.

# Contoh Penerapan

- Biology : taxonomy makhluk hidup : kingdom, phylum, class, order, family, genus dan species
- Information retrieval : clustering dokumen
- Pemanfaatan lahan : identifikasi area pemanfaatan lahan yang serupa berdasarkan data dalam database *earth observation*
- Marketing : membantu pelaku marketing untuk menemukan kelompok pelanggan tertentu, dan memanfaatkan pengetahuan ini untuk mengembangkan program terhadap kelompok pelanggan yang menjadi target.
- Perencanaan kota : identikasi kelompok rumah berdasarkan tipe, harga dan lokasi geografis.
- Iklim : mempelajari iklim dengan mencari pola pergerakan atmosfer dan samudra.
- Ilmu ekonomi : mencari pengetahuan tentang pasar dalam bidang marketing
- Dll.

# Metode Clustering: K-Means

1. Tentukan jumlah klaster k yang diinginkan.
2. Inisialisasi k pusat klaster (*centroid*) secara random.
3. Tempatkan setiap data atau objek ke klaster terdekat. Kedekatan dua objek ditentukan berdasar jarak. Jarak yang dipakai pada algoritma k-Means adalah *Euclidean distance* (*d*).

$$d_{Euclidean}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Dimana x = x1, x2, . . . , xn, dan y = y1, y2, . . . , yn merupakan banyaknya *n* atribut (kolom) antara 2 record

# Metode Clustering: K-Means

4. Hitung kembali pusat klaster dengan keanggotaan klaster yang sekarang. Pusat klaster adalah rata-rata (*mean*) dari semua data atau objek dalam klaster tertentu
5. Tugaskan lagi setiap objek dengan memakai pusat klaster yang baru. Jika pusat klaster sudah tidak berubah lagi, maka proses pengklasteran selesai. Atau, kembali lagi ke langkah nomor 3 sampai pusat klaster tidak berubah lagi (stabil) atau tidak ada penurunan yang signifikan dari nilai SSE (*Sum of Squared Errors*).

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} \left( d(p, c_i) \right)^2$$

# Contoh Algoritma K-Means (Diketahui satu atribut)

- Lakukan clustering ke dalam 2 kelompok klaster berdasarkan data berikut:

| Data | Nilai X |
|---|---|
| A | 2 |
| B | 3 |
| C | 4 |
| D | 10 |
| E | 11 |
| F | 12 |
| G | 20 |
| H | 25 |
| I | 35 |

# Tahap Clustering (Iterasi 1)

- Diketahui k = 2
- Pilih centroid awal (secara random dari dataset) untuk masing-masing klaster.
  - Misal: centroid C1 (m1) = 20, centroid C2 (m2) = 25
- Hitung jarak masing-masing data X terhadap nilai centroid masing-masing klaster menggunakan *Euclidean Distance*
- Hitung nilai SSE nya.

# Iterasi 1

| Centroid Awal | X |
|---|---|
| m1 | 20 |
| m2 | 25 |

| Data | Jarak ke m1 | Jarak ke m2 | Cluster |
|---|---|---|---|
| A | 18 | 23 | C1 |
| B | 17 | 22 | C1 |
| C | 16 | 21 | C1 |
| D | 10 | 15 | C1 |
| E | 9 | 14 | C1 |
| F | 8 | 13 | C1 |
| G | 0 | 5 | C1 |
| H | 5 | 0 | C2 |
| I | 15 | 10 | C2 |
| SSE | 1214 | | |

$SSE = 18^2+17^2+16^2+10^2+9^2+8^2+0^2+0^2+10^2 = 1214$

- Anggota klaster C1 = {A,B,C,D,E,F,G}
- Anggota klaster C2 = {H,I}

# Tahap Clustering (Iterasi 2)

- Hitung nilai centroid baru berdasarkan nilai rerata anggota masing-masing klaster.
  - centroid C1 (m1) = average(A,B,C,D,E,F,G),
  - centroid C2 (m2) = average(H,I)
- Hitung jarak masing-masing data X terhadap nilai centroid baru masing-masing klaster menggunakan *Euclidean Distance*
- Hitung nilai SSE nya.

# Iterasi 2

| Centroid Baru | X |
|---|---|
| m1 | 8.86 |
| m2 | 30 |

| Data | Jarak ke m1 | Jarak ke m2 | Cluster |
|---|---|---|---|
| A | 6.86 | 28 | C1 |
| B | 5.86 | 27 | C1 |
| C | 4.86 | 26 | C1 |
| D | 1.14 | 20 | C1 |
| E | 2.14 | 19 | C1 |
| F | 3.14 | 18 | C1 |
| G | 11.14 | 10 | C2 |
| H | 16.14 | 5 | C2 |
| I | 26.14 | 5 | C2 |
| SSE | 270.69 | | |

$SSE = 6.86^2+5.86^2+4.86^2+1.14^2+2.14^2+3.14^2+10^2+5^2+5^2 = 270.69$

- Anggota klaster C1 = {A,B,C,D,E,F}
- Anggota klaster C2 = {G,H,I}

# Tahap Clustering (Iterasi 3)

- Karena anggota berubah, maka lakukan iterasi ke-3.
- Hitung nilai centroid baru berdasarkan nilai rerata anggota masing-masing klaster.
  - centroid C1 (m1) = average(A,B,C,D,E,F),
  - centroid C2 (m2) = average(G,H,I)
- Hitung jarak masing-masing data X terhadap nilai centroid baru masing-masing klaster menggunakan *Euclidean Distance*
- Hitung nilai SSE nya.

# Iterasi 3

| Centroid Baru | X |
|---|---|
| m1 | 7.00 |
| m2 | 26.67 |

| Data | Jarak ke m1 | Jarak ke m2 | Cluster |
|---|---|---|---|
| A | 5.00 | 24.67 | C1 |
| B | 4.00 | 23.67 | C1 |
| C | 3.00 | 22.67 | C1 |
| D | 3.00 | 16.67 | C1 |
| E | 4.00 | 15.67 | C1 |
| F | 5.00 | 14.67 | C1 |
| G | 13.00 | 6.67 | C2 |
| H | 18.00 | 1.67 | C2 |
| I | 28.00 | 8.33 | C2 |
| SSE | 216.67 | | |

- Anggota klaster C1 = {A,B,C,D,E,F}
- Anggota klaster C2 = {G,H,I}
- Karena tidak ada perubahan data anggota klaster, maka iterasi dihentikan dengan nilai SSE sebesar 216.67

# Contoh Algoritma K-Means (Diketahui dua atribut)

- Lakukan clustering ke dalam 2 kelompok klaster berdasarkan data berikut:

| DATA | X1 | X2 |
|---|---|---|
| A | 1 | 3 |
| B | 3 | 3 |
| C | 4 | 3 |
| D | 5 | 3 |
| E | 1 | 2 |
| F | 4 | 2 |
| G | 1 | 1 |
| H | 2 | 1 |

# Tahap Clustering (Iterasi 1)

- Diketahui k = 2
- Pilih centroid awal (secara random dari dataset) untuk masing-masing klaster.
  - Misal: centroid C1 (m1) = (1,1), centroid C2 (m2) = (2,1)
- Hitung jarak masing-masing data X terhadap nilai centroid masing-masing klaster menggunakan *Euclidean Distance*
- Hitung nilai SSE nya.

# Iterasi 1

| Centroid Awal | X1 | X2 |
|---|---|---|
| m1 | 1 | 1 |
| m2 | 2 | 1 |

| Data | Jarak ke m1 | Jarak ke m2 | Cluster |
|---|---|---|---|
| A | 2.00 | 2.24 | C1 |
| B | 2.83 | 2.24 | C2 |
| C | 3.61 | 2.83 | C2 |
| D | 4.47 | 3.61 | C2 |
| E | 1.00 | 1.41 | C1 |
| F | 3.16 | 2.24 | C2 |
| G | 0.00 | 1.00 | C1 |
| H | 1.00 | 0.00 | C2 |
| SSE | 36.00 | | |

$SSE = 2^2+2.24^2+2.83^2+3.61^2+1^2+2.24^2+0^2+0^2 = 36$

- Anggota klaster C1 = {A,E,G}
- Anggota klaster C2 = {B,C,D,F,H}

# Tahap Clustering (Iterasi 2)

- Hitung nilai centroid baru berdasarkan nilai rerata anggota masing-masing klaster.
  - centroid C1 (m1) = average(A,E,G),
  - centroid C2 (m2) = average(B,C,D,F,H)
- Hitung jarak masing-masing data X terhadap nilai centroid baru masing-masing klaster menggunakan *Euclidean Distance*
- Hitung nilai SSE nya.

# Iterasi 2

| Centroid Baru | X1 | X2 |
|---|---|---|
| m1 | 1 | 2 |
| m2 | 3.6 | 2.4 |

| Data | Jarak ke m1 | Jarak ke m2 | Cluster |
|---|---|---|---|
| A | 1.00 | 2.67 | C1 |
| B | 2.24 | 0.85 | C2 |
| C | 3.16 | 0.72 | C2 |
| D | 4.12 | 1.52 | C2 |
| E | 0.00 | 2.63 | C1 |
| F | 3.00 | 0.57 | C2 |
| G | 1.00 | 2.95 | C1 |
| H | 1.41 | 2.13 | C1 |
| SSE | 7.88 | | |

$SSE = 1^2+0.85^2+0.72^2+1.52^2+0^2+0.57^2+1^2+1.41^2 = 7.88$

- Anggota klaster C1 = {A,E,G,H}
- Anggota klaster C2 = {B,C,D,F}

# Tahap Clustering (Iterasi 3)

- Karena anggota berubah, maka lakukan iterasi ke-3.
- Hitung nilai centroid baru berdasarkan nilai rerata anggota masing-masing klaster.
  - centroid C1 (m1) = average(A,E,G,H),
  - centroid C2 (m2) = average(B,C,D,F)
- Hitung jarak masing-masing data X terhadap nilai centroid baru masing-masing klaster menggunakan *Euclidean Distance*
- Hitung nilai SSE nya.

# Iterasi 3

| Centroid Baru | X1 | X2 |
|---|---|---|
| m1 | 1.25 | 1.75 |
| m2 | 4 | 2.75 |

| Data | Jarak ke m1 | Jarak ke m2 | Cluster |
|---|---|---|---|
| A | 1.27 | 3.01 | C1 |
| B | 2.15 | 1.03 | C2 |
| C | 3.02 | 0.25 | C2 |
| D | 3.95 | 1.03 | C2 |
| E | 0.35 | 3.09 | C1 |
| F | 2.76 | 0.75 | C2 |
| G | 0.79 | 3.47 | C1 |
| H | 1.06 | 2.66 | C1 |
| SSE | 6.25 | | |

$SSE = 1.27^2+1.03^2+0.25^2+1.03^2+0.35^2+0.75^2+0.79^2+1.06^2 = 6.25$

- Anggota klaster C1 = {A,E,G,H}
- Anggota klaster C2 = {B,C,D,F}
- Karena tidak ada perubahan data anggota klaster, maka iterasi dihentikan dengan nilai SSE sebesar 6.25

# Cukup mudahkan?

# Latihan

Lakukan clustering ke dalam 2 kelompok klaster:

| NAMA | B.IND | B.ING |
|---|---|---|
| JOKO | 8.54 | 8.4 |
| AGUS | 9.98 | 6.81 |
| SUSI | 6.2 | 9.15 |
| DYAH | 5.24 | 7.26 |
| WATI | 5.7 | 5.71 |
| IKA | 8.57 | 5.87 |
| EKO | 7.7 | 7.71 |
| YANTO | 6.6 | 5.7 |
| WAWAN | 9 | 8.12 |
| MAHMUD | 9.81 | 9.58 |