Chapter 7:

* K-means clustering:

Eg:

Instance	X	Y
1	1	1.5
2	1	4.5
3	2	1.5
4	2	3.5
5	3	2.5
6	3	4

Take (K = 2).

Soln:

Given the number of cluster to be created (K) = 2.
So, initially choose two points randomly as an initial cluster center, say data points 1 and 3 is choosen.

∴ centeriod of first cluster $c_1 = (1, 1.5)$

∴ centeriod of second cluster $c_2 = (2, 1.5)$

~~Let #1~~

Iteration 1:

$$d(c_1, 2) = \sqrt{(1-1)^2 + (1.5-4.5)^2} = 3$$

$$d(c_2, 2) = \sqrt{(2-1)^2 + (1.5-4.5)^2} = 3.16$$

So, the data belongs to c_1 as $d(c_1, 2) < d(c_2, 2)$

So, data point 2 belongs to c_1.

$$d(c_1, 4) = \sqrt{(1-2)^2 + (1.5-3.5)^2} = 2.23$$

$$d(c_2, 4) = \sqrt{(2-2)^2 + (1.5-3.5)^2} = 2$$

$d(c_1, 4) > d(c_2, 4)$ So, the data, belongs to c_2

point 4

$d(C_1, 5) = \sqrt{(1-3)^2 + (1.5-2.5)^2} = 2.23$

$d(C_2, 5) = \sqrt{(2-3)^2 + (1.5-2.5)^2} = 1.44$

$d(C_1, 5) > d(C_2, 5)$. So,

the data point 5 belongs to C_2.

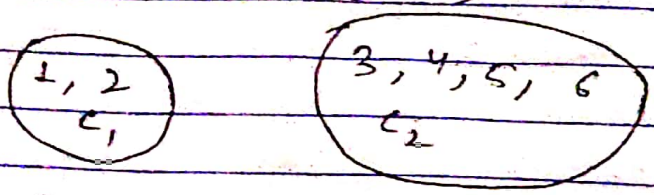
$d(C_1, 6) = \sqrt{(1-3)^2 + (1.5-4)^2} = 3.2$

$d(C_2, 6) = \sqrt{(2-3)^2 + (1.5-4)^2} = 2.69$

Here $d(C_1, 6) > d(C_2, 6)$ so,

data point 6 belongs to C_2.

So, the cluster becomes



Iteration 2:

Now, calculating new centroid for each clusters as

$C_1 = \left(\frac{1+1}{2}, \frac{1.5+4.5}{2}\right) = (1, 3)$

$C_2 = \left(\frac{2+2+3+3}{4} + \frac{1.5+3.5+2.5+4}{4}\right) = (2.5, 2.87)$

Now, again calculating similarity:

$d(C_1, 1) = \sqrt{(1-1)^2 + (3-1.3)^2} = 1.5$

$d(C_2, 1) = \sqrt{(2.5-1)^2 + (2.87-1.5)^2} = 2.03$

$d(C_1, 1) < d(C_2, 1)$ so,

data point 1 belongs to C_1.

$d(C_1, 2) = \sqrt{(1-1)^2 + (3-4.5)^2} = 1.5$

$d(C_2, 2) = \sqrt{(2.5-1)^2 + (2.87-4.5)^2} = 2.21$

$d(C_1, 2) < d(C_2, 2)$ so,

data point 2 belongs to C_1.

$d(C_1, 3) = \sqrt{(1-2)^2 + (3-1.5)^2} = 1.8$

$d(C_2, 3) = \sqrt{(2.5-2)^2 + (2.87-1.5)^2} = 1.46$

$d(c_1, 3) > d(c_2, 3)$ so

data point 3 belongs to c_2.

$d(c_1, 4) = \sqrt{(1-2)^2 + (3-3.5)^2} = 1.12$

$d(c_2, 4) = \sqrt{(2.5-2)^2 + (2.87-3.5)^2} = 0.8$

$d(c_1, 4) > d(c_2, 4)$ so,

data point 4 belongs to c_2.

$d(c_1, 5) = \sqrt{(1-3)^2 + (3-2.5)^2} = 2.06$

$d(c_2, 5) = \sqrt{(2.5-3)^2 + (2.87-2.5)^2} = 0.625$

$d(c_1, 5) > d(c_2, 5)$ so,

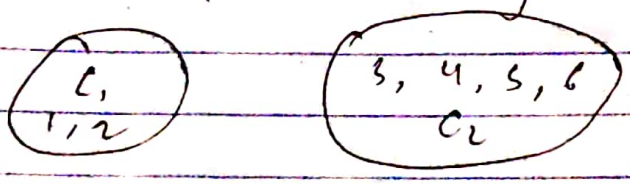
data point 5 belongs to c_1.

$d(c_1, 6) = \sqrt{(1-3)^2 + (3-4)^2} = 2.236$

$d(c_2, 6) = \sqrt{(2.5-3)^2 + (2.875-4)^2} = 0.176$

Here, $(c_2, 6) < d(c, 6)$

So, data points 6 belongs to c_2.



K-mediod Example:

i	x	y
x_1	2	6
x_2	3	4
x_3	3	8
x_4	4	7
x_5	6	2
x_6	6	4
x_7	7	3
x_8	7	4
x_9	8	5
x_{10}	7	6

K=2

Step 1:

We select two random representative objects

$c_1(3, 4)$, $c_2(7, 4)$

Step 2: for c_1,

i	x	y	c_1		distance/cost
x_1	2	6	3	4	$\|2-3\|+\|6-4\| = 3$
x_2	3	8	3	4	$\|3-3\|+\|8-4\| = 4$
x_4	4	7	3	4	$1+3 = 4$
x_5	6	2	3	4	$3+2 = 5$
x_6	6	4	3	4	$3+0 = 3$
x_7	7	3	3	4	$4+1 = 5$
x_9	8	5	3	4	$4+1 = 5$
x_{10}	7	6	3	4	$4+2 = 6$

for C_2

i	x	y	C_2		distance | cost
x_1	2	6	7	4	$\|2-7\|+\|6-4\| = 7$
x_3	3	8	7	4	$\|3-7\|+\|8-4\| = 8$
x_4	4	7	7	4	$\|4-7\|+\|7-4\| = 6$
x_5	6	2	7	4	3
x_6	6	4	7	4	L
x_7	7	3	7	4	L
x_9	8	5	7	4	L
x_{10}	7	6	7	4	2

Step 3:

Compair the cost of C_1 and C_2 for every i and select the minimum one.

Cluster 1 = {(2,6), {3,8}, {4,7}, {3,4}}

Cluster 2 = {(6,2), (6,4), (7,3), (7,4), (8,5), (7,6)}

Step 4:

Total cost = (3+4+4)+(3+1+1+1+2)

= 20

Step 5:

Select one of non-medoids O'. Let's select

O' = (7,3) from i.e x_7.

Step 6: Repeat step 2

for o'

i	x	y	o'	Distance / cost
x_1	2	6	7 3	8
x_3	3	8	7 3	9
x_4	4	7	7 3	7
x_5	6	2	7 3	2
x_6	6	4	7 3	2
x_8	7	4	7 3	1
x_9	6	5	7 3	3
x_{10}	7	6	7 3	3

for c_1

i	x	y	c_1	Distance / cost
x_1	2	6	3 4	3
x_3	3	8	3 4	4
x_4	4	7	3 4	4
x_5	6	2	3 4	5
x_6	6	4	3 4	3
x_8	7	4	3 4	4
x_9	6	5	3 4	6
x_{10}	7	6	3 4	6

Now comparing the cost of c_1 and o' for each i, so the new cluster becomes

Cluster 1 = {(2,6), (3,4), (4,7), (3,8)}

Cluster 2 = {(7,3), (6,2), (6,4), (7,4), (8,5), (7,6)}

Total cost = (3+4+4) + (2+2+1+3+5)

= 22.

~~Since~~

~~Total cost from c and o'~~

Step:6

$S \leq$ current total cost − past total cost

$= 22 - 22$

$= 2 > 0$

So, moving 0' would be a bad idea, so previous choice was good.