Grado de Ingeniero en Informática
Introducción a la minería de datos
Febrero 2014

---

1. **(2.5)** La siguiente tabla representa la matriz de similaridad o distancias entre los elementos de un cierto conjunto de datos compuesto por 6 puntos:

	p1	**p2**	**p3**	**p4**	**p5**	**p6**
p1	1.00	0.12	0.44	0.98	0.17	0.45
p2	0.12	1.00	0.77	0.88	0.10	0.01
p3	0.44	0.77	1.00	0.67	0.71	0.17
p4	0.98	0.88	0.67	1.00	0.23	0.82
p5	0.17	0.10	0.71	0.23	1.00	0.55
p6	0.45	0.01	0.17	0.82	0.55	1.00

   Realiza el dendograma correspondiente al clustering jerárquico mediante el método de *complete link*. ¿Es posible realizar con la información dada el clustering usando el método *average link*?.

2. **(2.5)** Considera la siguiente tabla que incluye la información de 10 transacciones realizadas en un cierto establecimiento:

Transacción	**Items**
1	{a, b, c, d, e}
2	{d, e}
3	{a, c, d, e}
4	{a, e}
5	{b, c, e}
6	{b, c, d, e}
7	{b, d, e}
8	{a, b, c}
9	{a, c, d, e}
10	{d, e}

   Considera una soporte mínimo del 30%. Construye la rejilla correspondiente a todos los posibles conjuntos de ítems. Marca en la rejilla cada nodo con una F si es frecuente, una I si es infrecuente y una N si es podado por el algoritmo *Apriori*. Adicionalmente marca los nodos maximalmente frecuentes con una M y los nodos cerrados y frecuentes con una C.

   A partir únicamente de los nodos cerrados y frecuentes obtén las 5 reglas con mayor confianza de todas las posibles, sin considerar las reglas triviales.

3. **(2.5)** Considera el conjunto de datos dónde cada instancia tiene 3 atributos, dos de ellos de tipo lógico y un tercero nominal y puede pertenecer a una de dos clases, c_1 ó c_2. La siguiente tabla indica el número de instancias para cada una de las dos clases en función de los valores posibles de los atributos:

Atributos			**Número de instancias**	
x_1	x_2	x_3	c_1	c_2
V	V	a	5	32
V	V	b	0	1
V	V	c	40	7
V	F	a	0	0
V	F	b	10	5
V	F	c	12	6
F	V	a	0	28
F	V	b	0	0
F	V	c	10	3
F	F	a	8	1
F	F	b	1	22
F	F	c	4	17

Construye un árbol de decisión binario utilizando una estrategia voraz y cómo criterio de división de cada nodo el error de clasificación. Para crear una nueva división es necesario que el nuevo subárbol mejore al nodo padre. Calcula el error de entrenamiento del árbol completo.

4. **(2.5)** La siguiente tabla muestra un conjunto de datos de 12 instancias representadas cada una de ellas por 5 variables de tipo lógico:

	x_1	x_2	x_3	x_4	x_5
p1	V	F	F	V	F
p2	V	F	F	F	F
p3	F	V	V	V	V
p4	F	V	V	F	V
p5	V	F	V	V	V
p6	F	F	F	F	V
p7	V	F	F	V	F
p8	F	V	F	F	F
p9	V	F	V	V	V
p10	V	F	V	F	F
p11	V	F	V	V	V
p12	F	V	V	F	V

Realiza el algoritmo k-medias paso a paso usando la distancia de Hamming y $k = 2$. Selecciona como centros iniciales las instancias p1 y p6. El centroide de cada clúster se construye usando la moda de cada variable que forma el clúster. El algoritmo ha de detenerse si no ha convergido después de 10 iteraciones.

Tiempo de realización: **6** horas. Calificación de cada ejercicio entre paréntesis.