Sistemas Inteligentes
Cuestiones y ejercicios del bloque 2, tema 3
Inducción de reglas y patrones: árboles de decisión

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

10 de noviembre de 2014

1. Cuestiones

1 ☐ (Examen de SIN del 18 de Enero de 2013)
Referente a los elementos que intervienen en un algoritmo de aprendizaje de árboles de decisión, indica cuál de estas afirmaciones es falsa:

A) Los *splits* son de la forma ¿$\mathbf{y} \in B$?, $B \subseteq E$.
B) La calidad de un *split* se mide mediante el decremento de impureza total producido por dicho *split*, y la impureza se puede calcular de diversas formas, entre ellas la probabilidad de error o la entropía.
C) En un árbol de decisión, es deseable que el error estimado por resustitución sea nulo.
D) Una buena etiqueta de clase para un nodo terminal $t \in \hat{T}$ es: $c^\star = \arg\max_c \hat{P}(c \mid t)$.

2 ☐ (Examen de SIN del 18 de enero de 2013)
Se tiene un problema de clasificación en $C = 4$ clases para el cual se está construyendo un árbol de clasificación T a partir de un cierto conjunto de datos de aprendizaje. Durante el proceso de construcción, se han obtenido 3 nodos con las siguientes distribuciones de probabilidades estimadas de las clases:

t	$\hat{P}(1 \mid t)$	$\hat{P}(2 \mid t)$	$\hat{P}(3 \mid t)$	$\hat{P}(4 \mid t)$
t_1	2^{-2}	2^{-2}	2^{-2}	2^{-2}
t_2	2^{-1}	2^{-1}	0	0
t_3	2^0	0	0	0

Indica cuál de los nodos es más impuro según el concepto de entropía:

A) t_1.
B) t_2.
C) t_3.
D) No hay un único nodo más impuro que el resto.

3 ☐ (Examen de SIN del 30 de enero de 2013)
En el algoritmo de aprendizaje de un árbol de decisión y clasificación, sea t un nodo con $N(t)$ elementos de $\mathbb{R}^D$, de los que $N_c(t)$ son de la clase c. Identificar cuál de las sigientes afirmaciones es falsa:

A) Una partición óptima de t es aquella para la que la Entropía, $H(t)$, es mínima.
B) t puede considerarse terminal u hoja si el decremento de impureza que se conseguiría con la mejor partición de ese nodo no es suficientemente grande; es decir, si $\quad \text{máx}_{j,r}\ \Delta(j, r, t) < \epsilon \quad$ donde $j \in \{1, 2, \ldots, D\}$ es una dimensión y $r \in \mathbb{R}$ es un umbral de partición unidimensional.
C) Una buena forma de evalúar la impureza de t es mediante la Entropía, $H(t)$, medida por el número de bits asociados a la decisión entre las clases representadas en ese nodo.
D) Si t se considera terminal u hoja, conviene asignarle una etiqueta de clase $c^\star$ tal que $N_{c^\star}(t)/N(t)$ sea máxima.

4 ☐ (Examen de SIN del 15 de enero de 2014; examen del bloque 2; cuestión 8)
Sea un problema de clasificación en C clases, $c = 1, \ldots, C$, para el que se ha aprendido un árbol de clasificación T. Sea t un nodo de T cuya impureza viene dada mediante la entropía, $H(t)$, asociada a las probabilidades a posteriori de las clases en t, $P(1 \mid t), \ldots, P(C \mid t)$. El nodo t será máximamente puro cuando:

A) Las clases sean equiprobables; esto es, $P(1 \mid t) = \cdots = P(C \mid t) = \frac{1}{C}$.
B) Exista una clase c^* de mayor probabilidad que el resto; esto es, $P(c^* \mid t) > P(c \mid t)$ para todo $c \neq c^*$.
C) Exista una clase c^* de probabilidad 1; esto es, tal que $P(c^* \mid t) = 1$.
D) Ninguna de las anteriores.

5 ☐ (Examen de SIN del 15 de enero de 2014; examen del bloque 2; cuestión 9)
Sea un problema de clasificación en 2 clases, $c = 1, 2$, para objetos representados mediante vectores de características reales bidimensionales; esto es, de la forma $\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2$. Se tienen 4 muestras de entrenamiento: $\mathbf{y_1} = (1, 0.2)^t$, perteneciente a la clase 1; y $\mathbf{y_2} = (2, 0.2)^t$, $\mathbf{y_3} = (3, 0.8)^t$ y $\mathbf{y_4} = (1, 0.8)^t$, pertenecientes a la clase 2. Queremos construir un árbol de clasificación empleando el decremento de impureza (medida en términos de entropía) para medir la calidad de una partición de un nodo. En el caso del nodo raíz y considerando sólo la característica y_1, ¿cuál de las siguientes afirmaciones es *cierta*? (Nota: $\log_2(1/3) = -1.585$ y $\log_2(2/3) = -0.585$).

A) La mejor partición es $y_1 \leq 1$.
B) La mejor partición es $y_1 \leq 2$.
C) La mejor partición es $y_1 \leq 3$.
D) Ninguna de las anteriores.

6 ☐ (Examen de SIN del 15 de enero de 2014; examen del bloque 2; cuestión 10)
Sea un problema de clasificación en C clases, $c = 1, \ldots, C$, para el que se ha aprendido un árbol de clasificación T. Sea t un nodo terminal de T en el que se han estimado las probabilidades a posteriori de las clases $\hat{P}(1 \mid t), \ldots, \hat{P}(C \mid t)$. Un criterio simple y eficaz para asignar una etiqueta de clase a t es:

A) La de una clase de probabilidad a posteriori mínima.
B) La de una clase de probabilidad a posteriori próxima a la media (i.e. $\frac{1}{C}$).
C) La de una clase de probabilidad a posteriori máxima.
D) Ninguna de las anteriores.

7 ☐ (Examen de SIN del 28 de enero de 2014; examen final; cuestión 4)
Sea un problema de clasificación en 2 clases, $c = 1, 2$, para objetos representados mediante vectores de características reales bidimensionales; esto es, de la forma $\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2$. Sea T un árbol de clasificación para este problema y sea t un nodo interno de T. Sean B_1 y B_2 las cajas de mínima inclusión de los objetos de la clase 1 y 2 en t, respectivamente. Dichas cajas están caracterizadas por las coordenadas de sus esquinas inferior izquierda y superior derecha de la forma $[\text{mín}\, y_1, \text{mín}\, y_2] \times [\text{máx}\, y_1, \text{máx}\, y_2]$, siendo $B_1 = [1.5, 0.6] \times [2.3, 3.5]$ y $B_2 = [2.5, 1.3] \times [3.8, 3.2]$. En términos de decremento de impureza (medida como entropía), ¿cuál de las siguientes particiones de t es mejor?

A) $y_1 \leq 3.8$
B) $y_1 \leq 2.3$
C) $y_2 \leq 1.3$
D) $y_2 \leq 3.5$