Data Mining

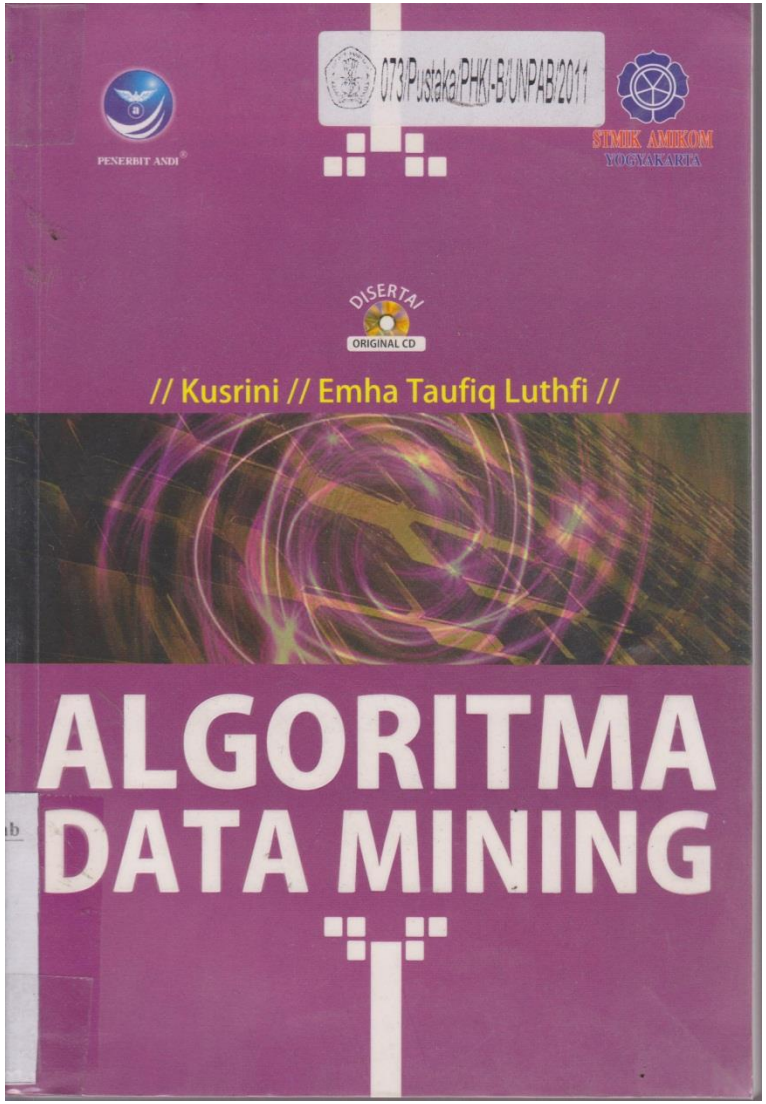
Junta Zeniarja, M.Kom, M.CS
junta@dsn.dinus.ac.id

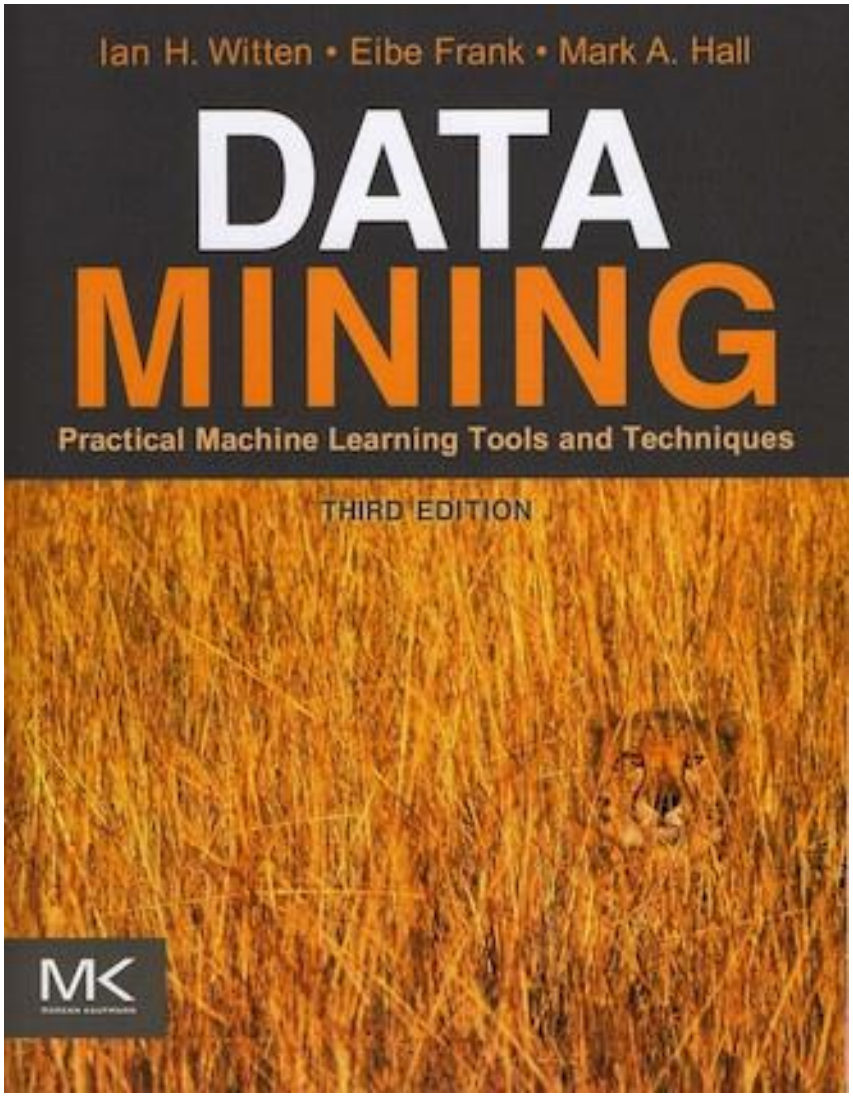
Profil

- Nama : Junta Zeniarja, M.Kom
- Alamat : Semarang
- Kontak
  - Phone : -
  - E-mail : junta@dsn.dinus.ac.id
  - Room : Ruang Dosen TI-S1 (H.2.3)

- Pendidikan
  - S1 => TI – UDINUS
  - S2 => TI – UDINUS
  - S2 => Computer Science UTeM (Universiti Teknikal Malaysia Melaka)
- Konsultasi - Sharing
  - 1:00 pm – 4:00 pm, Senin-Kamis.
  - Appointment via phone or e-mail preferred

Textbooks


PENERBIT ANDI
STMIK AMIKOM YOGYAKARTA
DISERTAI ORIGINAL CD
// Kusrini // Emha Taufiq Luthfi //
ALGORITMA DATA MINING


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

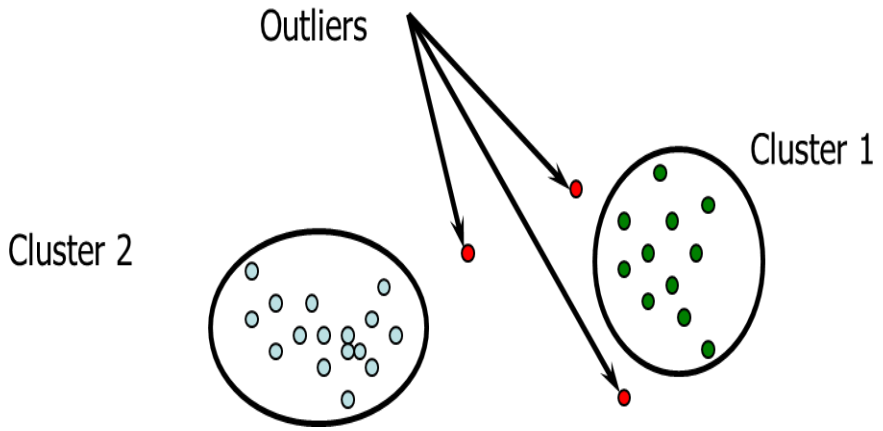
Clustering

DEFINISI CLUSTERING (DARI PAKAR DATA MINING)

- **Clustering** is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters (Bramer, 2007)
- Cluster analysis is the best-known descriptive data mining method. Given a data matrix composed of *n* observations (rows) and *p* variables(columns), the objective of cluster analysis is to cluster the observations into groups that are internally homogeneous (internal cohesion) and heterogeneous from group to group (external separation). (Guidici, 2009)
- By *clustering* we mean the method to divide a set of data (records/tuples/vectors/instances/objects/sample) into several groups (clusters), based on certain predetermined similarities. (Goronescu,2011)

DEFINISI CLUSTERING (DARI PAKAR DATA MINING)

- Clustering refers to the grouping of records, observations, or cases into classes of similar objects. A cluster is a collection of records that are similar to one another, and dissimilar to records in other clusters. Clustering algorithms seek to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized and the similarity to records outside the cluster is minimized. (Larose, 2005)



Algoritma Clustering

- Klastering adalah pengelompokkan data, hasil observasi dan kasus ke dalam class yang mirip
- Suatu klaster (cluster) adalah koleksi data yang mirip antara satu dengan yang lain, dan memiliki perbedaan bila dibandingkan dengan data dari klaster lain
- Perbedaan utama algoritma klastering dengan klasifikasi adalah klastering tidak memiliki target/class/label, jadi termasuk *unsupervised learning*
- Klastering sering digunakan sebagai tahap awal dalam proses data mining, dengan hasil klaster yang terbentuk akan menjadi input dari algoritma berikutnya yang digunakan

TIPE-TIPE CLUSTERING

William ([8]) membagi algoritma *clustering ke dalam kel ompok besar*

seperti berikut:

1. *Partitioning algorithms:* algoritma dalam kelompok ini membentuk bermacam partisi dan kemudian mengeval uasinya dengan berdasarkan beberapa kriteria.
2. *Hierarchy algorithms:* pembentukan dekomposisi hirarki dari sekumpulan data menggunakan beberapa kriteria.
3. *Density based:* pembentukan cluster berdasarkan pada koneksi dan fungsi densitas.
4. *Grid-based:* pembentukan cluster berdasarkan pada struktu r multiple level granularity
5. *Model-based:* sebuah model dianggap sebagai hipotesa untu k masing-masing cluster dan model yang baik dipilih diantara model hipotesa tersebut.

Contoh: Klastering Jenis Gaya Hidup

- Claritas, Inc. provide a demographic profile of each of the geographic areas in the country, as defined by zip code. One of the clustering mechanisms they use is the PRIZM segmentation system, which describes every U.S. zip code area in terms of distinct lifestyle types (66 segments). Just go to the company's Web site, enter a particular zip code, and you are shown the most common PRIZM clusters for that zip code.
- What do these clusters mean? For illustration, let's look up the clusters for zip code 90210, Beverly Hills, California. The resulting clusters for zip code 90210 are:
  1. *Cluster 01: Blue Blood Estates*
  2. *Cluster 10: Bohemian Mix*
  3. *Cluster 02: Winner's Circle*
  4. *Cluster 07: Money and Brains*
  5. *Cluster 08: Young Literati*

Nielsen PRIZM - Understanding Social and Lifestage G

- What is Nielsen PRIZM?
- Features and Benefits
- Lifestyle Segmentation
- Urbanization Classes
- Social Groups
- Lifestage Classes
- Lifestage Groups
- Summary

Features and Benefits

and market to them with tailored messages and products designed just for them. Captured by catchy names, images and behavior snapshots that bring the segments to life for marketers, PRIZM segments are memorable and summarize complex consumer profiles in a way that is intuitive and easy to communicate.

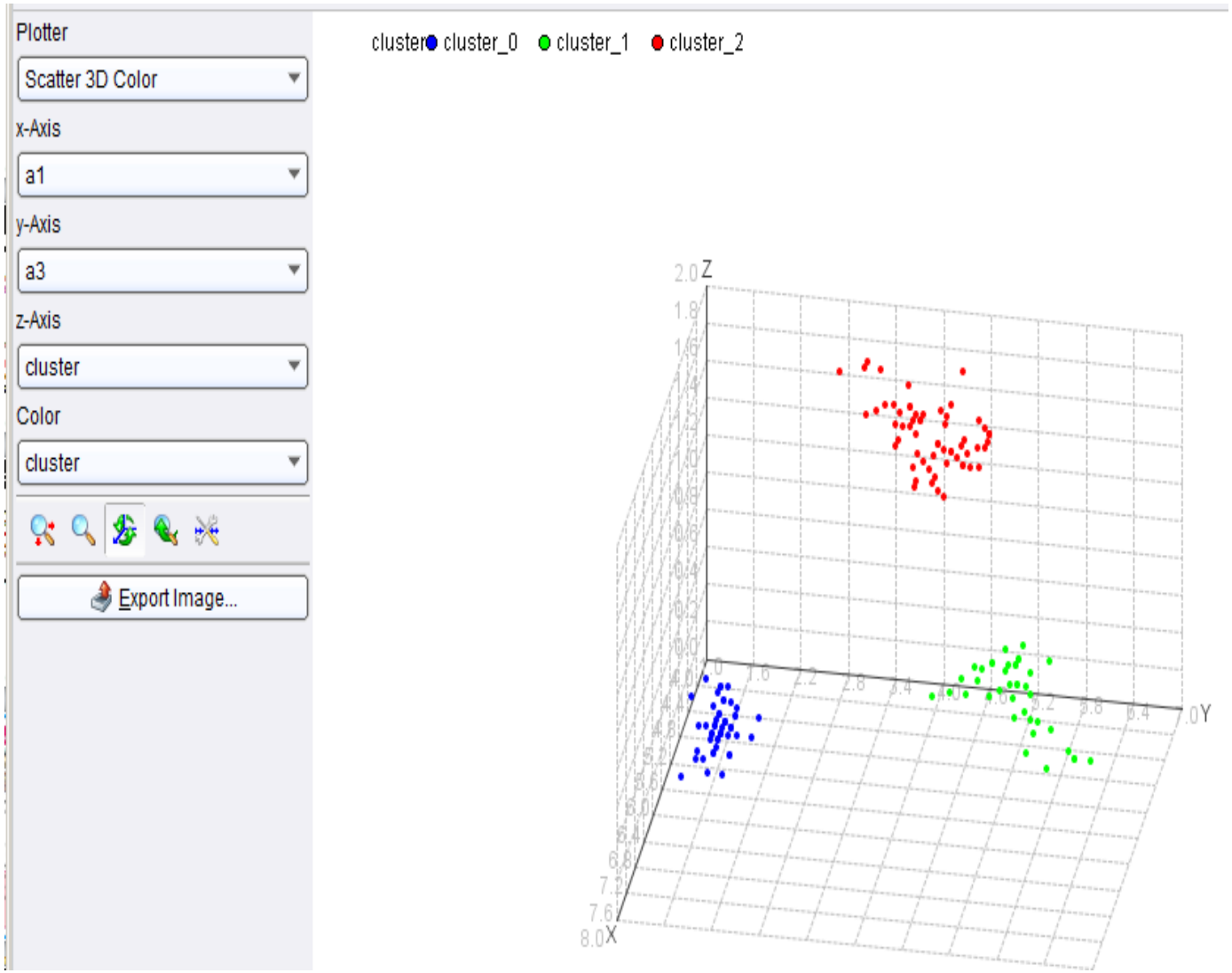
For example, PRIZM Segment number 16 is known as *Bohemian Mix*. We can describe both the demographic traits, as well as the lifestyle characteristics of the households in this segment. You can review these segment descriptors in the image at right.

Contoh: Klastering Bunga Iris

ExampleSet (150 examples, 2 special attributes, 4 regular attributes)

Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	4.800	3	1.400	0.100
14	id_14	Iris-setosa	4.300	3	1.100	0.100
15	id_15	Iris-setosa	5.800	4	1.200	0.200
16	id_16	Iris-setosa	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	5.100	3.800	1.500	0.300
21	id_21	Iris-setosa	5.400	3.400	1.700	0.200
22	id_22	Iris-setosa	5.100	3.700	1.500	0.400
23	id_23	Iris-setosa	4.600	3.600	1	0.200
24	id_24	Iris-setosa	5.100	3.300	1.700	0.500

Contoh: Klastering Bunga Iris (Plot)



CONTOH: KLASTERING BUNGA IRIS (TABLE)

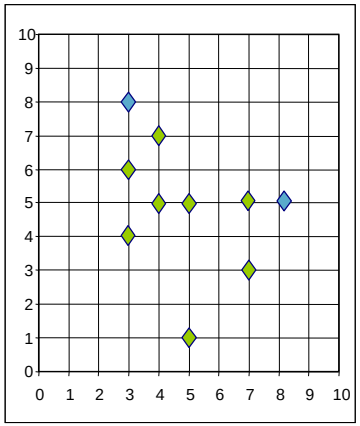
ExampleSet (150 examples, 3 special attributes, 4 regular attributes) View

Row No.	id	label	cluster	a1	a2	a3	a4
1	id_1	Iris-setosa	cluster_0	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	cluster_0	4.900	3	1.400	0.200
3	id_3	Iris-setosa	cluster_0	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	cluster_0	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	cluster_0	5	3.600	1.400	0.200
6	id_6	Iris-setosa	cluster_0	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	cluster_0	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	cluster_0	5	3.400	1.500	0.200
9	id_9	Iris-setosa	cluster_0	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	cluster_0	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	cluster_0	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	cluster_0	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	cluster_0	4.800	3	1.400	0.100
14	id_14	Iris-setosa	cluster_0	4.300	3	1.100	0.100
15	id_15	Iris-setosa	cluster_0	5.800	4	1.200	0.200
16	id_16	Iris-setosa	cluster_0	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	cluster_0	5.400	3.900	1.300	0.400
18	id_18	Iris-setosa	cluster_0	5.100	3.500	1.400	0.300
19	id_19	Iris-setosa	cluster_0	5.700	3.800	1.700	0.300
20	id_20	Iris-setosa	cluster_0	5.100	3.800	1.500	0.300

Cluster Model

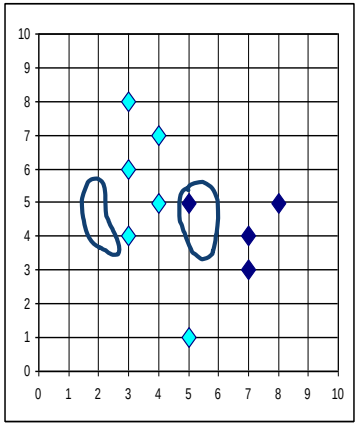
```
Cluster 0: 50 items
Cluster 1: 39 items
Cluster 2: 61 items
Total number of items: 150
```

CONTOH: KLASTERING K-MEANS

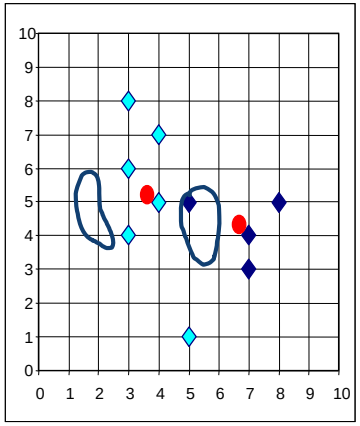


K=2

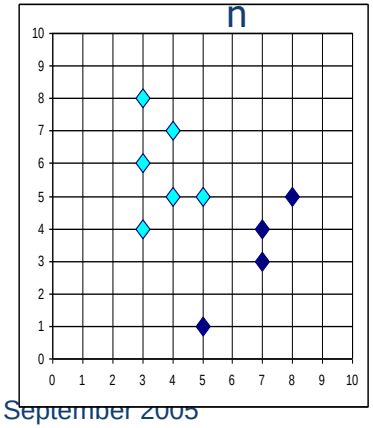
Arbitrarily choose K object as initial cluster center



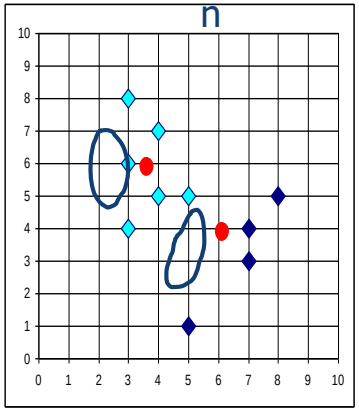
Assign each objects to most similar center



Update the cluster means



reassign



Update the cluster means

reassign

K-MEANS ALGORITHm

- *K-means pertama kali dipublikasikan oleh Stuart Lloyd pada tahun 1984* dan merupakan algoritma clustering yang banyak digunakan.
- *K-means bekerja* dengan mensegmentasi objek yang ada kedalam kelompok atau yang disebut dengan segmen sehingga objek yang berada dalam masing-masing kelompok lebih serupa satu sama lain dibandingkan dengan objek dalam kelompok yang berbeda.
- Algoritma Clustering adalah meletakkan nilai yang serupa dalam satu segmen, dan meletakkan nilai yang berbeda dalam cluster yang berbeda (Wu & Kumar, 2009).
- *K-Means memisahkan data dengan optimal dengan perulangan* yang memaksimalkan hasil dari partisi hingga tidak ada perubahan data dalam setiap segmentasi.

K-MEANS ALGORITHm … lanjt

- *K-Means bekerja dengan pendekatan Top-Down karena memulai dengan* segmentasi yang sudah ditentukan terlebih dahulu (Myatt, 2007). Sehingga hasil data sebuah segmen tidak mungkin tercampur antara satu segmen dengan segmen lainnya (Xu & Wunsch II, 2009). Pendekatan ini juga mempercepat proses komputasi untuk data dalam jumlah besar.
- Algoritma *K-means diterapkan pada objek yang diwakili dalam bentuk* titik didalam ruangan vektor berdimensi-d. *K-means mengcluster semua data* didalam setiap dimensi dimana titik dalam segmentasi yang sama diberi custer ID. Nilai dari k adalah masukan dasar dari algoritma yang menentukan jumlah segmentasi yang ingin dibentuk. Partisi akan dibentuk dari sekumpulan objek n kedalam cluster k sehingga terbentuk kesamaan objek dalam setiap segmentasi k.

K-MEANS ALGORITHm … lanjt

- Untuk menghasilkan cluster yang maksimal, titik awal partisi merupakan salah satu faktor yang berpengaruh.
- Karena *k-means memecahkan data kedalam* segmentasi berdasarkan nilai lokal maksimum. Karena itu pemilihan titik awal harus beralasan, ada beberapa metode yang diusulkan oleh beberapa peneliti seperti genetic *k-means, pemilihan titik awal secara acak sebanyak beberapa kali, metode yang* paling baik adalah dengan mengukur nilai titik tengah segmentasi berdasarkan jumlah jarak terpendek antar anggota kelas tersebut.

K-MEANS ALGORITHm … lanjt

Beberapa kelemahan dari algoritma *k-means antara lain:*

a. Perlu mengetahui jumlah segmen yang ingin dibentuk, penentuan jumlah segmentasi ini membutuhkan pengalaman dalam melihat dan menilai jumlah segmentasi yang ada dalam data.

b. Mudah terganggu dengan data yang tidak valid, karena cara kerja k-means merata-rata nilai dalam setiap segmen, maka data yang tidak relevan dapat mengacaukan pusat segmen.

c. Hasil partisi dari *k-means tidak memiliki nilai hirarki, sehingga tidak ada* segmentasi yang lebih baik dari segmentasi lainnya.

LANGKAH2 K-MEANS ALGORITHM

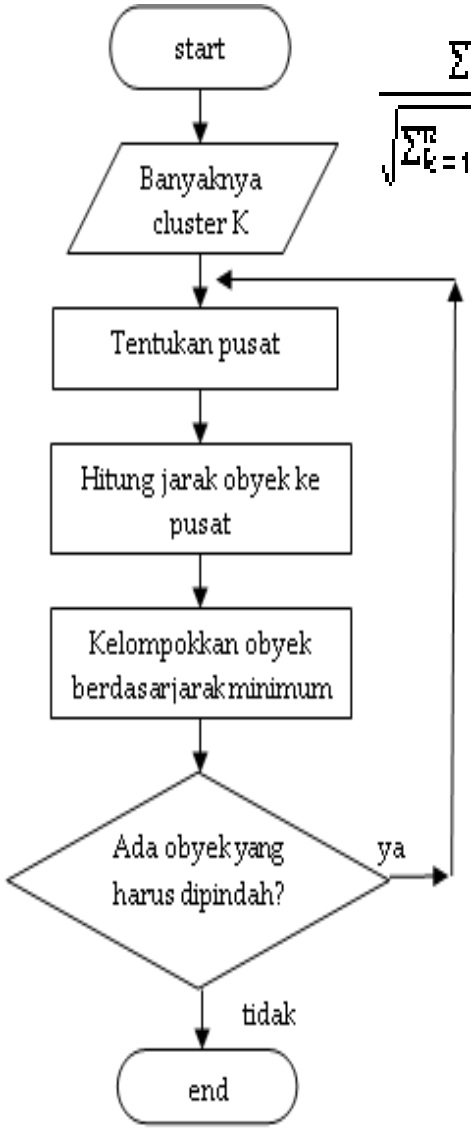
Langkah-langkah algoritma K-means adalah sebagai berikut:

1. Tentukan nilai k sebagai jumlah klaster yang ingin dibentuk
2. Bangkitkan k centroid (titik pusat klaster) awal secara random.
3. Hitung jarak setiap data ke masing-masing centroid menggunakan rumus korelasi antar dua objek yaitu Euclidean Distance dan kesamaan Cosine.
4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan centroidnya.
5. Tentukan posisi centroid baru (k C) dengan cara menghitung nilai rata-rata dari data-data yang ada pada centroid yang sama.

$$C_k = \left(\frac{1}{n_k}\right) \sum d_i$$

Dimana k n adalah jumlah dokumen dalam cluster k dan i d adalah dokumen dalam cluster k.

6. Kembali ke langkah 3 jika posisi centroid baru dengan centroid lama tidak sama.



Gambar 1. Flowchart algoritma K-Means

$$\frac{\sum_{k=1}^{n} x_k \times y_k}{\sqrt{\sum_{k=1}^{n} x_k^2} \times \sqrt{\sum_{k=1}^{n} y_k^2}}$$

EUCLIDEAN DISTANCE

Cara Menghitung Euclidean Distance

Euclidean distance adalah perhitungan jarak dari 2 buah titik dalam Euclidean space. Euclidean space diperkenalkan oleh seorang matematikawan dari Yunani sekitar tahun 300 B.C.E. untuk mempelajari hubungan antara sudut dan jarak. Euclidean ini biasanya diterapkan pada 2 dimensi dan 3 dimensi. Tapi juga sederhana jika diterapkan pada dimensi yang lebih tinggi.

1 dimensi

- Semisal ingin menghitung jarak Euclidean 1 dimensi. Titik pertama adalah 4, titik kedua adalah -10. Caranya adalah kurangkan -10 dengan 4. sehingga menghasilkan -14. Cari nilai absolut dari nilai -14 dengan cara mempangkatkannya sehingga mendapat nilai 196. Kemudian diakarkan sehingga mendapatkan nilai 14. Sehingga jarak euclidean dari 2 titik tersebut adalah 14.

2 dimensi

- Caranya hampir sama. Misalkan titik pertama mempunyai kordinat (3,5). Titik kedua ada di kordinat (5,-3). Caranya adalah kurangkan setiap kordinat titik kedua dengan titik yang pertama. Yaitu, (5-3,-3-5) sehingga menjadi (2,-8). Kemudian pangkatnya sehingga memperoleh (4,64). Kemudian tambahkan semuanya sehingga memperoleh nilai 64+4 = 68. Hasil ini kemudian diakarkan menjadi 8.25. Sehingga jarak euclideannya menjadi 8.25.

EUCLIDEAN DISTANCE … lanjt

Euclidean Distance adalah metrika yang paling sering digunakan untuk menghitung kesamaan 2 vektor. Euclidean distance menghitung akar dari kuadrat perbedaan 2 vektor (root of square differences between 2 vectors).

Rumus dari Euclidian Distance:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Contoh :

Terdapat 2 vektor ciri

$A = [0, 3, 4, 5]$

$B = [7, 6, 3, -1]$

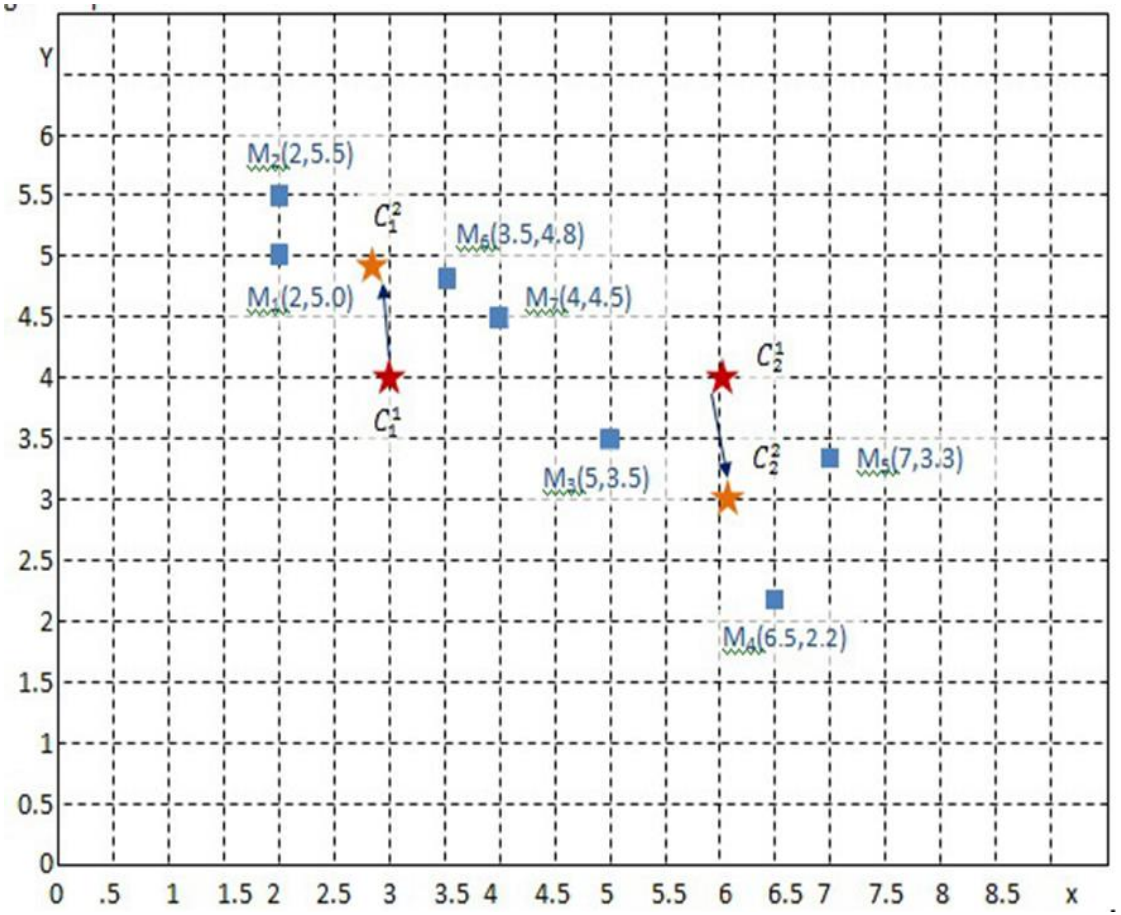
Euclidean Distance dan B adalah

$$d_{AB} = \sqrt{(0-7)^2 + (3-6)^2 + (4-3)^2 + (5-(-1))^2}$$

$$= \sqrt{49 + 9 + 1 + 36} = 9.747$$

Contoh kAsus

- Using K-means algorithm find the best groupings and means of two clusters of the 2D data below. Show all your work, assumptions, and regulations.
- 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)



Asumsi:

- Semua data akan dikelompokkan ke dalam dua kelas
- Center points of both clusters are **$C_1(3,4)$, $C_2(6,4)$**

Contoh kAsus (iterasi 1) ... LANJ

Iterasi 1

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

Sehingga didapatkan

D11=1.41,

D12=1.80,

D13=2.06,

D14=3.94,

D15=4.06,

D16=0.94,

D17=1.12,

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

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

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

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

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

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

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

Dengan cara yang sama hitung jarak tiap titik ke titik pusat kedua, dan kita akan mendapatkan :

D21 = 4.12, D22 = 4.27, D23 = 1.18, D24= 1.86,

D25 =1.22, D26 = 2.62, D27 = 2.06

Contoh kAsus (iterasi 1) … LANJ

b. Dari penghitungan *Euclidean distance*, kita dapat membandingkan:

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

{M1, M2, M6, M7} anggota C1 and {**M3, M4, M5**} anggota C2

c. Hitung titik pusat baru

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)

$$C1 = \left(\frac{2+2+3.5+4}{4}, \frac{5+5.5+4.8+4.5}{4}\right) = (2.85, 4.95)$$

$$C2 = \left(\frac{5+6.5+7}{3}, \frac{3.5+2.2+3.3}{3}\right) = (6.17, 3)$$

Contoh kAsus (iterasi 2) … LANJ

ITERASI 2

a) Hitung Euclidean distance dari tiap data ke titik pusat yang baru Dengan cara yang sama dengan iterasi pertama kita akan mendapatkan perbandingan sebagai berikut:

M_1		M1	M_2	M_3	M_4	M_5	M_6	M_7
Jarak ke	C_1	**0.76**	**0.96**	2.65	4.62	4.54	**0.76**	**1.31**
	C_2	4.62	4.86	**1.27**	**0.86**	**0.88**	3.22	2.63

b) Dari perbandingan tersebut kira tahu bahwa {M_1, M_2, M_6, M_7} anggota C_1 dan {**M_3, M_4, M_5**} anggota C_2

c) Karena anggota kelompok tidak ada yang berubah maka titik pusat pun tidak akan berubah.

KESIMPULAN

{M_1, M_2, M_6, M_7} anggota C_1 dan {**M_3, M_4, M_5**} anggota C_2

Latihan

- Tentukan anggota klasternya!

  M1 = (1, 4.5),

  M2 = (3, 6.5),

  M3 = (4, 4.5),

  M4 = (7.5, 3.2),

  M5 = (6, 2.3),

  M6 = (2.5, 3.8),

  M7 = (5, 5.5)

- Center points => **$C_1(3,4)$, $C_2(6,4)$**


Thank You!