Actividades resueltas

k-means: Algoritmo de *clustering* exclusivo

Descripción de la actividad

Una universidad online tiene la siguiente base de datos con siete registros correspondientes a las notas de los alumnos en dos asignaturas de un máster.

	Asignatura 1	Asignatura 2
A1	1.0	1.0
A2	1.5	2.0
A3	3.0	4.0
A4	5.5	7.0
A5	3.5	5.0
A6	4.5	5.0
A7	3.5	4.5

A partir de esta base de datos se quiere agrupar a los alumnos en **dos** agrupaciones exclusivas empleando el algoritmo **k-means** y la medida de distancia **Euclídea**.

Se sabe que en el primer paso del algoritmo k-means se han escogido como centroides de los dos clústeres a los elementos A1 y A4. Por tanto, el valor inicial del centroide del Clúster 1 es (1.0, 1.0) y el valor inicial del centroide del Clúster 2 es (5.5, 7.0).

Aplica el algoritmo k-means para encontrar los elementos que componen cada una de las agrupaciones de los alumnos (Clúster 1 y Clúster 2). Describe claramente los pasos que se realizan en la ejecución del algoritmo k-means.

Resolución de la actividad

Una vez en el paso 1 se han escogido los centroides de forma aleatoria, en el paso 2 se procede a asignar cada uno de los elementos al clúster más similar. Para ello se calcula

la distancia Euclídea entre el elemento A_i y los dos centroides y se asigna el elemento al clúster más cercano.

	Punto	Distancia al Centroide 1 (1.0, 1.0)	Distancia al Centroide 2 (5.5, 7.0)	Asignación
A1	(1.0, 1.0)	0	-	Clúster 1
A2	(1.5, 2.0)	$\sqrt{(1.5-1.0)^2+(2.0-1.0)^2} = \sqrt{0.5^2+1.0^2} = \sqrt{\mathbf{1.25}}$	$\sqrt{(1.5-5.5)^2+(2.0-7.0)^2} = \sqrt{(-4.0)^2+(-5.0)^2} = \sqrt{41}$	Clúster 1
A3	(3.0, 4.0)	$\sqrt{2^2+3^2} = \sqrt{\mathbf{13}}$	$\sqrt{2.5^2+3^2} = \sqrt{15.25}$	Clúster 1
A4	(5.5, 7.0)	-	0	Clúster 2
A5	(3.5, 5.0)	$\sqrt{2.5^2+4^2} = \sqrt{22.25}$	$\sqrt{2^2+2^2} = \sqrt{\mathbf{8}}$	Clúster 2
A6	(4.5, 5.0)	$\sqrt{3.5^2+4^2} = \sqrt{28.25}$	$\sqrt{1^2+2^2} = \sqrt{\mathbf{5}}$	Clúster 2
A7	(3.5, 4.5)	$\sqrt{2.5^2+3.5^2} = \sqrt{18.5}$	$\sqrt{2^2+2.5^2} = \sqrt{\mathbf{10.25}}$	Clúster 2

Nota: No es necesario calcular las raíces cuadradas para ver que distancia es menor. Tampoco calculamos la distancia para A1 y A4 porque estos elementos son los que se ha cogido como centroides.

En el paso 3 del algoritmo se recalculan las posiciones de los centroides para la asignación actual de los clústeres.

El Clúster 1 se compone de los elementos A1, A2 y A3, entonces su centroide será el valor medio de estos puntos que conforman el clúster.

	Punto
A1	(1.0, 1.0)
A2	(1.5, 2.0)
A3	(3.0, 4.0)

$$coordenada\ Asignatura\ 1 = \frac{1.0+1.5+3.0}{3} = 1.833$$

$$coordenada\ Asignatura\ 2 = \frac{1.0+2.0+4.0}{3} = 2.333$$

Por tanto el centroide del Clúster 1 será el punto (1.833, 2.333).

De la misma forma se calcula el centroide para el Clúster 2.

	Punto
A4	(5.5, 7.0)
A5	(3.5, 5.0)
A6	(4.5, 5.0)
A7	(3.5, 4.5)

$$coordenada\ Asignatura\ 1 = \frac{5.5+3.5+4.5+3.5}{4} = 4.25$$

$$coordenada\ Asignatura\ 2 = \frac{7.0+5.0+5.0+4.5}{4} = 5.375$$

Por tanto el centroide del Clúster 2 será el punto (4.25, 5.375).

Finaliza la primera iteración y se repiten los pasos 2 y 3 hasta que las posiciones de los centroides no varíen.

Empezamos con la segunda iteración:

En el paso 2 se procede a volver a asignar cada uno de los elementos al clúster más similar con los nuevos centroides: centroide del Clúster 1 (1.833, 2.333) y centroide del Clúster 2 (4.25, 5.375).

	Punto	Distancia al Centroide 1 (1.8, 2.3)	Distancia al Centroide 2 (4.3, 5.4)	Asignación
A1	(1.0, 1.0)	$\sqrt{(1.0-1.8)^2+(1.0-2.3)^2} = \sqrt{0.8^2+1.3^2} = \sqrt{\mathbf{2.33}}$	$\sqrt{(1.0-4.3)^2+(1.0-5.4)^2} = \sqrt{3.3^2+4.4^2} = \sqrt{30.25}$	Clúster 1
A2	(1.5, 2.0)	$\sqrt{0.3^2+0.3^2} = \sqrt{\mathbf{0.18}}$	$\sqrt{2.8^2+3.4^2} = \sqrt{19.4}$	Clúster 1
A3	(3.0, 4.0)	$\sqrt{1.2^2+1.7^2} = \sqrt{4.33}$	$\sqrt{1.3^2+1.4^2} = \sqrt{\mathbf{3.65}}$	**Clúster 2**
A4	(5.5, 7.0)	$\sqrt{3.7^2+4.7^2} = \sqrt{35.78}$	$\sqrt{1.2^2+1.6^2} = \sqrt{\mathbf{4}}$	Clúster 2
A5	(3.5, 5.0)	$\sqrt{1.7^2+2.7^2} = \sqrt{10.18}$	$\sqrt{0.8^2+0.4^2} = \sqrt{\mathbf{0.8}}$	Clúster 2
A6	(4.5, 5.0)	$\sqrt{2.7^2+2.7^2} = \sqrt{14.58}$	$\sqrt{0.2^2+0.4^2} = \sqrt{\mathbf{0.2}}$	Clúster 2
A7	(3.5, 4.5)	$\sqrt{1.7^2+2.2^2} = \sqrt{7.73}$	$\sqrt{0.8^2+0.9^2} = \sqrt{\mathbf{1.45}}$	Clúster 2

Ha habido un cambio en la asignación del elemento A3 con respecto a la iteración anterior (antes pertenecía al Clúster 1 y ahora pertenece al Clúster 2). Por lo tanto los centroides van a cambiar de posición y debemos continuar con la aplicación del algoritmo.

En el paso 3 del algoritmo se recalculan las posiciones de los centroides para la asignación actual de los clústeres.

Se calcula el centroide para el Clúster 1.

	Punto
A1	(1.0, 1.0)
A2	(1.5, 2.0)

$$coordenada\ Asignatura\ 1 = \frac{1.0+1.5}{2} = 1.25$$

$$coordenada\ Asignatura\ 2 = \frac{1.0+2.0}{2} = 1.5$$

Por tanto el centroide del Clúster 1 será el punto (1.25, 1.5).

Se calcula el centroide para el Clúster 2.

	Punto
A3	(3.0, 4.0)
A4	(5.5, 7.0)
A5	(3.5, 5.0)
A6	(4.5, 5.0)
A7	(3.5, 4.5)

$$coordenada\ Asignatura\ 1 = \frac{3.0 + 5.5 + 3.5 + 4.5 + 3.5}{5} = 4$$

$$coordenada\ Asignatura\ 2 = \frac{4.0 + \ 7.0 + 5.0 + 5.0 + 4.5}{5} = 5.1$$

Por tanto el centroide del Clúster 2 será el punto (4, 5.1).

Finaliza la segunda iteración y se repiten los pasos 2 y 3 hasta que las posiciones de los centroides no varíen.

Empezamos con la tercera iteración:

En el paso 2 se procede a volver a asignar cada uno de los elementos al clúster más similar con los nuevos centroides: centroide del Clúster 1 (1.25, 1.5) y centroide del Clúster 2 (4, 5.1).

	Punto	Distancia al Centroide 1 (1.25, 1.5)	Distancia al Centroide 2 (4, 5.1)	Asignación
A1	(1.0, 1.0)	$\sqrt{0.25^2 + 0.5^2} = \sqrt{\mathbf{0.31}}$	$\sqrt{3.0^2 + 4.1^2} = \sqrt{25.81}$	Clúster 1
A2	(1.5, 2.0)	$\sqrt{0.25^2 + 0.5^2} = \sqrt{\mathbf{0.31}}$	$\sqrt{2.5^2 + 3.1^2} = \sqrt{15.86}$	Clúster 1
A3	(3.0, 4.0)	$\sqrt{1.75^2 + 2.5^2} = \sqrt{9.31}$	$\sqrt{1.0^2 + 1.1^2} = \sqrt{\mathbf{2.21}}$	Clúster 2
A4	(5.5, 7.0)	$\sqrt{4.25^2 + 5.5^2} = \sqrt{48.31}$	$\sqrt{1.5^2 + 1.9^2} = \sqrt{\mathbf{5.86}}$	Clúster 2
A5	(3.5, 5.0)	$\sqrt{2.25^2 + 3.5^2} = \sqrt{17.31}$	$\sqrt{0.5^2 + 0.1^2} = \sqrt{\mathbf{0.26}}$	Clúster 2
A6	(4.5, 5.0)	$\sqrt{3.25^2 + 3.5^2} = \sqrt{22.81}$	$\sqrt{0.5^2 + 0.1^2} = \sqrt{\mathbf{0.26}}$	Clúster 2
A7	(3.5, 4.5)	$\sqrt{2.25^2 + 3.0^2} = \sqrt{14.06}$	$\sqrt{0.5^2 + 0.6^2} = \sqrt{\mathbf{0.61}}$	Clúster 2

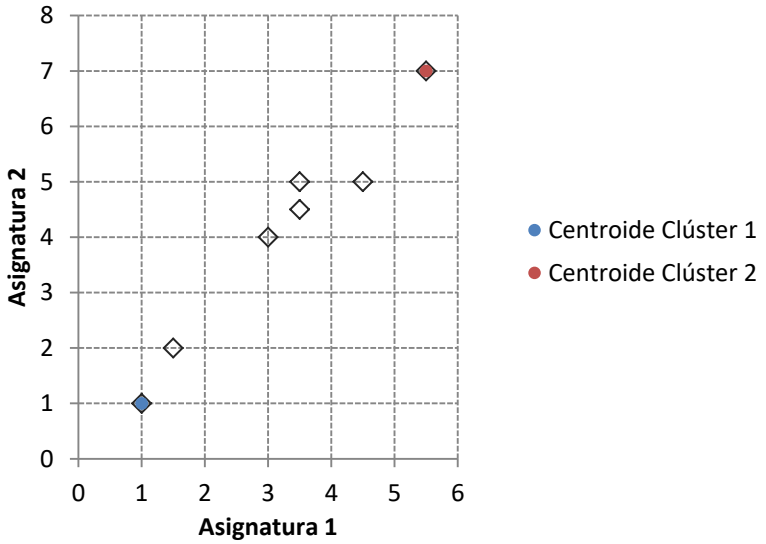
No ha habido ningún cambio en la asignación de los elementos con respecto a la iteración, por lo tanto los centroides no van a cambiar de posición. Entonces podemos concluir con la aplicación del algoritmo k-means.

En resumen, el Clúster 1 contiene a los alumnos A1 y A2 y el Clúster 2 a los alumnos A3, A4, A5, A6 y A7. Para cada uno de los clústeres los alumnos tienen una serie de características en común que les caracterizan. Los alumnos del Clúster 1 sacan como media un 1.25 en la Asignatura 1 y un 1.5 en la Asignatura 2, mientras que los alumnos del Clúster 1 sacan como media un 4 en la Asignatura 1 y un 5.1 en la Asignatura 2.

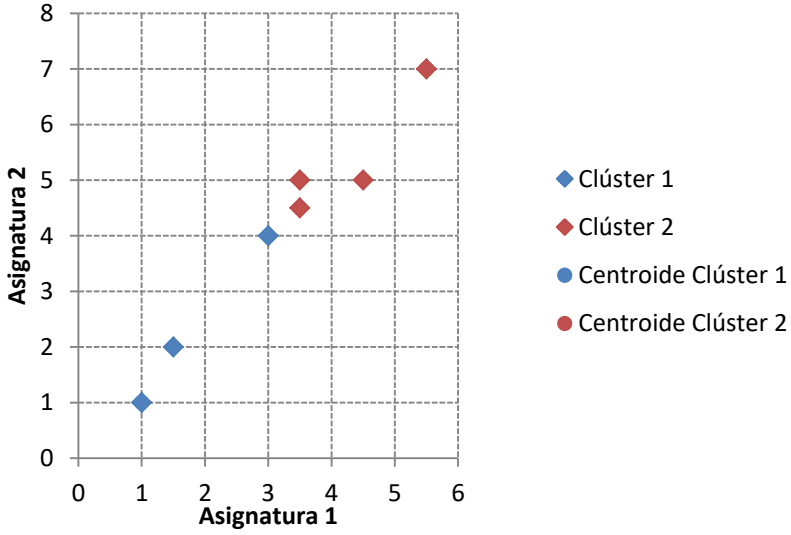
Anexo: Representación gráfica

Aunque no se pedía en el ejercicio, en los siguientes gráficos se representa la asignación de los elementos a cada clúster y los centroides para cada paso de cada una de las iteraciones del algoritmo.

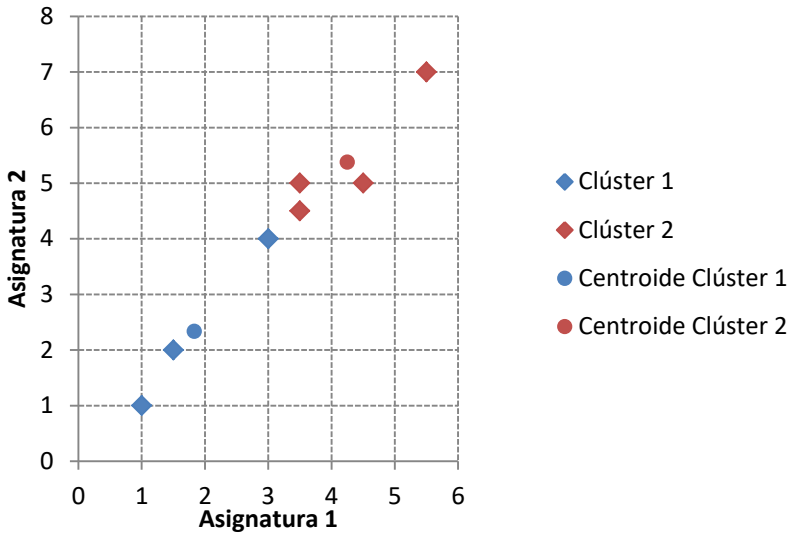
Paso 1: Seleccionar las posiciones iniciales de los centroides (1.0, 1.0) y (5.5, 7.0).



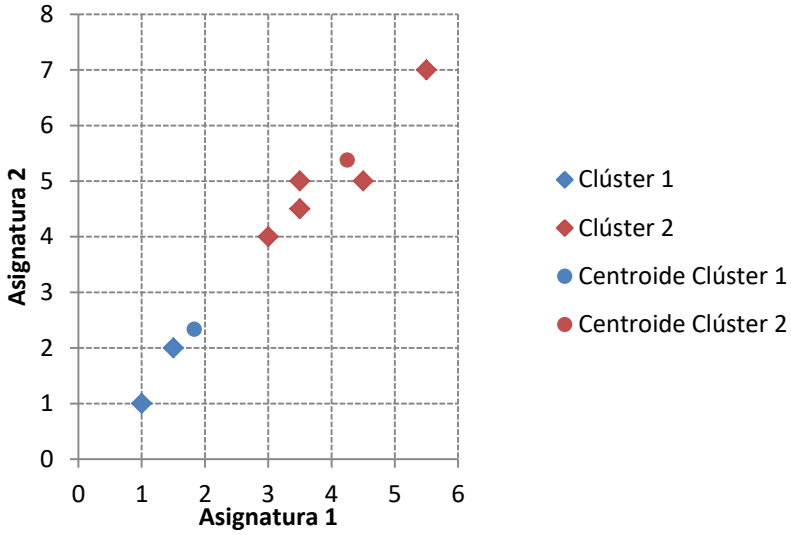
Primera Iteración. Paso 2: Asignar los elementos a los centroides



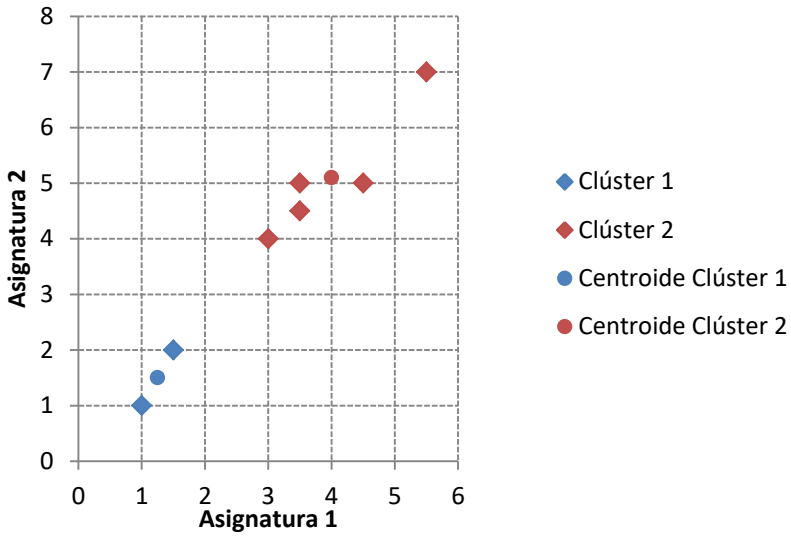
Primera Iteración. Paso 3: Recalcular los centroides



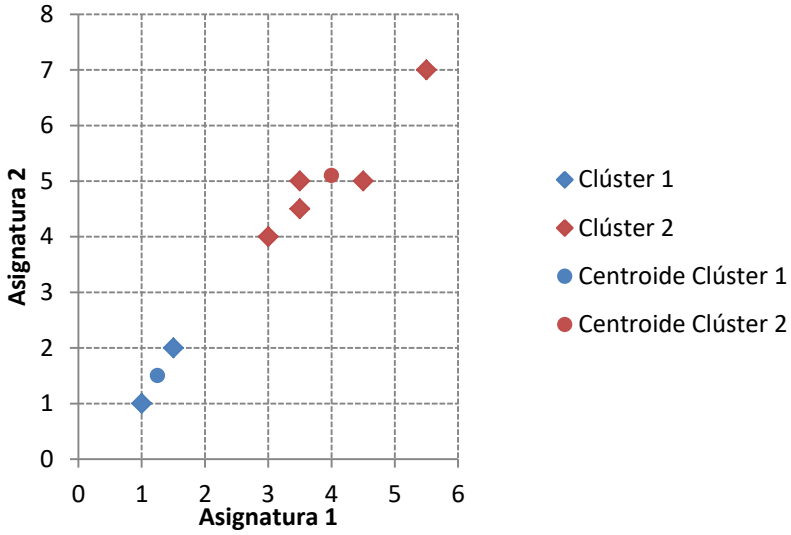
Segunda Iteración. Paso 2: Asignar los elementos a los centroides (1.8, 2.3) y (4.3, 5.4)



Segunda Iteración. Paso 3: Recalcular los centroides



Tercera Iteración. Paso 2: Asignar los elementos a los centroides (1.25, 1.5) y (4, 5.1)



No hay cambios en la asignación de los elementos, por tanto finaliza el algoritmo.