Introduction to Data Mining

Homework 1A

Homework group 4:

Arngrimur Einarsson
Troy Macasland
Nick Geertjens
Guðmundur Palsson
Stefan Jonsson

Clustering with k-means:

We have the following dataset in two dimensional Euclidean space

Ω	A	B	C	D	E	F	G	H
x	0	1	1	3	5	5	8	11
y	9	8	5	1	2	0	5	5

We are going to cluster this dataset, using the standard Euclidean distance:

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

α. Cluster the dataset Ω with the k-means algorithm, for k=3, using records A, B and C as the initial centroids. If during the algorithm, a particular record has equal distance to two cluster centroids, choose the one that is closest to the origin (i.e.: the point with the coordinate (0,0)).

Solution: We have our initial centroids $\vec{m}_1 = \begin{Bmatrix} 0 \\ 9 \end{Bmatrix}$, $\vec{m}_2 = \begin{Bmatrix} 1 \\ 8 \end{Bmatrix}$, $\vec{m}_3 = \begin{Bmatrix} 1 \\ 5 \end{Bmatrix}$, then we compute the clustering by these initial centroids.

$$C(i) = \operatorname{argmin}_{1 \le l \le 3} \| \vec{x}_i - \vec{m}_l \|^2$$

$$C(D) = \operatorname{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(3-0)^2 + (1-9)^2} = 8,5 \\ \sqrt{(3-1)^2 + (1-8)^2} = 7.2 \\ \sqrt{(3-1)^2 + (1-5)^2} = 4.5 \end{cases}$$

so we assign point D to cluster k_3

We use the same method for point E

$$C(E) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(5-0)^2 + (2-9)^2} = 8.6 \\ \sqrt{(5-1)^2 + (2-8)^2} = 7.2 \\ \sqrt{(5-1)^2 + (2-5)^2} = 5 \end{cases}$$

so we assign point E to cluster k_3

We use the same method for point F

$$C(F) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(5-0)^2 + (0-9)^2} = 10.29 \\ \sqrt{(5-1)^2 + (0-8)^2} = 8,944 \\ \sqrt{(5-1)^2 + (0-5)^2} = 6.4 \end{cases}$$

so we assign point F to cluster k_3

We use the same method for point G

$$C(G) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(8-0)^2 + (5-9)^2} = 8,9 \\ \sqrt{(8-1)^2 + (5-8)^2} = 7.6 \\ \sqrt{(8-1)^2 + (5-5)^2} = 7 \end{cases}$$

so we assign point G to cluster k_3

We use the same method for point

$$C(H) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(11-0)^2 + (5-9)^2} = 11.7 \\ \sqrt{(11-1)^2 + (5-8)^2} = 10.44 \\ \sqrt{(11-1)^2 + (5-5)^2} = 10 \end{cases}$$

so we assign point H to cluster k_3

Then we recompute the centroids based on the new clustering $C(i)$, we move each cluster center to the mean of its observations

$$\vec{m}_1' = \begin{Bmatrix} m_x' \\ m_y' \end{Bmatrix} = \begin{Bmatrix} 0 \\ 9 \end{Bmatrix}$$ unchanged!

$$\vec{m}_2' = \begin{Bmatrix} m_x' \\ m_y' \end{Bmatrix} = \begin{Bmatrix} 1 \\ 8 \end{Bmatrix}$$ unchanged!

$$\vec{m}_3' = \begin{Bmatrix} m_x' \\ m_y' \end{Bmatrix} = \begin{Bmatrix} (1+3+5+5+8+11)/6 \\ (5+1+2+0+5+5)/6 \end{Bmatrix}$$

$$= \begin{Bmatrix} 5.5 \\ 3 \end{Bmatrix}$$ new center for k_3

The next iteration is calculated the same way, we now have new cluster centroids and we need to reassign cluster for each point

$$C(D) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(3-0)^2+(1-9)^2} = 8.5 \\ \sqrt{(3-1)^2+(1-8)^2} = 7.2 \\ \sqrt{(3-5.5)^2+(1-3)^2} = 3.2 \end{cases}$$

so we assign point D to cluster k_3

$$C(E) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(5-0)^2+(2-9)^2} = 8.6 \\ \sqrt{(5-1)^2+(2-8)^2} = 7.2 \\ \sqrt{(5-5.5)^2+(2-3)^2} = 1.11 \end{cases}$$

so we assign point E to cluster k_3

$$C(F) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(5-0)^2+(0-9)^2} = 10.29 \\ \sqrt{(5-1)^2+(0-8)^2} = 8,944 \\ \sqrt{(5-5.5)^2+(0-3)^2} = 3.04 \end{cases}$$

so we assign point E to cluster k_3

$$C(G) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(8-0)^2+(5-9)^2} = 8,9 \\ \sqrt{(8-1)^2+(5-8)^2} = 7.6 \\ \sqrt{(8-5.5)^2+(5-3)^2} = 3.2 \end{cases}$$

so we assign point E to cluster k_3

$$C(C) = \text{argmin}_{1 \le \ell \le 3} \begin{cases} \sqrt{(1-0)^2+(5-9)^2} = 4.1 \\ \sqrt{(1-1)^2+(5-8)^2} = 3 \\ \sqrt{(1-5.5)^2+(5-3)^2} = 4.9 \end{cases}$$

so we assign point C to cluster k_2

$$C(H) = \text{argmin}_{1 \le \ell \le 3} \begin{cases} \sqrt{(11-0)^2+(5-9)^2} = 11.7 \\ \sqrt{(11-1)^2+(5-8)^2} = 10.44 \\ \sqrt{(11-5.5)^2+(5-3)^2} = 5.85 \end{cases}$$

so we assign point H to cluster k_3

Then we recompute the centroids

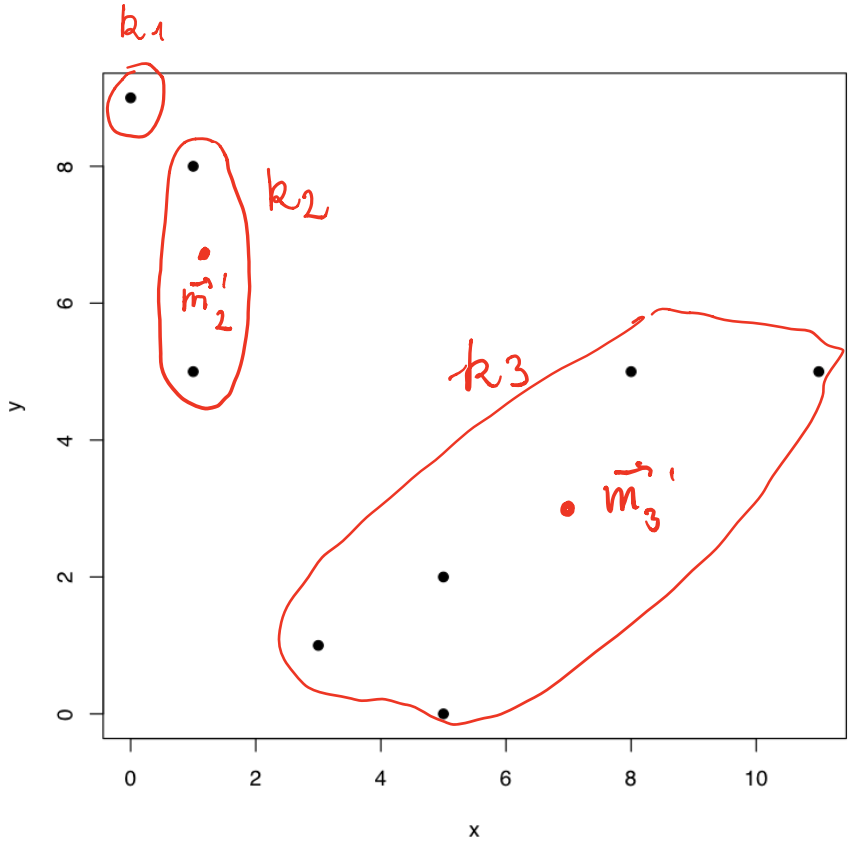
$$\vec{m}_1' = \begin{Bmatrix} m_x' \\ m_y' \end{Bmatrix} = \begin{Bmatrix} 0 \\ 9 \end{Bmatrix}$$ unchanged!

$$\vec{m}_2' = \begin{Bmatrix} m_x' \\ m_y' \end{Bmatrix} = \begin{Bmatrix} (1+1)/2 \\ (8+5)/2 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 6.5 \end{Bmatrix}$$

$$\vec{m}_3' = \begin{Bmatrix} m_x' \\ m_y' \end{Bmatrix} = \begin{Bmatrix} (3+5+5+8+11)/5 \\ (1+2+0+5+5)/5 \end{Bmatrix}$$

$$= \begin{Bmatrix} 6.4 \\ 2.6 \end{Bmatrix}$$

After two iterations!



The next iteration is calculated the same way, we now have new cluster centroids and we need to reassign cluster for each point

$$C(D) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(3-0)^2+(1-9)^2} = 8.5 \\ \sqrt{(3-1)^2+(1-6.5)^2} = 5.8 \\ \sqrt{(3-6.4)^2+(1-2.6)^2} = 3.75 \end{cases}$$

so we assign point D to cluster k_3

$$C(E) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(5-0)^2+(2-9)^2} = 8.6 \\ \sqrt{(5-1)^2+(2-6.5)^2} = 6.02 \\ \sqrt{(5-6.4)^2+(2-2.6)^2} = 1.52 \end{cases}$$

so we assign point E to cluster k_3

$$C(F) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(5-0)^2+(0-9)^2} = 10.29 \\ \sqrt{(5-1)^2+(0-6.4)^2} = 7.6 \\ \sqrt{(5-6.4)^2+(0-2.6)^2} = 2.95 \end{cases}$$

so we assign point E to cluster k_3

$$C(G) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(8-0)^2+(5-9)^2} = 8,9 \\ \sqrt{(8-1)^2+(5-6.4)^2} = 7.2 \\ \sqrt{(8-6.4)^2+(5-2.6)^2} = 2.88 \end{cases}$$

so we assign point E to cluster k_3

$$C(C) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(1-0)^2+(5-9)^2} = 4,12 \\ \sqrt{(1-1)^2+(5-6.4)^2} = 1.5 \\ \sqrt{(1-6.4)^2+(5-2.6)^2} = 5.9 \end{cases}$$

so we assign point C to cluster k_2

$$C(H) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(11-0)^2+(5-9)^2} = 11.7 \\ \sqrt{(11-1)^2+(5-6.4)^2} = 10.11 \\ \sqrt{(11-6.4)^2+(5-2.6)^2} = 5.18 \end{cases}$$

so we assign point H to cluster k_3

$$C(B) = \text{argmin}_{1 \le l \le 3} \begin{cases} \sqrt{(1-0)^2+(8-9)^2} = 1.41 \\ \sqrt{(1-1)^2+(8-6.4)^2} = 1.5 \\ \sqrt{(1-6.4)^2+(8-2.6)^2} = 7.6 \end{cases}$$

So we assign point B to cluster k_1

Then we recompute the centroids

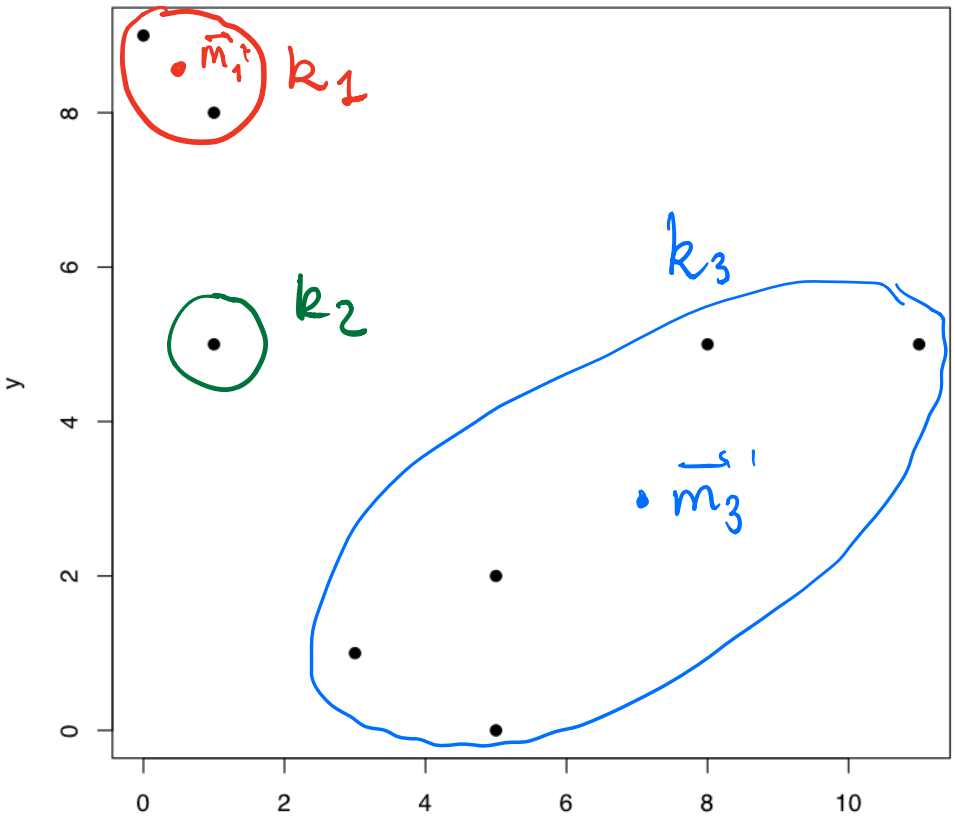
$$\vec{m}_1' = \begin{Bmatrix} m_x' \\ m_y' \end{Bmatrix} = \begin{Bmatrix} (0+1)/2 \\ (9+8)/2 \end{Bmatrix} = \begin{Bmatrix} 0.5 \\ 8.5 \end{Bmatrix}$$

$$\vec{m}_2' = \begin{Bmatrix} m_x' \\ m_y' \end{Bmatrix} = \begin{Bmatrix} 1 \\ 8 \end{Bmatrix}$$

$$\vec{m}_3' = \begin{Bmatrix} m_x' \\ m_y' \end{Bmatrix} = \begin{Bmatrix} 6.4 \\ 2.6 \end{Bmatrix}$$

After this, our solution converges!

Final result:



β: Again, cluster the dataset Ω with the k-means algorithm, for k=3, but this time using E, F and G as the initial centroids What do we observe?

Solution.

Now we use $c_1 = \begin{Bmatrix} 5 \\ 2 \end{Bmatrix}$, $c_2 = \begin{Bmatrix} 5 \\ 0 \end{Bmatrix}$, $c_3 = \begin{Bmatrix} 8 \\ 5 \end{Bmatrix}$.

Lets make a distance matrix where the distance is calculated using

$$d\left((x_1, y_1), (x_2, y_2)\right) = \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2}$$

and each column represents a distant of a point to some centroid.

$$D^0 = \begin{array}{c} \begin{matrix} A & B & C & D & E & F & G & H \end{matrix} \\ \begin{bmatrix} 8.6 & 7.2 & 5 & 2.24 & 0 & 2 & 4.2 & 6.7 \\ 10.3 & 8.9 & 6.4 & 2.24 & 2 & 0 & 5.8 & 7.8 \\ 8.9 & 7.6 & 7 & 6.4 & 4.2 & 5.8 & 0 & 3 \end{bmatrix} \end{array} \begin{array}{l} c_1 = (5,2) \\ c_2 = (5,0) \\ c_3 = (8,5) \end{array}$$

Now if we look at each column and arrange every point to the cluster with the lowest distance we get clusters k_1, k_2, k_3:

$k_1 = A(0,9), B(1,8), C(1,5), E(5,2)$

$k_2: D(3,1), F(5,0)$

$k_3: G(8,5), H(11,5)$

When we recompute the centroids we get:

$$C_1 = \left(\left(\frac{0+1+1+5}{4} \right), \left(\frac{9+8+5+2}{4} \right) \right) = (1{,}75 ; 6)$$

$$C_2 = \left(\left(\frac{3+5}{2} \right), \left(\frac{1+0}{2} \right) \right) = (4 ; 0{,}5)$$

$$C_3 = \left(\left(\frac{8+11}{2} \right), \left(\frac{5+5}{2} \right) \right) = (9{,}5 ; 5)$$

Now lets use our new centroids and make a new distance matrix.

$$D' = \begin{array}{c} \begin{matrix} A & B & C & D & E & F & G & H \end{matrix} \\ \left[\begin{matrix} 3.5 & 2.2 & 1.25 & 5.2 & 5.2 & 6.8 & 6.3 & 9.3 \\ 9.4 & 8.1 & 5.4 & 1.1 & 1.8 & 1.1 & 6 & 8.3 \\ 10 & 9 & 8.5 & 7.6 & 5.4 & 6.7 & 4.5 & 1.5 \end{matrix}\right] \end{array} \begin{matrix} C_1 = (1.75; 6) \\ C_2 = (4, 0.5) \\ C_3 = (9.5; 5) \end{matrix}$$

we get clusters:

$K_1 = A(0,9), B(1,8), C(1,5)$

$K_2 = D(3,1), E(5,2), F(5,0)$

$K_3 = G(8,5), H(11,5)$

When we recompute the centroids we get

$$C_1 = \left(\left(\frac{1+1}{3}\right), \left(\frac{9+8+5}{3}\right)\right) = (0,66 ; 7,33)$$

$$C_2 = \left(\left(\frac{3+5+5}{3}\right), \left(\frac{1+2+0}{3}\right)\right) = (4,3 , 1)$$

$$C_3 = \left(\frac{8+11}{2}, \frac{5,5}{2}\right) = (9,5, 5)$$

We get the new distance matrix

$$D^2 = \begin{array}{c} \begin{matrix} A & B & C & D & E & F & G & H \end{matrix} \\ \begin{bmatrix} 1,8 & 0,8 & 2,4 & 6,7 & 6,9 & 8,5 & 7,7 & 10,6 \\ 9,1 & 7,7 & 5,2 & 1,3 & 1,2 & 1,2 & 5,4 & 7,8 \\ 10,3 & 9 & 8,5 & 7,6 & 5,4 & 6,7 & 4,5 & 4,5 \end{bmatrix} \end{array} \begin{matrix} C_1 = (0,66 : 7,33) \\ C_2 = (4,3;1) \\ C_3 = (9,5; 5) \end{matrix}$$

We can see we get the same clusters as befor:

$K_1 = A(0,9), B(1,8), C(1,5)$

$K_2 = D(3,1), E(5,2), F(5,0)$

$K_3 = G(8,5), H(11,5)$

So the algorithm has reached equalibrium.

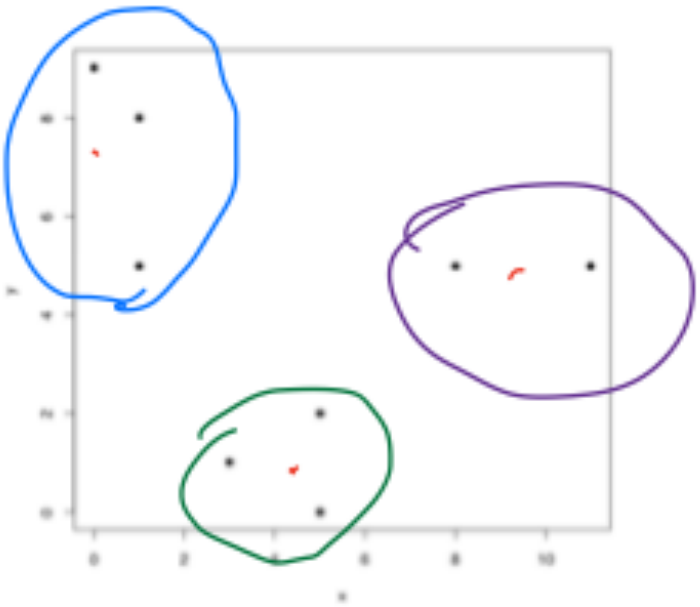


Figure 1: The Flat-Table Dataset Ω.

γ. For both clusterings you have obtained in exercises α, and β, compute the within cluster distance $W(C)$ and the between cluster dissimilarity $B(C)$. Briefly interpret your results: what does this tell us about the quality of the clusterings?

Clustering in α: We use the formula

$$W(C) = \frac{1}{2} \sum_{k=1}^{k} \sum_{C(i)=k} \sum_{C(i')=k} D(x_i, x_{i'})$$

$$B(C) = \frac{1}{2} \sum_{k=1}^{k} \sum_{C(i)=k} \sum_{C(i')\neq k} D(x_i, x_{i'})$$

Lets first compute $W(C)$ for α and β using

$$d\left((x_1,y_1),(x_2,y_2)\right) = \sqrt{(x_2-x_1)^2-(y_2-y_1)^2}$$

and

$$W(C) = \frac{1}{2}\sum_{l=1}\sum_{C(i)=l}\sum_{C(i')=l} D(\overline{x_i}, \overline{x_{i'}})$$

Lets calculate the distances within each cluster first and then add the sum up. Starting with α:

$k_1 : A(0,9), B(1,8)$

$$S_1 = 2\cdot\sqrt{(0-1)^2-(9-8)^2} = 2\sqrt{2}$$

$k_2 = C(1,5)$

$S_2 = 0$

$k_3 = D(3,1), E(5,2), F(5,0), G(8,5), H(11,5$

Here lets use a distance table where a square represents a distance between two points Then we sum up all the distances (squares)

S_3:

	D	E	F	G	H
D	0	2,23	2,23	6,4	8,9
E	2,23	0	2	4,24	6,7
F	2,23	2	0	5,8	7,8
G	6,4	4,24	5,8	0	3
H	8,9	6,7	7,8	3	0

$= 98,6$

$$S_{nm} = S_1 + S_2 + S_3 = 2\sqrt{2} + 0 + 98,6$$

$$= 101,43$$

so $w(C) = \frac{1}{2} \cdot 101,43 = 50,71$

B:

$K_1 = A(0,9) \cdot B(1,8) \cdot C(1,5)$

	A	B	C
A	0	1,4	4,1
B	1,4	0	3
C	4,1	3	0

$S_1 = 17$

$k_2 = D(3,1), E(5,2), F(5,0)$

	D	E	F
D	0	2,24	2,24
E	2,24	0	2
F	2,24	2	0

$S_2 = 12,96$

$k_3 = G(8,5), H(11,5)$

$S_3 = 2 \cdot d(G, H) = 6$

$$Sum = S_1 + S_2 + S_3 = 35,96$$

So

$$w(c) = \frac{1}{2} Sum = 17,98$$

	D	E	F	G	H
A	8.54	8.6	10.3	8.94	11.7
B	7.28	7.21	8.94	7.62	10.44
C	4.47	5	6.4	7	10

$$k_1 = 122,414$$

	A	B	C	G	H
D	8,54	7,28	4.47	6,4	8,94
E	8,6	7,2	5	4,24	6,70
F	10.29	8,94	6,4	5,89	7,8

$$k_2 = 199,69$$

	A	B	C	D	E	F
G	8,94	7,62	7	6,4	4,24	5,89
H	11,7	10,44	10	8,94	6,7	7,8

$$k_3 = 95,72$$

$$\frac{k_1 + k_2 + k_3}{2} = \frac{B(C)2}{208,93}$$

For the clustering in β we get

	D	E	F	G	H
A	8.5	8.6	10.3	8.9	11.7
B	7.3	7.2	8.9	7.6	10.4

$\sum = 89.4$

	A	B	D	E	F	G	H
C	4.1	3	4.5	5	6.4	7	10

$\sum = 40$

	A	B	C
D	8.5	7.3	4.5
E	8.6	7.2	5
F	10.3	8.9	6.4
G	8.9	7.6	7
H	11.7	10.4	10

$\sum = 122.3$

$$\frac{1}{2}\sum_{tot} = 122.05 = B(C)$$

	W(C)	B(C)
α	50.7	125,85
β	17,98	208,93

One should minimize the distance within clusters, W(C), and maximize the distanc between clusters, B(C). From the table we see that β is clustering of better quality.