Inteligenţă artificială

Tema

Algoritmul *K-means clustering* este o metodă de determinare a clusterelor pe care le formează mai multe pattern-uri. Procedura este una de instruire nesupervizată. Se presupune cunoscut numărul *K* al clusterelor, acesta fiind un parametru stabilit *a priori*.

Fiecare cluster are un centroid. Algoritmul lucrează cu *K* clustere, deci *K* dintre punctele folosite la instruire vor fi centriozii celor *K* clustere. Întrucât iniţializarea centroizilor se face aleator, există posibilitatea ca mai multe rulări ale algoritmului să conducă la rezultate diferite.

Fiecare punct este asociat clusterului determinat de cel mai apropiat centriod. Distanţa dintre punct şi centriod poate fi calculată, de exemplu, ca distanţă euclidiană, dar se poate opta şi pentru alte variante.

Algoritmul este următorul:

1. *Se aleg aleator K puncte ca centroizi iniţiali.*
2. *Se formează K clustere prin asignarea tuturor punctelor celor mai apropiaţi centroizi.*
3. *Se recalculează centroizii astfel: noul centriod va fi centrul de greutate determinat de punctele clusterului.*
4. *Se reiau paşii 2 şi 3 până când centroizii nu se mai modifică.*

Aplicaţi algoritmul *K-means clustering* pentru următoarele puncte, cu *K*=3:

```
P1: 45 85
P2: 50 43
P3: 40 80
P4: 55 42
P5: 200 43
P6: 48 40
P7: 195 41
P8: 43 87
P9: 190 40
```

Listaţi cei 3 centroizi şi punctele asociate fiecărui cluster.

Mai jos găsiți un exemplu numeric de aplicare a algoritmului k-means clustering.

K-Means Clustering – Example

Problem: Cluster the following eight points (with (x, y) representing locations) into three clusters A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9). Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points $a=(x1, y1)$ and $b=(x2, y2)$ is defined as: $\rho(a, b) = |x2 - x1| + |y2 - y1|$.
Use k-means algorithm to find the three cluster centers after the second iteration.

Solution:

Iteration 1

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)				
A2	(2, 5)				
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

First we list all points in the first column of the table above. The initial cluster centers – means, are (2, 10), (5, 8) and (1, 2) - chosen randomly. Next, we will calculate the distance from the first point (2, 10) to each of the three means, by using the distance function:

point	mean1
$x1, y1$	$x2, y2$
(2, 10)	(2, 10)

$$\rho(a, b) = |x2 - x1| + |y2 - y1|$$

$$\begin{aligned}\rho(point, mean1) &= |x2 - x1| + |y2 - y1| \\ &= |2 - 2| + |10 - 10| \\ &= 0 + 0\end{aligned}$$

$= 0$

point	mean2
x1, *y1*	*x2*, *y2*
(2, 10)	(5, 8)

$$\rho(a, b) = |x2 - x1| + |y2 - y1|$$

$$\begin{aligned}\rho(point, mean2) &= |x2 - x1| + |y2 - y1| \\ &= |5 - 2| + |8 - 10| \\ &= 3 + 2 \\ &= 5\end{aligned}$$

point	mean3
x1, *y1*	*x2*, *y2*
(2, 10)	(1, 2)

$$\rho(a, b) = |x2 - x1| + |y2 - y1|$$

$$\begin{aligned}\rho(point, mean2) &= |x2 - x1| + |y2 - y1| \\ &= |1 - 2| + |2 - 10| \\ &= 1 + 8 \\ &= 9\end{aligned}$$

So, we fill in these values in the table:

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)	0	5	9	1
A2	(2, 5)				
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

So, which cluster should the point (2, 10) be placed in? The one, where the point has the shortest distance to the mean – that is mean 1 (cluster 1), since the distance is 0.

Cluster 1	Cluster 2	Cluster 3
(2, 10)		

So, we go to the second point (2, 5) and we will calculate the distance to each of the three means, by using the distance function:

point	mean1
x1, y1	*x2, y2*
(2, 5)	(2, 10)

$$\rho(a, b) = |x2 - x1| + |y2 - y1|$$

$$\begin{aligned}\rho(point, mean1) &= |x2 - x1| + |y2 - y1| \\ &= |2 - 2| + |10 - 5| \\ &= 0 + 5 \\ &= 5\end{aligned}$$

point	mean2
x1, y1	*x2, y2*
(2, 5)	(5, 8)

$$\rho(a, b) = |x2 - x1| + |y2 - y1|$$

$$\begin{aligned}\rho(point, mean2) &= |x2 - x1| + |y2 - y1| \\ &= |5 - 2| + |8 - 5| \\ &= 3 + 3 \\ &= 6\end{aligned}$$

point	mean3
x1, y1	*x2, y2*
(2, 5)	(1, 2)

$$\rho(a, b) = |x2 - x1| + |y2 - y1|$$

$$
\begin{aligned}
\rho(point, mean2) &= |x2 - x1| + |y2 - y1| \\
&= |1 - 2| + |2 - 5| \\
&= 1 + 3 \\
&= 4
\end{aligned}
$$

So, we fill in these values in the table:

Iteration 1

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

So, which cluster should the point (2, 5) be placed in? The one, where the point has the shortest distance to the mean – that is mean 3 (cluster 3), since the distance is 0.

Cluster 1	Cluster 2	Cluster 3
(2, 10)		(2, 5)

Analogically, we fill in the rest of the table, and place each point in one of the clusters:

Iteration 1

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	(5, 8)	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	(1, 2)	9	10	0	3
A8	(4, 9)	3	2	10	2

Cluster 1	Cluster 2	Cluster 3
(2, 10)	(8, 4)	(2, 5)
	(5, 8)	(1, 2)

(7, 5)
(6, 4)
(4, 9)

Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

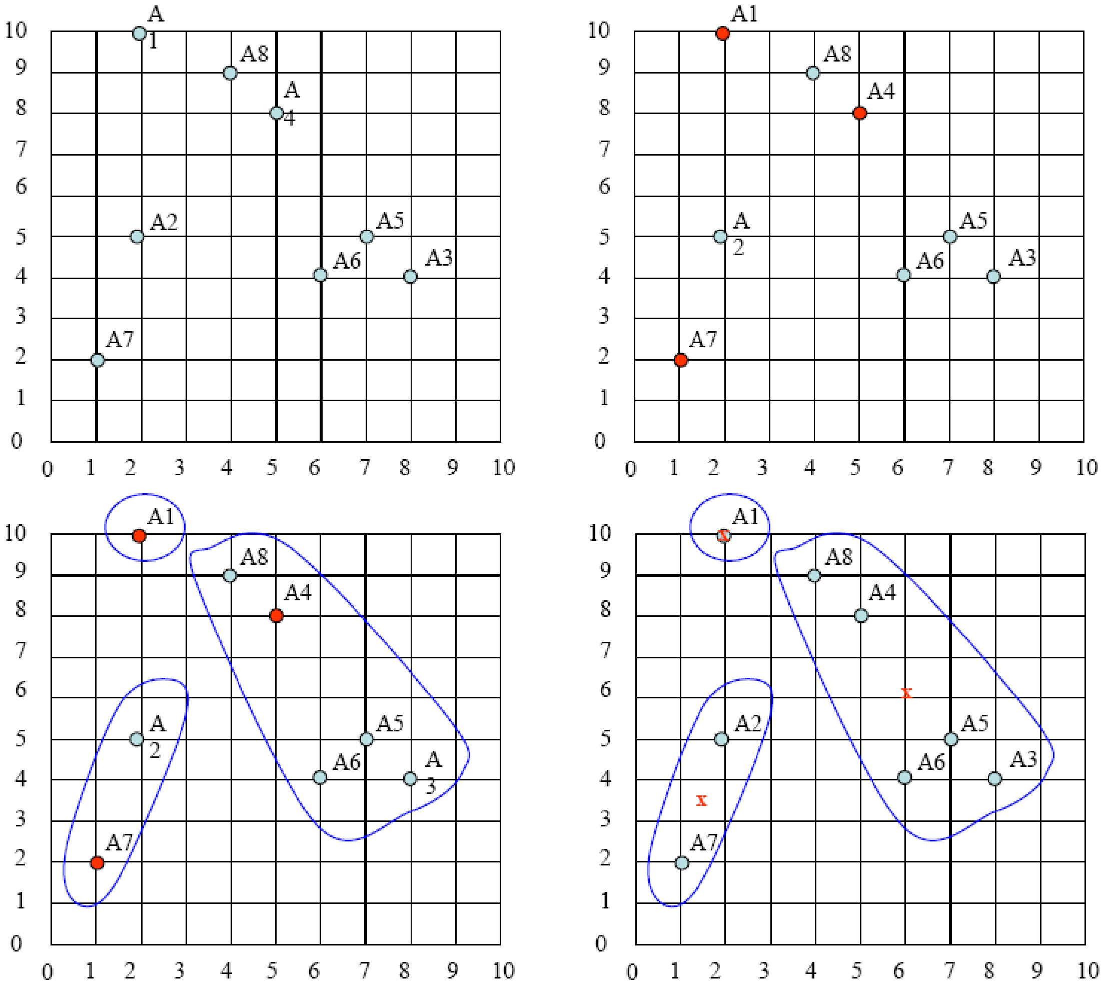
For Cluster 1, we only have one point A1(2, 10), which was the old mean, so the cluster center remains the same.

For Cluster 2, we have ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)

For Cluster 3, we have ((2+1)/2, (5+2)/2) = (1.5, 3.5) new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

b) centers of the new clusters:
C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

c)

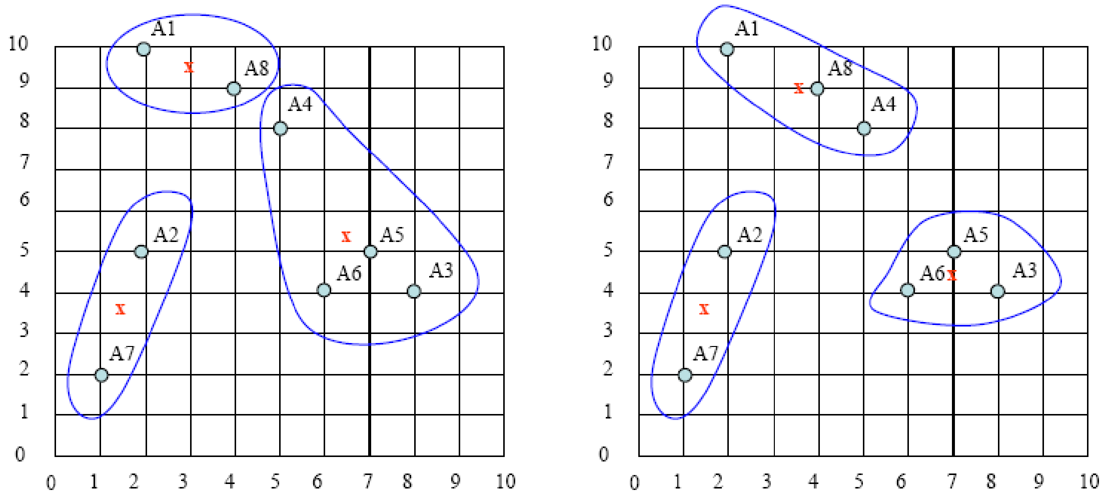


The initial cluster centers are shown in red dot. The new cluster centers are shown in red x.

That was Iteration1 (epoch1). Next, we go to Iteration2 (epoch2), Iteration3, and so on until the means do not change anymore.
In Iteration2, we basically repeat the process from Iteration1 this time using the new means we computed.

d)

We would need two more epochs. After the 2nd epoch the results would be:
1: {A1, A8}, 2: {A3, A4, A5, A6}, 3: {A2, A7}
with centers C1=(3, 9.5), C2=(6.5, 5.25) and C3=(1.5, 3.5).
After the 3rd epoch, the results would be:
1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7}
with centers C1=(3.66, 9), C2=(7, 4.33) and C3=(1.5, 3.5).



https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxzdWpveXNhaGFzaXRlfGd4OjJhMjI5ZTU1ZTgxMDRhYjk