Grado de Ingeniero en Informática
Introducción a la minería de datos
Febrero 2016

Cuestiones

Responde brevemente a las siguientes cuestiones justificando las respuestas:

1. **(2)** ¿Por qué el error de clasificación no es una buena medida en problemas de clasificación de dos clases con un gran desequilibrio de clases?

2. **(2)** ¿En qué consiste el sobreaprendizaje (overfitting) en la construcción de un clasificador? ¿Es posible evitarlo?

3. **(2)** ¿Cómo podría resolver un problema de clasificación de N clases ($N > 2$) si tengo un método de clasificación que solo puede distinguir entre dos clases?

4. **(2)** ¿Es perjudicial la existencia de ruido o de outliers para el método de boosting de construcción de agrupaciones de clasificadores?

5. **(2)** Si se le da como dato de un problema la matriz de similaridad entre todos los elementos de un conjunto de datos, ¿es posible realizar con dicha información el clustering usando los métodos *complete link*, *single link* y *average link*?.

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	–	–	–	–	–
p2	0.12	1.00	–	–	–	–
p3	0.34	0.73	1.00	–	–	–
p4	0.68	0.45	0.65	1.00	–	–
p5	0.27	0.34	0.34	0.35	1.00	–
p6	0.34	0.11	0.11	0.56	0.21	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	{c, d, e}
2	{d, e}
3	{b, c, d}
4	{a, c, e}
5	{c, d, e}
6	{a, c, e}
7	{b, d, e}
8	{c, d}
9	{b, c, d}
10	{a, d}

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:

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

3. **(4)** 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	F	F	F	F	V
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 p2. El centroide de cada clúster se construye usando la moda de cada variable que forma el clúster.

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