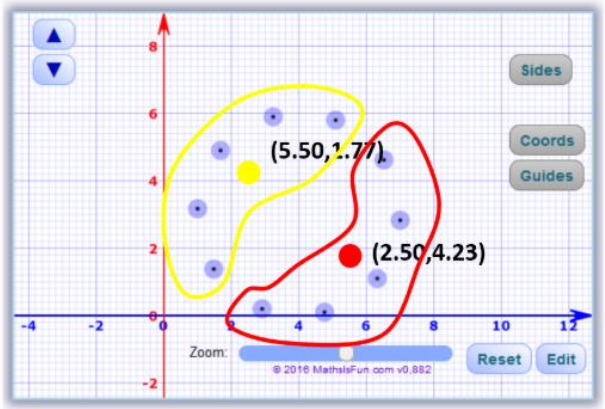
K-MEANS CLUSTERING ALGORITHM

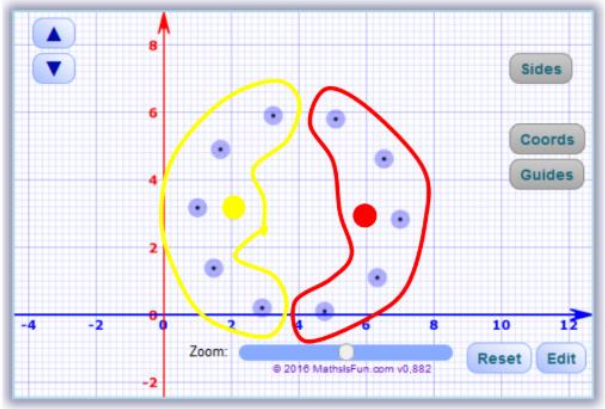
Centroid Awal
(1.48,1.38), (4.76,0.1)



(5.09,5.80), (3.24,5.90), (1.68,4.90), (1.00,3.17), (1.48,1.38), (2.91,0.20), (4.76,0.10), (6.32,1.10), (7.00,2.83), (6.52,4.62), **(2.50,4.23), (5.50,1.77)**

Centroid Akhir
(2.50,4.23), (5.50,1.77)

Centroid Awal
(5.09,5.8), (3.24,5.9)



(5.09,5.80), (3.24,5.90), (1.68,4.90), (1.00,3.17), (1.48,1.38), (2.91,0.20), (4.76,0.10), (6.32,1.10), (7.00,2.83), (6.52,4.62), **(5.94,2.89), (2.06,3.11)**

Centroid Akhir
(5.94,2.89), (2.06,3.11)

Konsep Clustering

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.

Konsep Clustering

- Pengelompokkan sejumlah data atau objek ke dalam klaster (group) sehingga dalam setiap klaster akan berisi data yang semirip mungkin
- Termasuk unsupervised learning
- Data pada teknik pengklasteran tidak diketahui keluarannya (outputnya atau labelnya)
- Metode untuk mengukur kualitas klaster : jumlah dari kesalahan kuadrat (sum of squared-error, SSE) :

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

$p \in C_i$ = tiap data poin pada cluster i, m_i = centroid dari cluster i, d = jarak/ distances/ variance terdekat pada masing-masing cluster i.

- Nilai SSE tergantung pada jumlah klaster dan bagaimana data dikelompokkan ke dalam klaster-klaster. Semakin kecil nilai SSE semakin bagus hasil klastering yang dibuat

Algoritma K-Means Clustering

- Termasuk partitioning clustering
- Objek-objek dikelompokkan ke dalam *k* klaster
- Untuk melakukan klastering ini, nilai *k* harus ditentukan terlebih dahulu
- Kluster-kluster tersebut mempunyai suatu nilai tengah (nilai pusat) yang disebut dengan centroid
- Menggunakan ukuran kemiripan untuk mengelompokkan objek.
- Kemiripan diterjemahkan dalam konsep jarak (distance (*d*))
- Jika jarak dua objek atau data, semakin dekat berarti semakin tinggi kemiripannya
- Tujuan dari *k-Means* : meminimalisir total dari jarak elemen-elemen antar kluster (jarak antara suatu elemen dalam sebuah kluster dengan nilai centroid kluster tersebut)

Algoritma K-Means Clustering

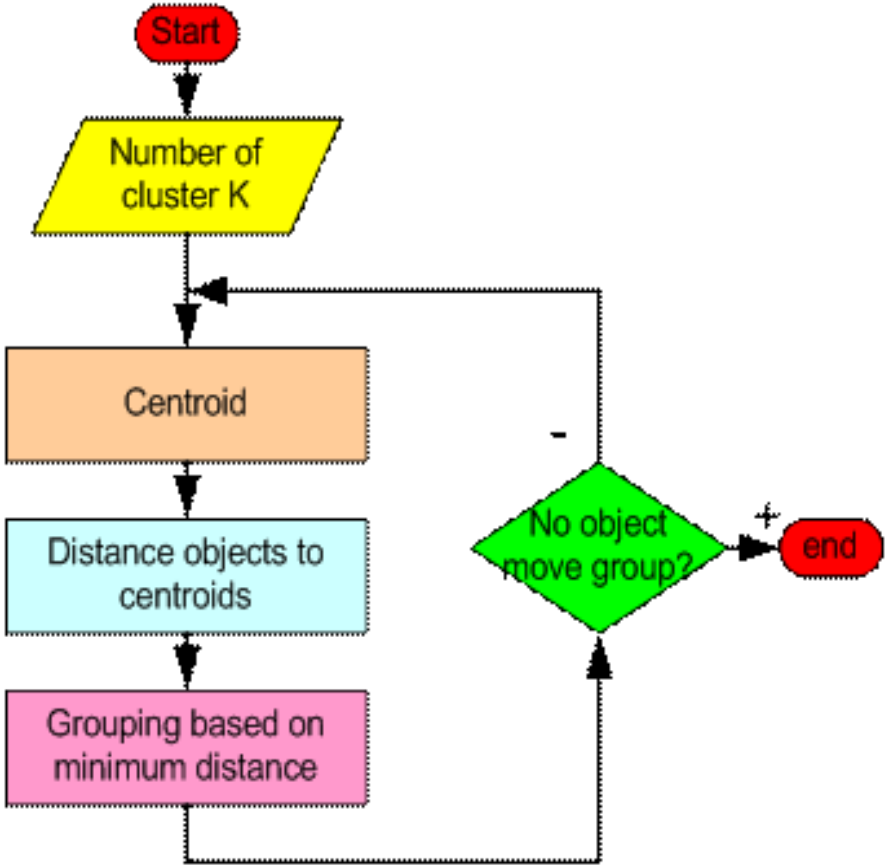
1. Pilih jumlah klaster *k* yang diinginkan
2. Inisialisasi *k* pusat klaster (centroid) secara random/ acak
3. Tempatkan setiap data atau objek ke klaster terdekat. Kedekatan dua objek ditentukan berdasar jarak. Jarak yang dipakai pada algoritma *k-Means* adalah *Euclidean distance (d)*.

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

x = *x*1, *x*2, . . . , *xn*, dan **y** = *y*1, *y*2, . . . , *yn* merupakan banyaknya n atribut(kolom) antara 2 record.

4. Hitung kembali pusat klaster dengan keanggotaan klaster yang sekarang. Pusat klaster adalah rata-rata (mean) dari semua data atau objek dalam klaster tertentu.

Algoritma K-Means Clustering

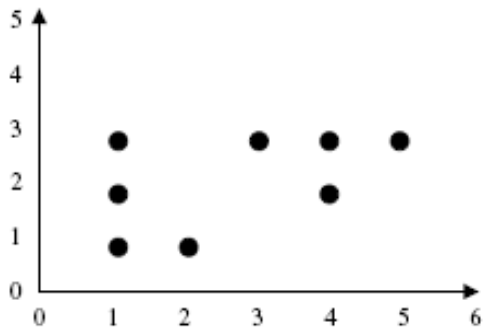


Contoh 1

Tabel 1 Data point

Instances	X	Y
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1

tampilan data awal



1. Tentukan jumlah klaster *k*=2
2. Tentukan centroid awal secara acak misal dari data disamping m1 =(1,1), m2=(2,1)
3. Tempatkan tiap objek ke klaster terdekat berdasarkan nilai centroid yang paling dekat selisihnya(jaraknya). Pada tabel 2.Didapatkan hasil: anggota cluster1 = {A,E,G}, cluster2={B,C,D,F,H}. Nilai SSE yaitu :

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

$$= 2^2 + 2{,}24^2 + 2{,}83^2 + 3{,}61^2 + 1^2 + 2{,}24^2 + 0^2 + 0^2 = 36$$

Contoh 1

Tabel 2

Point	Distance from m_1	Distance from m_2	Cluster Membership
a	2.00	2.24	C_1
b	2.83	2.24	C_2
c	3.61	2.83	C_2
d	4.47	3.61	C_2
e	1.00	1.41	C_1
f	3.16	2.24	C_2
g	0.00	1.00	C_1
h	1.00	0.00	C_2

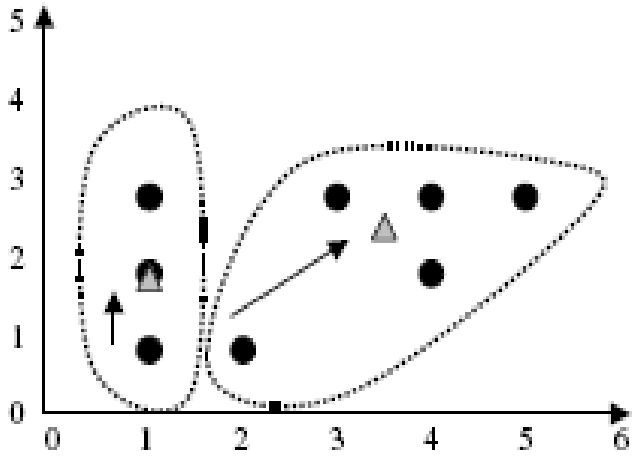
4. Menghitung nilai centroid yang baru :

$$m_1 = \left[(1+1+1)/3, (3+2+1)/3\right] = (1,2)$$

$$m_2 = \left[(3+4+5+4+2)/5, (3+3+3+2+1)/5\right] = (3{,}6;2{,}4)$$

5. Tugaskan lagi setiap objek dengan memakai pusat klaster yang baru. Pada tabel 3. Nilai SSE yang baru :

Clusters dan centroid *setelah tahap pertama*



$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p,m_i)^2 = 1^2 + 0.85^2 + 0.72^2 + 1.52^2 + 0^2 + 0.57^2 + 1^2$$

$$+1.41^2 = 7.88$$

Contoh 1

Tabel 3

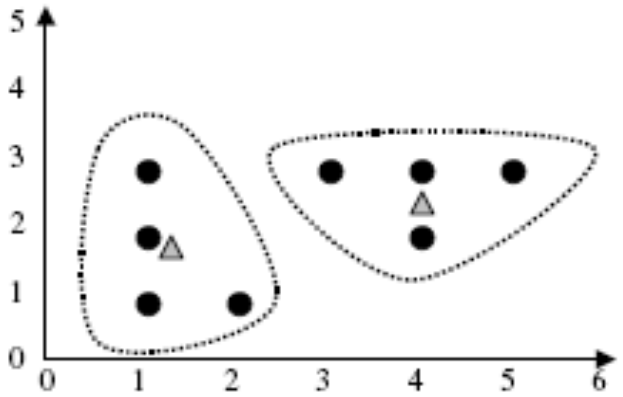
Point	Distance from m_1	Distance from m_2	Cluster Membership
a	1.00	2.67	C_1
b	2.24	0.85	C_2
c	3.16	0.72	C_2
d	4.12	1.52	C_2
e	0.00	2.63	C_1
f	3.00	0.57	C_2
g	1.00	2.95	C_1
h	1.41	2.13	C_2

- Terdapat perubahan anggota cluster yaitu cluster1={A,E,G,H}, cluster2={B,C,D,F}, maka cari lagi nilai centroid yang baru yaitu : m1=(1,25;1,75) dan m2=(4;2,75)
- Tugaskan lagi setiap objek dengan memakai pusat klaster yang baru. Pada tabel 4. Nilai SSE yang baru :

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} d(p,m_i)^2 = 1.27^2 + 1.03^2 + 0.25^2 + 1.03^2 + 0.35^2 + 0.75^2$$

$$+0.79^2 + 1.06^2 = 6.25$$

Clusters dan centroid *setelah tahap kedua*.



Contoh 1

Tabel 4

Point	Distance from m_1	Distance from m_2	Cluster Membership
a	1.27	3.01	C_1
b	2.15	1.03	C_2
c	3.02	0.25	C_2
d	3.95	1.03	C_2
e	0.35	3.09	C_1
f	2.76	0.75	C_2
g	0.79	3.47	C_1
h	1.06	2.66	C_2

- Dapat dilihat pada tabel 4.Tidak ada perubahan anggota lagi pada masing-masing cluster
- Hasil akhir yaitu :

  cluster1={A,E,G,H}, dan cluster2={B,C,D,F} dengan nilai SSE = 6,25 dan jumlah iterasi 3

K-Means Clustering Visual Basic Code

```
Sub kMeanCluster (Data() As Variant, numCluster As Integer)
' main function to cluster data into k number of Clusters
' input:
' + Data matrix (0 to 2, 1 to TotalData);
' Row 0 = cluster, 1 =X, 2= Y; data in columns
' + numCluster: number of cluster user want the data to be clustered
' + private variables: Centroid, TotalData
' ouput:
' o) update centroid
' o) assign cluster number to the Data (= row 0 of Data)

Dim i As Integer
Dim j As Integer
Dim X As Single
Dim Y As Single
Dim min As Single
Dim cluster As Integer
Dim d As Single
Dim sumXY()

Dim isStillMoving As Boolean
isStillMoving = True
if totalData <= numCluster Then
'only the last data is put here because it designed to be interactive
Data(0, totalData) = totalData ' cluster No = total data
Centroid(1, totalData) = Data(1, totalData) ' X
Centroid(2, totalData) = Data(2, totalData) ' Y
Else
'calculate minimum distance to assign the new data
min = 10 ^ 10 'big number
X = Data(1, totalData)
Y = Data(2, totalData)
For i = 1 To numCluster
d = dist(X, Y, Centroid(1, i), Centroid(2, i))
If d < min Then
min = d
cluster = i
End If
Next i
Data(0, totalData) = cluster
```

K-Means Clustering Visual Basic Code

```
Do While isStillMoving
' this loop will surely convergent
'calculate new centroids
' 1 =X, 2=Y, 3=count number of data
ReDim sumXY(1 To 3, 1 To numCluster)
For i = 1 To totalData
sumXY(1, Data(0, i)) = Data(1, i) + sumXY(1, Data(0, i))
sumXY(2, Data(0, i)) = Data(2, i) + sumXY(2, Data(0, i))
Data(0, i))
sumXY(3, Data(0, i)) = 1 + sumXY(3, Data(0, i))
Next i
For i = 1 To numCluster
Centroid(1, i) = sumXY(1, i) / sumXY(3, i)
Centroid(2, i) = sumXY(2, i) / sumXY(3, i)
Next i
'assign all data to the new centroids
isStillMoving = False
For i = 1 To totalData
min = 10 ^ 10 'big number
X = Data(1, i)
Y = Data(2, i)
For j = 1 To numCluster
d = dist(X, Y, Centroid(1, j), Centroid(2, j))
If d < min Then
min = d
cluster = j
End If
Next j
If Data(0, i) <> cluster Then
Data(0, i) = cluster
isStillMoving = True
End If
Next i
Loop
End If
End Sub
```

Latihan 1

Tabel berikut adalah dataset dari 15 mahasiswa yang memprogramkan mata kuliah Data mining. Dari 15 mahasiswa tersebut akan dikelompokkan menjadi 3 bagian yaitu kelompok pintar, sedang dan kurang. Hitung pula nilai SSE nya.

NO	NAMA MAHASISWA	UTS	TUGAS	UAS
1	Roy	89	90	75
2	Sintia	90	71	95
3	Iqbal	70	75	80
4	Dilan	45	65	59
5	Ratna	65	75	53
6	Merry	80	70	75
7	Rudi	90	85	81
8	Hafiz	70	70	73
9	Gede	96	93	85
10	Christian	60	55	48
11	Justin	45	60	58
12	Jesika	60	70	72
13	Ayu	85	90	88
14	Siska	52	68	55
15	Reitama	40	60	7

Latihan 2

Lakukan proses clustering terhadap data berikut. Lakukan pula beberapa eksperimen untuk menentukan k (jumlah klaster) yang paling optimal berdasarkan nilai SSE yang paling minimal. Implementasikan proses clustering tersebut dengan menggunakan MATLAB. Gambarkan pula grafik scatter pada setiap nilai k.

Tabel 1 Data Mahasiswa

No	Nama	Jurusan	Kota Asal	IPK
1	Ade Supryan Stefanus	IS	Jakarta	3,16
2	Adelina Ganardi Putri Hardi	ACC	Semarang	3,22
3	Adeline Dewita	BF	Bekasi	3,29
4	Adiputra	IB	Jakarta	2,83
5	Afrieska Laura Trisyana	PR	Jakarta	3,15
6	Agam Khalilullah	IB	Banda Aceh	3,25
7	Agus Mulyana Jungjungan	IB	Bogor	3,43
8	Agusman	PR	Bekasi	3,06
9	Aidil Friadi	BF	Banda Aceh	3,36
10	Ajeng Putri Ariandhani	ACC	Bandung	3,28

Latihan 2

Transformasi Data Agar data di atas dapat diolah dengan menggunakan metode k-means clustering, maka data yang berjenis data nominal seperti kota asal dan jurusan harus diinisialisasikan terlebih dahulu dalam bentuk angka

Tabel 2 Inisialisasi Data Wilayah Kota Asal

Wilayah	Frekuensi	Inisial
Jakarta	84	1
Jawa Barat	82	2
Sumatera Utara	28	3
Sulawesi	14	4
Jawa Timur	13	5
Sumatera Selatan	13	6
Bali	8	7
Kalimantan	1	8

Setelah semua data mahasiswa ditransformasi ke dalam bentuk angka, maka data-data tersebut telah dapat dikelompokan dengan menggunakan algoritma K-Means Clustering.

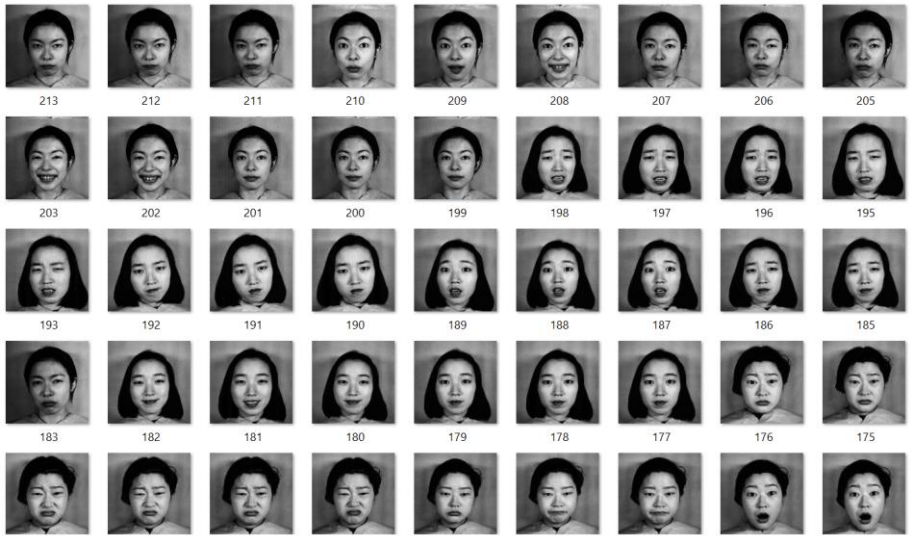
Tabel 3 Inisialisasi Data Jurusan

Jurusan	Singkatan	Frekuensi	Inisial
Accounting	ACC	46	1
Management, concentration in International Business	IB	37	2
Public Relation	PR	35	3
Management, concentration in Banking & Finance	BF	28	4
Industrial Engineering	IE	23	5
Information Technology	IT	20	6
Management, concentration in Marketing	MKT	18	7
Visual Communication Design	VCD	12	8
Management, concentration in Hotel & Tourism Management	HTM	9	9
Electrical Engineering	EE	6	10
Business Administration	BA	4	11
International Relations	IR	2	12
Management, concentration in Human Resources Management	HRM	1	13
Information System	IS	1	14
Management	MGT	1	15

Soal UAS

Lakukan proses clustering terhadap dataset ekspresi wajah yang terdapat di database JAFFE yang berjumlah 213 data dengan menggunakan Algoritma K-Means Clustering. Lakukan pula beberapa eksperimen untuk menentukan nilai k (jumlah klaster) yang paling optimal berdasarkan nilai SSE yang paling minimal. Implementasikan proses clustering tersebut dengan menggunakan MATLAB. Gambarkan pula grafik scatter pada setiap nilai k.

Dataset Ekspresi Wajah (JAFFE Database)



Ketentuan :

Tugas dikerjakan secara berkelompok sesuai dengan data kelompok yang sudah ada. Hasinya dipresentasikan sesuai dengan jadwal UAS matakuliah Sistem Cerdas Kel. E13701.