Guía de Ejercicios Aprendizaje de Máquinas Inteligencia Artificial (CC52A)

6 de noviembre de 2007

1. Arboles de Decision

1. Investigue las ventajas y desventajas de los árboles de decisión versus los siguientes tipos de modelos: (a) redes Neuronales, (b) redes Bayesianas, (c) discriminantes lineales.

2. En la construcción de un árbol de decisión, diga los tipo de splits se favorecen usando GiniSplit versus Entropía.

3. Dé como ejemplo un conjunto de datos y un split deficiente sobre ellos que se obtendría usando tanto GiniSplit como Entropía, debido a que éstos dos criterios favorecen splits con muchos hijos. Investigue dos técnicas para solucionar el problema.

4. Explique el fenómeno de sobreajuste de un árbol de decisión dando un gráfico que muestre cómo se comporta el error de un árbol versus su tamaño.

5. Corra J48 (Weka) sobre whether.nominal sin poda y obtenga un modelo. Seleccione un nodo y diga si podarlo o no (sin usar Weka) en base a la técnica de poda basada en estimación del error usando datos de prueba.

6. Sobre el árbol obtenido anteriormente aplique poda usando test de Chi-cuadrado y poda basada en reglas.

7. Construya un árbol y conjunto de datos en los cuales no se pode un nodo usando poda basada en reducción del error y se pode el nodo usando poda basada en PDM.

8. ¿Qué ventajas tiene la poda basada en reglas sobre poda basada en reducción del error sin transformar a reglas?

9. Dé un ejemplo concreto de un árbol de decisión y datos de prueba para los cuales la poda basada en reglas es ventajosa sobre la poda basada en reducción del error sin reglas.

10. Dado un conjunto D de N objetos con un atributo numérico A y un atributo de la clase C, con dos clases. Supongamos que estamos evaluando el GINISPLIT de "splits"binarios de la foma $(A \leq v, A > v)$ y ya sabemos el número de objetos en cada clase. ¿Cuántos contadores y operaciones de incremento de éstos se necesitan para computar los GINISPLITS en una única lectura de D, (1) asumiendo que D está ordenado por A, y (2) asumiendo que D no está ordenado por A?

11. Diga cuánto sería el costo $Costo(D \mid M)$ para (a) un árbol a cuyas hojas están asociado sólo objetos de una única clase, y (b) un árbol cuyas hojas tienen objetos en partes iguales en dos clases.

12. Considere el árbol de la figura 1. Tras aplicar al modelo a un conjunto de datos de poda, obtenemos las siguientes frecuencias de ocurrencia de las clases para las hojas del subárbol bajo el nodo t_1:

Nodo	YES	NO
t_3	20	5
t_4	4	30
t_5	4	20

    Muestre cómo quedaría el subárbol bajo t_1 después de aplicar poda basada en reducción del error (use 80 % de nivel de confianza).

13. Para el árbol y datos de poda del ejercicio 11, si conviene podar el nodo t_2 aplicando test de chi-cuadrado con un 95 % de nivel de confianza.

14. Para el árbol y datos de poda del ejercicio 11, considere la regla asociada a la hoja t_4. Diga si conviene o no podar la condición $\#\texttt{CHILDREN} > 0$ de esta regla (use nivel de confianza 80 % para estimar el error).

15. Para el árbol y datos de poda del ejercicio 11, diga si conviene reemplazar el nodo t_1 por una hoja, usando el test de chi-cuadrado con un 95 % de nivel de confianza.

16. Para el árbol y datos de poda del ejercicio 11 diga usando poda basada en el principio de descripción mínima si conviene o no podar el subárbol bajo el nodo t_2.

2. Evaluación de Modelos de Clasificación

1. Si en un experimento de validación cruzada sobre 100 datos con 6 “folds” se obtienen las siguientes tasas de errores: 0,2, 0,2, 0,3, 0,4, 0,25, 0,7, dé una estimación del error con 90 % de confianza.

2. Usando la estadística “t-test”, compare el modelo del ejemplo anterior con otro que obtuvo las siguientes tasas de errores (en el mismo experimento de validación cruzada): 0,25, 0,34, 0,1, 0,3, 0,5, 0,7.

3. Supongamos que evaluamos un modelo y encontramos $r = 300$ datos mal clasificados. Estime el error, con nivel de confianza 90 %, si tenemos $n = 1000$, $n = 2000$ y $n = 3000$ datos. ¿Cuántos datos se requieren para tener un error en un rango de 0,1 con nivel de confianza 90 %?

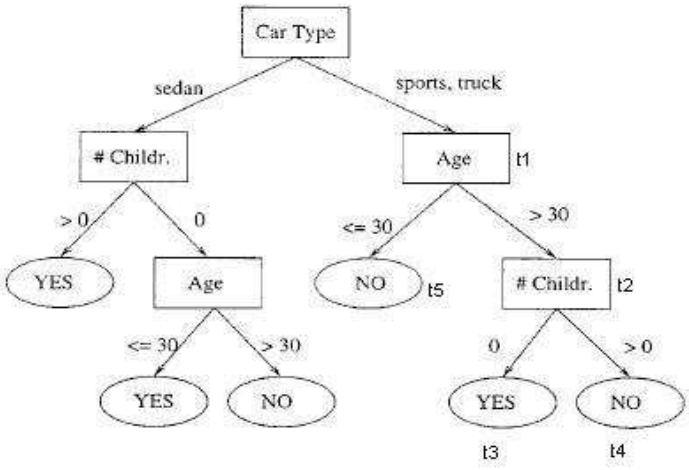

Figura 1: Arbol de Decisión

4. Diga las ventajas y desventajas de las siguientes estrategias para evaluar un modelo: "hold out", validación cruzada, "leave-One-Out Holdout". ¿En qué situaciones usaría cada una de ellas?.

5. ¿Por qué es importante contabilizar el costo de los errores en modelos de clasificación? Dé un ejemplo realista. Explique dos técnicas para generar modelos que minimicen el costo de los errores.

3. Resolución de algunos ejercicios

Ejercicio 1.12

De la fórmula vista en clases para el estimar el error, el error si podamos un nodo t es el siguiente:

$$\texttt{ErrorPodado}(t) = \frac{R_t + \frac{Z^2}{2N_t} + Z\sqrt{\frac{R}{N_t} - \frac{R^2}{N_t} + \frac{Z^2}{4{N_t}^2}}}{1 + \frac{Z^2}{N_t}}, \tag{1}$$

done, N_t es el número de datos en el nodo t, R_t es la frecuencia de datos mal clasificados en el nodo t, y Z es tal que $P(X \geq Z) = \frac{1-NivelConfianza}{2}$, donde X distribuye $Normal(O, 1)$. Z se obtiene consultando una tabla de "valores z" para la distribución normal unitaria. Para $NivelConfianza = 80\,\%$, obtenemos $Z = 1{,}28$.

Si no podamos un nodo t, su error sería:

$$\texttt{Error}(t) = \begin{cases} \Sigma_i \frac{N_{t_i}}{N_t} \texttt{Error}(t_i) & \text{si } t \text{ es un nodo interior} \\ \texttt{ErrorPodado}(t) & \text{si } t \text{ es un nodo hoja} \end{cases} \tag{2}$$

donde los t_i's son los nodos hijos de t.

Para asegurarnos que el resultado de la poda sea el subárbol de t_1 con menor error, procedemos podando en forma "bottom-up".

Luego, primero verificamos si podamos el nodo t_2. Para t_2, tenemos $N_{t_2} = 59$, $R_{t_2} = \frac{24}{59} = 0{,}4$ (recordemos que si podamos asignamos la clase mayoritaria al nodo que para t_2 podado sería NO). Aplicando la formula 1 obtenemos $\texttt{ErrorPodado}(t_2)$.

Ahora, aplicando la formula 2 el error de t_2 si no podamos es

$$\texttt{Error}(t_2) = \tfrac{25}{59}\texttt{ErrorPodado}(t_3) + \tfrac{34}{59}\texttt{ErrorPodado}(t_4).$$

Luego, podamos t_2 ssi $\texttt{Error}(t_2) \geq \texttt{ErrorPodado}(t_2)$.

El ejercicio se completa calculando los errores y continuando el procedimiento de poda evaluando si podamos el nodo t_1 (en base a si se podó o no el nodo t_2).

Ejercicio 1.13

Denotemos X la variable binaria que define el split, donde:

$$X = \begin{cases} 0 & \text{si } \#\texttt{CHILDREN} = 0 \\ 1 & \text{si } \#\texttt{CHILDREN} > 0. \end{cases} \tag{3}$$

Luego usamos el test de chi-cuadrado para decidir si existe dependencia X y la variabl de la clase, para los datos asociados al nodo t_2. La tabla de contingencia entre las dos variable seria:

	$X = 0$	$X = 1$	Total
$Clase = \texttt{YES}$	20	4	24
$Clase = \texttt{NO}$	5	30	35
Total	25	34	59

El ejercicio se completa calculando el estimador χ^2 para esta tabla, si este es menor que 3,84 conviene podar el nodo.

Ejercicio 1.15

Debemos evaluar si el subárbol completo bajo t_1 es significativo de acuerdo al estimador chi-cuadrado. Definimos una variable X ahora asociada a los splits combinados del subárbol.

$$X = \begin{cases} 0 & \text{si } \texttt{AGE} \leq 30 \\ 1 & \text{si } \texttt{AGE} > 30 \text{ y } \#\texttt{CHILDREN} = 0 \\ 2 & \text{si } \texttt{AGE} > 30 \text{ y } \#\texttt{CHILDREN} > 0. \end{cases} \tag{4}$$

	$X = 0$	$X = 1$	$X = 2$	Total
Clase = `YES`	4	20	4	28
Clase = `NO`	20	5	30	55
Total	24	25	34	83

El ejercicio se completa calculando el estimador χ^2 para esta tabla, si este es menor que 3,84 conviene reemplazar el nodo por una hoja.

Ejercicio 1.16

El principo de descripción mínima dice que el costo de un modelo M sobre un conjunto de datos D se define como:

$$C(M, D) = C(M) + C(D|M),$$

donde $C(M)$ es el costo en bits de trasmitir el modelo y $C(D|M)$ es el costo en bits de trasmitir los datos codificados usando el modelo.

Si el nodo t se poda, el modelo sólo contiene la distribución de probabilidades de las clases asociada al nodo (suponemos que el árbol guarda esta distribución en cada hoja, en lugar de guardar sólo el valor de la clase mayoritaria). Luego, $C(M)$ es el costo de enviar esta distribución y $C(D|M)$ es el costo de enviar las clases asumiendo que la codificamos eficientemente usando la distribución de prob. asociada al nodo.

Luego, definimos el costo si el nodo se poda:

$$\texttt{CostoPodado}(t) = N_t \ \mathit{Entropia}(t) + \log \binom{N_t + m - 1}{m - 1}, \tag{5}$$

donde N_t es el número de datos en t y m es el número de clases. La entropía del nodo se calcula según se explicó en clases (ver apuntes) (y como se explica arriba representa $C(D|M)$).

Como se explicó en clases, la expresión $\binom{N_t+m-1}{m-1}$ representa el número posible de distribuciones de probabilidad que puede tener un nodo con N_t elementos y m clases. Este número es equivalente al número de soluciones enteras positivas de la ecuación:

$$N_{t,c_1} + N_{t,c_2} + \ldots + N_{t,c_m} = N_t,$$

donde N_{t,c_i} es el número de datos que pertenecen a la clase c_i en el nodo t.

De la fórmula 5, tenemos el siguiente costo si t_2 se poda:

$$\texttt{CostoPodado}(t) = 59\ \mathit{Entropia}(t_2) + \log \binom{59+2-1}{2-1}$$

Y tenemos que

$$\mathit{Entropia}(t_2) = -\tfrac{35}{59} \log \tfrac{35}{59} - \tfrac{24}{59} \log \tfrac{24}{59} = 0,97$$

Luego

$$\texttt{CostoPodado}(t_2) = 59 \times 0,97 + \log 60 = 63,13$$

Ahora debemos estimar el costo de un nodo t si no lo podamos. En este caso debemos incluír el costo en bits de codificar el split del nodo. Luego tenemos

$$\texttt{Costo}(t) = \begin{cases} 1 + \log(A) + \log(S) + \Sigma_i \texttt{Costo}(t_i) & \text{si } t \text{ es un nodo interior} \\ 1 + \texttt{CostoPodado}(t) & \text{si } t \text{ es un nodo hoja} \end{cases} \quad (6)$$

donde A es el número de atributos que considera el modelo, S es el número posible de splits que se podrían tener usando el atributo del nodo t, y los t_i's son los nodos hijo de t. El 1 de cada fórmula es el costo de codificar si el nodo es interior u hoja (con este bit codificamos la estructura del árbol).

Para el nodo t_2 tenemos $A = 3$ y $S = 10$ (consideramos 10 posibles splits, es decir consideramos que la condición de la izquierda de cada split puede ser uno de $\#\texttt{CHILDREN} = 0$, $\#\texttt{CHILDREN} = 1 \ldots$, hasta $\#\texttt{CHILDREN} = 9$).

$$\texttt{Costo}(t_2) = 1 + \log(3) + \log(10) + 1 + \texttt{CostoPodado}(t_3) + 1 + \texttt{CostoPodado}(t_4).$$

Usando la fórmula 5, tenemos que $\texttt{CostoPodado}(t_3) = 25 \times 0,72 + \log 26 = 22,70$ y tenemos que $\texttt{CostoPodado}(t_4) = 34 \times 0,52 + \log 35 = 22,81$. Luego

$\texttt{Costo}(t_2) = 1 + 1,58 + 3,32 + 1 + 22,70 + 1 + 22,81 = 53,42$.

Luego $\texttt{Costo}(t_2) < \texttt{CostoPodado}(t_2)$, entonces no conviene podar t_2.