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

Cuestiones

Responde brevemente a las siguientes cuestiones justificando las respuestas:

1. **(2)** ¿Qué significado tienen desde el punto de vista intuitivo las medidas de error sensibilidad y especificidad para problemas de clasificación de dos clases?

   RESPUESTA: La sensibilidad tiene como objeto medir el ratio de verdaderos positivos. Mide la capacidad que tiene un clasificador para no errar en la identificación de positivos clasificándolos como negativos. La especificidad hace justo lo contrario, mide la capacidad de no clasificar los datos erróneamente como positivos si son negativos.

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

   RESPUESTA: El sobreaprendizaje ocurre cuando un clasificador aprende muy bien el conjunto de entrenamiento a costa de perder su capacidad de generalización. No existen métodos para evitarlo de forma consistente aunque si hay técnicas para tratar de atenuar su efecto, como el uso de modelos más simples o la detención prematura del entrenamiento mediante validación cruzada.

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

   RESPUESTA: Sí, se puede transformar el problema de N clases en M problemas de dos clases. Métodos conocidos son el one-vs.-one, el one-vs.all o los códigos ECOC.

4. **(2)** Indique cómo llevaría a cabo la comparación de los métodos siguientes de clasificación:

   (a) Comparación de dos métodos sobre un conjunto de N problemas.
   (b) Comparación de un método contra un serie de métodos estándar sobre un conjunto de N problemas para ver si es mejor que todos ellos.

   RESPUESTA: Para el primer caso tendríamos el test de Wilcoxon. Para el segundo caso aplicaríamos primero un test de Friedman o Iman-Davenport para ver si hay diferencias significativas globales. En caso de que sí las haya podemos aplicar el procedimiento de Holm para comparar nuestro método con cada uno de los métodos estándar pasa a paso.

5. **(2)** ¿Qué tipo de clústers tiende a generar un metodo de clustering particional como por ejemplo k-medias?

   REPSUESTA: Genera normalmente clústers homgéneos y de forma globular, es por ello que funciona pobremente si nuestros clústers no corresponden a esta forma.

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 *complete 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	{a, d, e}
3	{b, c, d}
4	{a, c, e}
5	{a, c, e, d}
6	{a, b, c, e}
7	{b, d, e}
8	{c, d}
9	{b, c, d, e}
10	{a, 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:

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

3. **(4)** Considera el conjunto de datos dónde cada instancia tiene 3 atributos, dos de ellos de tipo lógico y un tercero nominal y dos clases, c_1 y c_2. 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
V	V	a	36	7
V	V	b	0	44
V	V	c	7	1
V	F	a	0	16
V	F	b	6	8
V	F	c	6	21
F	V	a	28	0
F	V	b	22	0
F	V	c	3	4
F	F	a	1	18
F	F	b	22	6
F	F	c	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.