



Computing Science at University of Stirling

T6 DM Tutorial on “Clustering & Association Rules” – Answers

1. K-Means clustering

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 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).

The distance matrix based on the Euclidean distance is given below:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{72}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{4}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Suppose that the initial seeds (centres of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch only. At the end of this epoch show:

- The new clusters (i.e. the examples belonging to each cluster)
- The centres of the new clusters
- Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.
- How many more iterations are needed to converge? Draw the result for each epoch.

Solution:

a) $d(a,b)$ denotes the Euclidean distance between a and b. It is obtained directly from the distance matrix or calculated as follows: $d(a,b) = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2}$

seed1=A1=(2,10), seed2=A4=(5,8), seed3=A7=(1,2)

epoch1 – start:

A1:

$d(A1, \text{seed1}) = 0$ as A1 is seed1

$d(A1, \text{seed2}) = \sqrt{13} > 0$

$d(A1, \text{seed3}) = \sqrt{65} > 0$

→ A1 ∈ cluster1

A2:

$d(A2, \text{seed1}) = \sqrt{25} = 5$

$d(A2, \text{seed2}) = \sqrt{18} = 4.24$

$d(A2, \text{seed3}) = \sqrt{10} = 3.16 \leftarrow \text{smaller}$

→ A2 ∈ cluster3

A3:

$$d(A3, \text{seed1}) = \sqrt{72} = 8.49$$

$$d(A3, \text{seed2}) = \sqrt{25} = 5 \leftarrow \text{smaller}$$

$$d(A3, \text{seed3}) = \sqrt{53} = 7.28$$

□ A3 ∈ cluster2

A4:

$$d(A4, \text{seed1}) = \sqrt{13}$$

$$d(A4, \text{seed2}) = 0 \text{ as A4 is seed2}$$

$$d(A4, \text{seed3}) = \sqrt{52} > 0$$

→ A4 ∈ cluster2

A5:

$$d(A5, \text{seed1}) = \sqrt{50} = 7.07$$

$$d(A5, \text{seed2}) = \sqrt{13} = 3.60 \leftarrow \text{smaller}$$

$$d(A5, \text{seed3}) = \sqrt{45} = 6.70$$

→ A5 ∈ cluster2

A6:

$$d(A6, \text{seed1}) = \sqrt{52} = 7.21$$

$$d(A6, \text{seed2}) = \sqrt{17} = 4.12 \leftarrow \text{smaller}$$

$$d(A6, \text{seed3}) = \sqrt{29} = 5.38$$

→ A6 ∈ cluster2

A7:

$$d(A7, \text{seed1}) = \sqrt{65} > 0$$

$$d(A7, \text{seed2}) = \sqrt{52} > 0$$

$$d(A7, \text{seed3}) = 0 \text{ as A7 is seed3}$$

→ A7 ∈ cluster3

A8:

$$d(A8, \text{seed1}) = \sqrt{5}$$

$$d(A8, \text{seed2}) = \sqrt{2} \leftarrow \text{smaller}$$

$$d(A8, \text{seed3}) = \sqrt{58}$$

→ A8 ∈ cluster2

end of epoch1

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

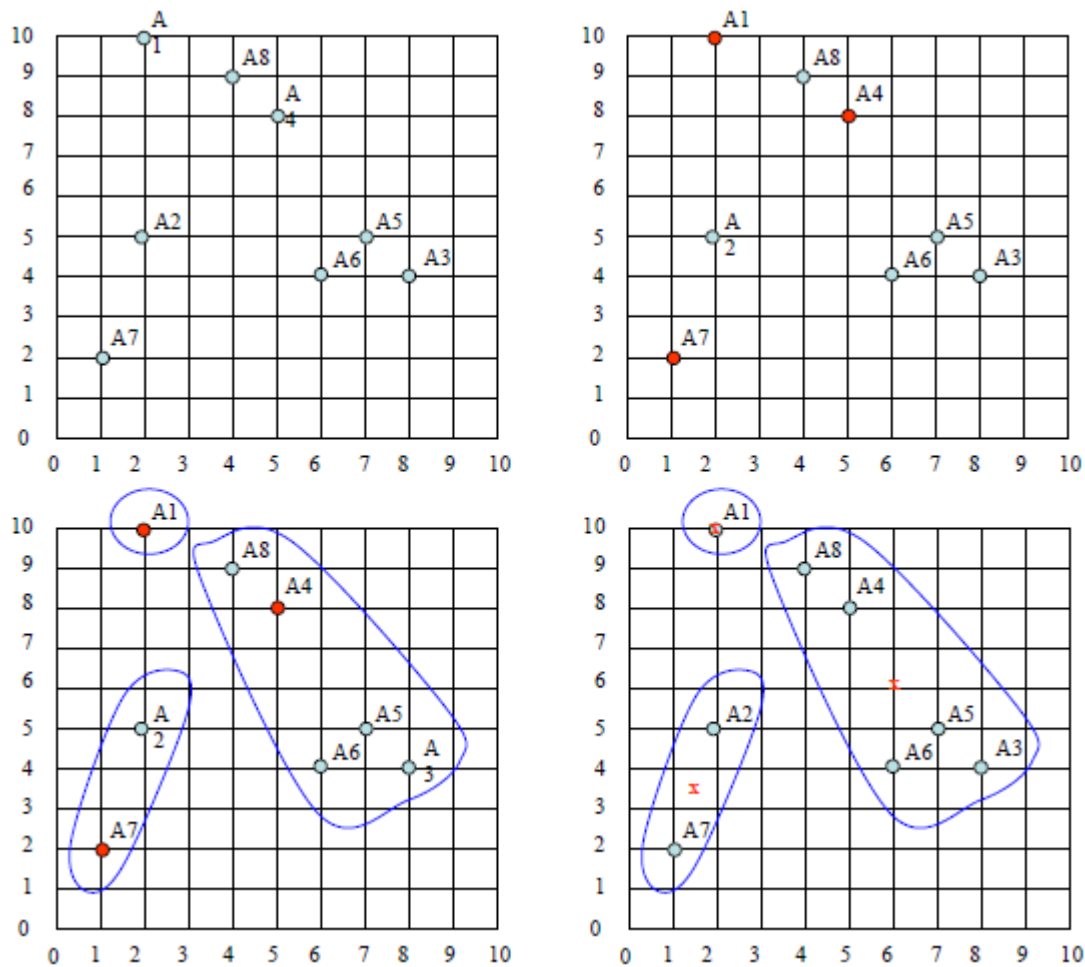
b) Centres of the new clusters:

$$C1 = (2, 10)$$

$$C2 = ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)$$

$$C3 = ((2+1)/2, (5+2)/2) = (1.5, 3.5)$$

c)



d) We would need two more epochs. After the 2nd epoch the results would be:

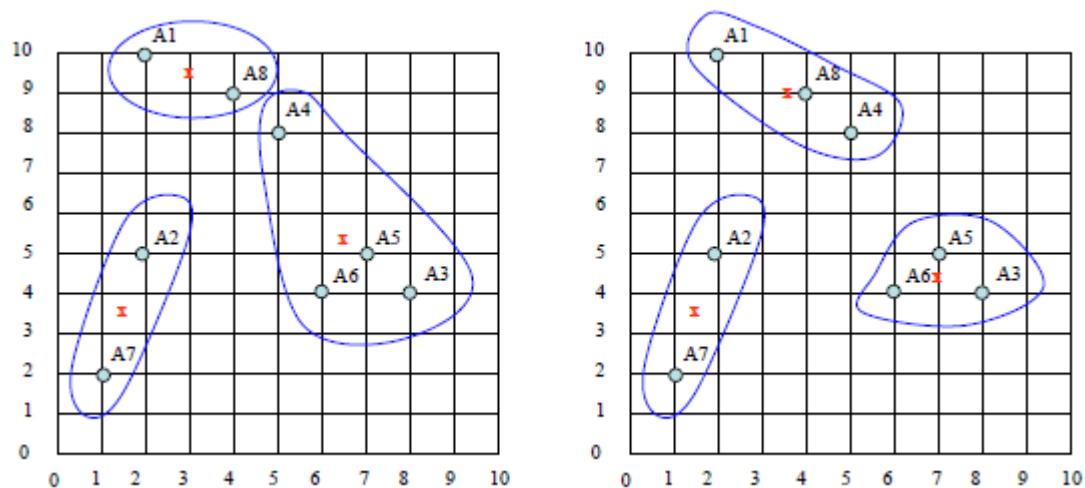
1: {A1, A8}, 2: {A3, A4, A5, A6}, 3: {A2, A7}

with centres $C1=(3, 9.5)$, $C2=(6.5, 5.25)$ 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 centres $C1=(3.66, 9)$, $C2=(7, 4.33)$ and $C3=(1.5, 3.5)$.



2. Association Rules

Find all association rules in the following database:

(minimum support = 40%, minimum confidence = 70%)

TransID	Items
1	a, b, c
2	b, c, d, e
3	c, d
4	a, b, d
5	a, b, c

Solution:

- First find all itemsets with support ≥ 2

C_1

set	supp
{a}	3
{b}	4
{c}	4
{d}	3
{e}	1

C_2

set	supp
{a,b}	3
{a,c}	2
{a,d}	1
{b,c}	3
{b,d}	2
{c,d}	2

C_3

set	supp
{a,b,c}	2
{b,c,d}	1

TID	Items
1	a, b, c
2	b, c, d, e
3	c, d
4	a, b, d
5	a, b, c

(minsup = 2, minconf = 70%)

- $F = \{ \{a\}, \{b\}, \{c\}, \{d\}, \{a,b\}, \{a,c\}, \{b,c\}, \{b,d\}, \{c,d\}, \{a,b,c\} \}$

- Then split all frequent itemsets in all possible ways

(Trivial rules: $X \Rightarrow \{ \}$ 100%)

ac \Rightarrow b	100%
a \Rightarrow b	100%
b \Rightarrow a	75%
b \Rightarrow c	75%
c \Rightarrow b	75%
$\{ \} \Rightarrow$ b	80%
$\{ \} \Rightarrow$ c	80%

set	supp
$\{ \}$	5
{a}	3
{b}	4
{c}	4
{d}	3
{a,b}	3
{a,c}	2
{b,c}	3
{b,d}	2
{c,d}	2
{a,b,c}	2