Sistemas Inteligentes
Cuestiones y ejercicios del bloque 2, tema 4
Aprendizaje no supervisado: algoritmo k-medias

Escola Tècnica Superior d'Informàtica
Dep. de Sistemes Informàtics i Computació
Universitat Politècnica de València

10 de noviembre de 2014

1. Cuestiones

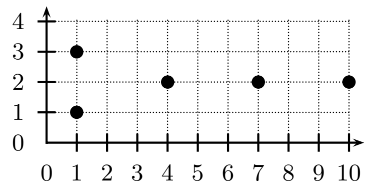
1 $\boxed{\text{C}}$ Durante la ejecución del algoritmo *c-means* se obtiene la siguiente partición en dos grupos $X_1 = \{(0,0),(1,0),(2,1)\}$ y $X_2 = \{(0,1),(1,2),(2,2)\}$. Calcula la Suma de Errores Cuadráticos (SEC) de dicha partición.

A) 8/3
B) 4/3
C) 16/3
D) 5/3

2 $\boxed{\text{B}}$ Indica cual de las siguientes afirmaciones con respecto a la Suma de Errores Cuadráticos (SEC) es la correcta:

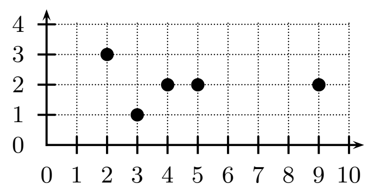
A) La versión de Duda-Hart del *c-means* garantiza un mínimo global de la SEC.
B) No existe ningún algoritmo de coste polinómico que garantice un mínimo global de la SEC.
C) La versión de Duda-Hart del *c-means* garantiza una SEC nula.
D) La versión "popular" del *c-means* garantiza un mínimo local de la SEC.

3 $\boxed{\text{B}}$ La menor suma de errores cuadráticos con la que los puntos de la figura a la derecha pueden agruparse en dos clústers es:



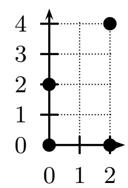
A) Menor que 10
B) Entre 10 y 15
C) Entre 15 y 20
D) Mayor que 20

4 $\boxed{\text{B}}$ La menor suma de errores cuadráticos con la que los puntos de la figura a la derecha pueden agruparse en dos clústers es:



A) Menor que 5.
B) Mayor que 5 y menor que 10.
C) Mayor que 10 y menor que 15.
D) Mayor que 15.

5 $\boxed{\text{A}}$ Los puntos de la figura a la derecha están siendo agrupados mediante el algoritmo C-Medias y, tras cierta iteración del algoritmo, se tiene la partición $\Pi = \{X_1 = \{(0,0),(0,2)\}, X_2 = \{(2,0),(2,4)\}\}$, medias $\mathbf{m}_1 = (0,1)$ y $\mathbf{m}_2 = (2,2)$, y SEC (suma de errores cuadráticos) $J = 10$. Si el punto $(2,0)$ se cambia de grupo, entonces:



A) La nueva SEC será menor que 6.
B) La nueva SEC estará entre 6 y 10.
C) La nueva SEC será mayor que 10.
D) No conviene cambiar ese punto de grupo porque los grupos se quedarían con tallas desequilibradas.

6 $\boxed{\text{A}}$ Sean dos clases, A y B, de las que se dispone de los siguientes prototipos: $A = \{(0,2),(1,1),(1,3),(2,2)\}$; y $B = \{(3,2),(3,3),(4,2),(4,3)\}$. Supóngase estos dos conjuntos de prototipos constituyen dos clústers que resultan de un proceso de agrupamiento no supervisado. La SEC, J, correspondiente a dicho agrupamiento es:

A) $J \leq 6$
B) $6 < J \leq 8$
C) $8 < J \leq 10$
D) $J > 10$

7 $\boxed{\text{D}}$ La diferencia principal entre el aprendizaje supervisado (AS) y no-supervisado (ANS) es que:

A) en el AS se conocen las clases correctas de los datos de test, mientras que en el ANS solo se conocen las de entrenamiento.
B) en el AS siempre hay un operador humano que supervisa los resultados de forma que el sistema solo sirve de ayuda o asistencia, mientras que en el ANS todo se realiza de forma totalmente automática.
C) el ANS es un proceso iterativo mientras que el AS se realiza en un paso.
D) en el AS se conocen las etiquetas de clase correctas de todos los datos de aprendizaje, mientras que en el ANS no se conocen.

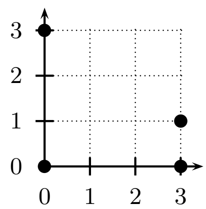
8 $\boxed{\text{B}}$ El algoritmo C-medias es una técnica de *clustering* particional que aplicamos en reconocimiento del habla para...

A) Transformar la señal al dominio temporal-frecuencial.
B) Diseñar *codebooks*.
C) Entrenar modelos de Markov.
D) Ninguna de las anteriores.

9 $\boxed{\text{B}}$ Sean dos clases, A y B, de las que se dispone de los siguientes prototipos: $A = \{(2,1),(1,2),(2,3),(3,2)\}$ y $B = \{(4,3),(5,3),(3,5),(6,5)\}$. Supóngase estos dos conjuntos de prototipos constituyen dos clústers que resultan de un proceso de agrupamiento particional. La SEC sería:

A) $SEC < 4$
B) $SEC > 12$
C) $SEC = 11$
D) $4 < SEC < 10$

10 $\boxed{\text{C}}$ Los puntos de la figura a la derecha están siendo agrupados mediante el algoritmo C-Medias y, tras cierta iteración del algoritmo, se tiene la partición $\Pi = \{X_1 = \{(0,0),(0,3),(3,0)\}, X_2 = \{(3,1)\}\}$. Sea J' la suma de errores cuadráticos de esta partición y sea J la suma de errores cuadráticos de la partición que se obtiene al cambiar de grupo el punto $(3,0)$. Entonces:

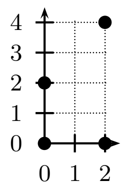


A) $J \geq J'$
B) $\frac{1}{2}J' \leq J < J'$
C) $\frac{1}{4}J' \leq J < \frac{1}{2}J'$
D) $J < \frac{1}{4}J'$

11 $\boxed{\text{C}}$ Cuál de la siguientes afirmaciones en relación al aprendizaje no supervisado es falsa:

A) El objetivo del aprendizaje no supervisado es agrupar en grupos “naturales” los datos disponibles
B) Una medida muy empleada para medir la calidad de un agrupamiento particional es la Suma de Errores Cuadráticos (SEC)
C) El algoritmo c-medias garantiza un mínimo global del SEC
D) Se emplea, por ejemplo, en Reconocimiento Automático del habla para representar una señal acústica como una secuencia de símbolos asociados a los “codewords”

12 $\boxed{\text{B}}$ Los puntos de la figura a la derecha están siendo agrupados mediante el algoritmo C-Medias y, tras cierta iteración del algoritmo, se tiene la partición $\Pi = \{X_1 = \{(0,0),(0,2)\}, X_2 = \{(2,0),(2,4)\}\}$, medias $\mathbf{m}_1 = (0,1)$ y $\mathbf{m}_2 = (2,2)$, y SEC (suma de errores cuadráticos) $J = 10$. Si el punto $(2,0)$ se cambia de grupo, entonces:



A) La nueva SEC será menor que 5.
B) La nueva SEC estará entre 5 y 7.
C) La nueva SEC será mayor que 7 pero menor que 10
D) Ese punto no se puede cambiar porque deja uno de los grupos con sólo un punto.

13 $\boxed{\text{C}}$ Sea $X = \{1, 3, 4.5\}$ un conjunto de 3 datos unidimensionales a agrupar en dos clústers mediante alguna técnica de clustering particional. Más concretamente, se desea optimizar el criterio SEC (suma de errores cuadráticos) y emplear el algoritmo C-medias, si bien no se ha decidido si emplearemos la versión *popular* o la versión de *Duda y Hart (DH)*. Sea $\Pi^0 = \{X_1 = \{1, 3\}, X_2 = 4.5\}$ una partición inicial en dos clústers y SEC $J^0 = 2$. Indica cuál de las siguientes afirmaciones es cierta:

A) Tanto la versión popular como la DH terminarán sin modificar la partición inicial.
B) La versión popular terminará con una partición mejorada, pero no la versión DH, que terminará sin modificar la partición inicial.
C) La versión DH terminará con una partición mejorada, pero no la versión popular, que terminará sin modificar la partición inicial.
D) Ambas versiones terminarán con particiones mejoradas respecto a la partición inicial.

14 $\boxed{\text{A}}$ (Examen de SIN del 18 de enero de 2013)
El criterio de clustering particional "Suma de Errores Cuadráticos" es apropiado cuando los datos forman clústers:

A) Hiperesféricos y de tamaño similar.
B) Hiperesféricos y de cualquier tamaño.
C) Alargados y de tamaño similar.
D) Alargados y de cualquier tamaño.

15 $\boxed{\text{C}}$ (Examen de SIN del 30 de enero de 2013)
Se tienen 3 datos unidimensionales: $x_1 = 0$, $x_2 = 20$ y $x_3 = 35$. Se ha establecido una partición de estos datos en dos clústers, $\Pi = \{X_1 = \{x_1, x_2\}, X_2 = \{x_3\}\}$. La Suma de Errores Cuadráticos (SEC) de esta partición es:

A) $J(\Pi) = 0$
B) $0 < J(\Pi) \leq 150$
C) $150 < J(\Pi) \leq 300$ $\quad J(\Pi) = (x_1 - m_1)^2 + (x_2 - m_1)^2 + (x_3 - m_2)^2 = (0 - 10)^2 + (20 - 10)^2 + (35 - 35)^2 = 200$
D) $J(\Pi) > 300$

16 $\boxed{\text{B}}$ (Examen de SIN del 30 de enero de 2013)
Tras aplicar la versión "correcta" ("Duda y Hart") del algoritmo C-medias a la partición de la cuestión anterior (Π), la partición resultante (Π^*) es: $\quad \Delta J = \frac{n_2}{n_2+1}|x_2 - m_2|^2 - \frac{n_1}{n_1-1}|x_2 - m_1|^2$

A) $\Pi^* = \Pi$. $\quad \Delta J = 0$
B) $\Pi^* = \{X_1 = \{x_1\}, X_2 = \{x_2, x_3\}\}$. $\quad \Delta J = \frac{1}{2}|20 - 35|^2 - \frac{2}{1}|20 - 10|^2 = 112.5 - 200 = -87.5$
C) $\Pi^* = \{X_1 = \{x_2\}, X_2 = \{x_1, x_3\}\}$. $\quad \Delta J = \frac{1}{2}|0 - 35|^2 - \frac{2}{1}|0 - 10|^2 = 612.5 - 200 = 412.5$
D) Ninguna de las anteriores.

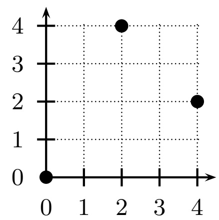
17 $\boxed{\text{D}}$ (Examen de SIN del 15 de enero de 2014; examen del bloque 2; cuestión 11)
Indica cuál de la siguientes afirmaciones sobre *Clustering* es correcta:

A) Se suele emplear el algoritmo *Perceptrón* a partir de datos de entrenamiento *con* etiquetas de clase.
B) Se suele emplear el algoritmo *Perceptrón* a partir de datos de entrenamiento *sin* etiquetas de clase.
C) Se suele emplear el algoritmo *C-Medias* a partir de datos de entrenamiento *con* etiquetas de clase.
D) Se suele emplear el algoritmo *C-Medias* a partir de datos de entrenamiento *sin* etiquetas de clase.

18 $\boxed{\text{D}}$ (Examen de SIN del 15 de enero de 2014; examen del bloque 2; cuestión 12)
El criterio de clustering particional "Suma de Errores Cuadráticos" es apropiado cuando los datos forman clústers:

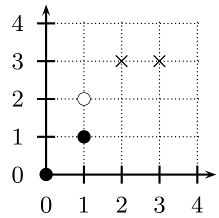
A) No alargados.
B) Alargados y de cualquier tamaño.
C) Alargados y de tamaño similar.
D) Ninguna de las anteriores.

19 $\boxed{\text{B}}$ (Examen de SIN del 15 de enero de 2014; examen del bloque 2; cuestión 13)
La menor suma de errores cuadráticos con la que pueden agruparse en dos clústers los puntos a la derecha es un valor:



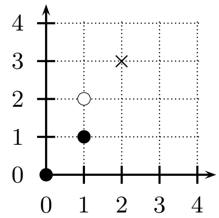
A) Entre 0 y 3.
B) Entre 3 y 6. $\quad J = 4$
C) Entre 6 y 9.
D) Mayor que 9.

20 $\boxed{\text{B}}$ (Examen de SIN del 15 de enero de 2014; examen del bloque 2; cuestión 14)
La figura a la derecha muestra una partición de 5 puntos bidimensionales en 3 clústers (representados mediante los símbolos $\bullet$, $\circ$ y $\times$). Considera todas las posibles transferencias de clúster de cada punto (en un clúster no unitario). En términos de suma de errores cuadráticos (J):



A) Ninguna transferencia permite mejorar J.
B) Sólo se puede mejorar J transfiriendo $(1,1)^t$ del clúster $\bullet$ al $\circ$.
C) Sólo se puede mejorar J transfiriendo $(2,3)^t$ del clúster $\times$ al $\circ$.
D) Las dos transferencias anteriores permiten mejorar J.

21 $\boxed{\text{C}}$ (Examen de SIN del 28 de enero de 2014; examen final; cuestión 5)
La figura a la derecha muestra una partición de 4 puntos bidimensionales en 3 clústers (representados mediante los símbolos $\bullet$, $\circ$ y $\times$). La suma de errores cuadráticos de esta partición es $J = 1$. Si se ejecuta el algoritmo C-medias (de Duda y Hart) a partir de la misma:



A) No se realizará niguna transferencia de clúster.
B) Se transferirá un único punto, obteniéndose una partición de J entre $\frac{2}{3}$ y 1.
C) Se transferirá un único punto, obteniéndose una partición de J entre 0 y $\frac{2}{3}$. J=0.5
D) Se realizarán dos transferencias de clúster, obteniéndose una partición de J nula.

2. Problemas

1. Se tienen los siguientes 5 vectores bidimensionales:

$$\mathbf{x}_1 = \begin{pmatrix}1\\7\end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix}4\\2\end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix}4\\6\end{pmatrix}, \quad \mathbf{x}_4 = \begin{pmatrix}8\\2\end{pmatrix} \quad \text{y} \quad \mathbf{x}_5 = \begin{pmatrix}8\\6\end{pmatrix}$$

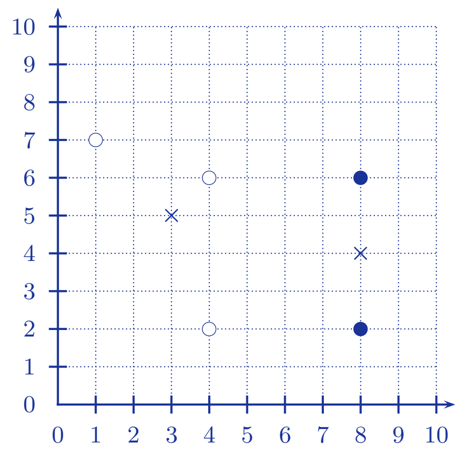
Se desea agrupar estos vectores de manera no-supervisada en 2 clases. Partiendo de la partición

$$\Pi = \{X_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}, X_2 = \{\mathbf{x}_4, \mathbf{x}_5\}\}$$

realiza una traza de ejecución de una iteración del bucle principal del algoritmo *c-medias*.

$$\mathbf{m}_1 = \frac{1}{3}(\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3) = \begin{pmatrix}3\\5\end{pmatrix}$$

$$\mathbf{m}_2 = \frac{1}{2}(\mathbf{x}_4 + \mathbf{x}_5) = \begin{pmatrix}8\\4\end{pmatrix}$$



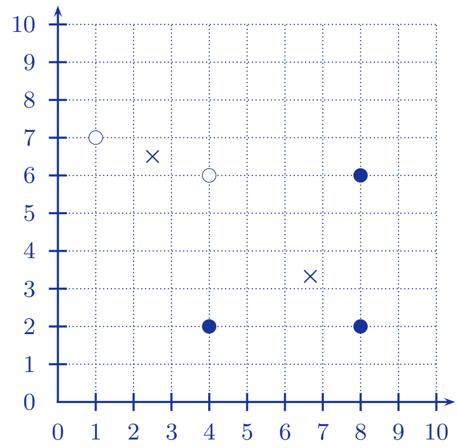
$$J_1 = \|\mathbf{x}_1 - \mathbf{m}_1\|^2 + \|\mathbf{x}_2 - \mathbf{m}_1\|^2 + \|\mathbf{x}_3 - \mathbf{m}_1\|^2 = 8 + 10 + 2 = 20$$
$$J_2 = \|\mathbf{x}_4 - \mathbf{m}_2\|^2 + \|\mathbf{x}_5 - \mathbf{m}_2\|^2 = 4 + 4 = 8$$
$$J = J_1 + J_2 = 28$$

Si transferimos $\mathbf{x}_n \in X_i$ a X_j, entonces $\quad \Delta J = \frac{|X_j|}{|X_j|+1}\|\mathbf{x}_n - \mathbf{m}_j\|^2 - \frac{|X_i|}{|X_i|-1}\|\mathbf{x}_n - \mathbf{m}_i\|^2$

¿Transferimos $\mathbf{x}_1$ de X_1 a X_2? : $\quad \Delta J = \frac{2}{3}\cdot 58 - \frac{3}{2}\cdot 8 = \frac{80}{3} > 0 \;\Rightarrow\;$ NO

¿Transferimos $\mathbf{x}_2$ de X_1 a X_2? : $\quad \Delta J = \frac{2}{3}\cdot 20 - \frac{3}{2}\cdot 10 = -\frac{5}{3} < 0 \;\Rightarrow\;$ SÍ

$$\mathbf{m}_1 = \mathbf{m}_1 - \frac{\mathbf{x}_2 - \mathbf{m}_1}{|X_1|-1} = \begin{pmatrix}5/2\\13/2\end{pmatrix} = \begin{pmatrix}2.5\\6.5\end{pmatrix}$$
$$\mathbf{m}_2 = \mathbf{m}_2 + \frac{\mathbf{x}_2 - \mathbf{m}_2}{|X_2|+1} = \begin{pmatrix}20/3\\10/3\end{pmatrix} = \begin{pmatrix}6.67\\3.33\end{pmatrix}$$
$$J = J + \Delta J = \frac{79}{3} = 26.33$$



$$¿\text{Transferimos } \mathbf{x}_3 \text{ de } X_1 \text{ a } X_2? \;:\quad \Delta J = \frac{3}{4}\cdot\frac{128}{9} - \frac{2}{1}\cdot\frac{10}{4} = \frac{17}{3} = 5.67 > 0 \;\Rightarrow\; \text{NO}$$

$$¿\text{Transferimos } \mathbf{x}_4 \text{ de } X_2 \text{ a } X_1? \;:\quad \Delta J = \frac{2}{3}\cdot\frac{151}{2} - \frac{3}{2}\cdot\frac{32}{9} = \frac{805}{16} = 50.31 > 0 \;\Rightarrow\; \text{NO}$$

$$¿\text{Transferimos } \mathbf{x}_5 \text{ de } X_2 \text{ a } X_1? \;:\quad \Delta J = \frac{2}{3}\cdot\frac{61}{2} - \frac{3}{2}\cdot\frac{80}{9} = 7 > 0 \;\Rightarrow\; \text{NO}$$

(AQUÍ TERMINA LA RESPUESTA AL ENUNCIADO). El algoritmo continúa como sigue:

$$¿\text{Transferimos } \mathbf{x}_1 \text{ de } X_1 \text{ a } X_2? \;:\quad \Delta J = \frac{3}{4}\cdot\frac{410}{9} - \frac{2}{1}\cdot\frac{5}{2} = \frac{175}{6} = 29.17 > 0 \;\Rightarrow\; \text{NO}$$

$$¿\text{Transferimos } \mathbf{x}_2 \text{ de } X_2 \text{ a } X_1? \;:\quad \Delta J = \frac{2}{3}\cdot\frac{45}{2} - \frac{3}{2}\cdot\frac{80}{9} = \frac{5}{3} = 1.67 > 0 \;\Rightarrow\; \text{NO}$$

No es necesario seguir ya que no se realizará ninguna transferencia. La partición optimizada es:

$$\Pi = \{X_1 = \{\mathbf{x}_1, \mathbf{x}_3\}, X_2 = \{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5\}\}$$