K-means (clustering)

A centroid - based Technique

Example (2 dimensions)

Individual	variable 1	variable 2
1	1	1
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Euclidean Distance.

$= \sqrt{(x_i - m_i)^2 + (x_j - m_j)^2}$

$x_i\ x_j$ = observed data point

$m_i\ m_j$ = mean / centroid

① $= \sqrt{(1-5)^2 + (1-7)^2}$

$= \sqrt{16 + 36} = \sqrt{52} = 7.21$

② $= \sqrt{(1.5-1)^2 + (2-1)^2}$

$= 1.12$

$\sqrt{(1.5-5)^2 + (2-7)^2}$

$= 6.10$

Step-1 Decide how many clusters

$k = 2$

Step-2 Decide centroids
choose random centroids.

Individual Data point	Distance (E) from C_1 (1,1)	Distance (E) from C_2 (5,7)
1 (1,1)	0	7.21
2 (1.5,2)	1.12	6.10
3 (3,4)	3.61	3.64
4 (5,7)	7.21	0
5 (3.5,5)	4.72	2.5
6 (4.5,5)	5.31	2.06
7 (3.5,4.5)	4.30	2.92

$K_1 = C_1\ (1,1) = \{1, 2, 3\}$

$K_2 = C_2\ (5,7) = \{4, 5, 6, 7\}$

Step-4 New centroid mean C_1 & C_2

No. of data points in k1

$$m_1 = \frac{1}{3}[1+1.5+3] + \frac{1}{3}[1+2+4]$$

$$= 1.83, 2.33$$

No. of data points in k2

$$m_2 = \frac{1}{4}[5+3.5+4.5+3.5] + \frac{1}{4}[7+5+5+4.5]$$

$$= 4.12, 5.38$$

Individual Data points	Distance C_1 (1.83, 2.33)	Distance C_2 (4.12, 5.38)
1(1,1)	1.57	5.32
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

$K_1 = \{1, 2\}$

$K_2 = \{3, 4, 5, 6, 7\}$

Step-5 New centoroids

$$m_1 = \frac{1}{2}[1+1.5] + \frac{1}{2}[1+2]$$

$$= 1.25, 1.5$$

$$m_2 = \frac{1}{5}[3+5+3.5+4.5+3.5] + \frac{1}{5}[4+7+5+5+4.5]$$

$$= 3.9, 5.1$$

Scanned with CamScanner

Individual Data points	Distance $c_1(1.25, 1.5)$	Distance $c_2(3.9, 5.1)$
1	0.56	5.02
2	0.56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

$K_1 = \{1, 2\}$

$K_2 = \{3, 4, 5, 6, 7\}$