-CLUSTERING-

medicine example (slide 52 lec 2 week 6) hand calculations

1- Given values: • input data (see table)

• initial ~~clusters~~ centroids $C_1 = (1,1)$ $C_2 = (2,1)$

4 samples	x_1	x_2	
Medicine	weight index	PH	
A	1	1	$\rightarrow (1,1)$ ← same position as C_1
B	2	1	$\rightarrow (2,1)$ ← same position as C_2
C	4	3	$\rightarrow (4,3)$
D	5	4	$\rightarrow (5,4)$

Iteration #1:

2- Calculate the distance of each data point to each centroid C_1 & C_2

We don't need to calculate this for medicine A since it's position is the same as $C_1 = (1,1)$ & also for medicine B which is similar in position to $C_2 = (2,1)$.

position of medicine C from the data table is (4,3)

d is the distance between medicine C & centroid C_1

$d(\text{med } C, C_1) = (4-1)^2 + (3-1)^2 = 3^2 + 2^2 = 9 + 4 = \underline{13}$ ($C_1 = (1,1)$)

$d(\text{med } C, C_2) = (4-2)^2 + (3-1)^2 = 2^2 + 2^2 = 4 + 4 = \underline{8}$ ($C_2 = (2,1)$)

$d(\text{med } D, C_1) = (5-1)^2 + (4-1)^2 = 4^2 + 3^2 = 16 + 9 = \underline{25}$

$d(\text{med } D, C_2) = (5-2)^2 + (4-1)^2 = 3^2 + 3^2 = 9 + 9 = \underline{18}$

* for medicine C, comparing 13 & 8, we can see that this data point is closer to centroid C_2 rather than C_1. so it will be assigned to C_2 (the second cluster). The same is true for med D.

3- create / update the assignment matrix

based on the distance calculations, our assignment matrix after 1 iteration is:

$$X = \begin{matrix} \\ A \\ B \\ C \\ D \end{matrix} \begin{matrix} K_1 & K_2 \leftarrow \text{the clusters} \\ \end{matrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

our samples

1 means that the sample belongs to the cluster. 0 means that it's doesn't belong to the cluster. each data point can only be assigned to one cluster.

4- update the position of each centroid C_1 & C_2: by taking the average position of all of the data points it contains.

for C_1 we only have one data point @ (1,1) so it stays as (1,1)

for $C_2 = \left(\frac{\text{med B} + \text{med C} + \text{med D}}{3}\right) = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = (3.67, 2.67)$

(med B = (2,1); med C = (4,3))

C_2 is was in position (2,1) but is now in position (3.67, 2.67)

Iteration #2: (repeat same steps)

1- calculate distances

Med A is still the same position as $c_1 = (1,1)$

For med B = $d(\text{med B}, c_1) = (2-1)^2 + (1-1)^2 = 1$

$d(\text{med B}, c_2) = (2-3.67)^2 + (1-2.67)^2 = 5.57$

] med B is now closer to C1

$c_2 = (3.67, 2.67)$

med B = (2,1)

for med C = $d(\text{med C}, c_1) = (4-1)^2 + (3-1)^2 = 13$

$d(\text{med C}, c_2) = (4-3.67)^2 + (3-2.67)^2 = 0.22$

] med C is closer to c_2

for med D = $d(\text{med D}, c_1) = (5-1)^2 + (4-1)^2 = 25$

$d(\text{med D}, c_2) = (5-3.67)^2 + (4-2.67)^2 = 3.53$

] med D is closer to c_2

update assignment matrix accordingly

2- update assignment matrix:

$$X = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} K_1 & K_2 \end{array} \leftarrow \text{clusters} \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \end{array}$$

samples ↗

⇒ compared to the previous assignment matrix, med B now moved to cluster 1.

3- update the position of each centroid:

cluster #1 has med A & med B, we will take the average position of both data points & that will be the new position of c_1.

medians positions from the data table. They don't change. only the centroids positions change.

med A = (1,1)

med B = (2,1)

$$c_1 = \left(\frac{\text{med A} + \text{med B}}{2}\right) = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1.5, 1)$$

$$c_2 = \left(\frac{\text{med C} + \text{med D}}{2}\right) = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4.5, 3.5)$$

since the centroid positions changed as well as the assignment matrix, we proceed calculating the 3rd iteration. In that iteration you will see that the assignment matrix & centroid positions doesn't change. so you terminate the algorithm.