Grado de Ingeniero en Informática
Introducción a la minería de datos
Enero 2016, Segundo llamamiento

Cuestiones

Responde brevemente a las siguientes cuestiones justificando las respuestas:

1. **(2)** ¿Qué diferencias hay entre los tipos típicos de clústers generados por los métodos jerárquicos de *single link* y *complete link*?

2. **(2)** ¿En qué se basa el algoritmo *a priori* para la generación de conjuntos frecuentes de ítems en análisis de reglas de asociación para poder reducir el número de posibles conjuntos de ítems a visitar?

3. **(2)** ¿Qué efectos puede tener la existencia de ruido o de outliers en un método de boosting de construcción de agrupaciones de clasificadores?

4. **(2)** Disponemos de tres métodos de clasificación para resolver una serie de problemas. Estos métodos son un árbol de decisión, un máquina de vectores soporte (SVM) y un clasificador por vecino más cercano (1-NN). Tenemos los siguientes tres problemas:

   (a) Un problema con un número moderado de variables pero cientos de miles de instancias.
   (b) Un problema con un número moderado de instancias pero miles de variables.
   (c) Un problema con patrones que contienen mucho ruido.

   Indica cuál de los tres métodos anteriores sería el más adecuado para cada problema.

5. **(2)** Indica una ventaja del clustering jerárquico con respecto al particional.

Problemas

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

	p1	**p2**	**p3**	**p4**	**p5**	**p6**
p1	1.00	0.12	0.34	0.18	0.97	0.45
p2	0.12	1.00	0.73	0.48	0.21	0.11
p3	0.34	0.73	1.00	0.65	0.01	0.77
p4	0.18	0.48	0.65	1.00	0.35	0.80
p5	0.97	0.21	0.01	0.35	1.00	0.65
p6	0.45	0.11	0.77	0.80	0.65	1.00

   Realiza el dendograma correspondiente al clustering jerárquico mediante el método de *single link*.

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

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

Obtén la confianza de las reglas siguientes:

$$a, b \rightarrow c$$
$$b, c \rightarrow a$$
$$a, b, c \rightarrow e$$
$$\emptyset \rightarrow a, b$$
$$d \rightarrow a, e$$

3. **(4)** Considera el conjunto de datos dónde cada instancia tiene 3 atributos, dos de ellos de tipo lógico y un tercero nominal y tres clases, c_1, c_2 y c_3. La siguiente tabla indica el número de instancias para cada una de las tres clases en función de los valores posibles de los atributos:

Atributos			**Número de instancias**		
x_1	x_2	x_3	c_1	c_2	c_3
V	V	a	5	36	7
V	V	b	10	0	44
V	V	c	40	7	1
V	F	a	0	0	16
V	F	b	10	6	8
V	F	c	1	6	21
F	V	a	9	28	0
F	V	b	38	22	0
F	V	c	10	3	4
F	F	a	8	1	18
F	F	b	15	22	6
F	F	c	4	17	0

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.

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