Guía de Ejercicios: Árboles de Decisión

Profesor Carlos Hurtado

25 de agosto de 2008

1. Construcción de Árboles de Decisión

Problema 1 *Dé un ejemplo concreto de un conjunto de datos donde usando índice Gini se selecciona un split distinto al split seleccionado usando entropía.*

Problema 2 *Explique el fenómeno de sobreajuste (overfitting) de un árbol de decisión dando un gráfico que muestre cómo se comporta el error de un árbol en datos de entrenamiento y datos objetivos versus su tamaño.*

Problema 3 *Dado un conjunto D de N objetos con una variable numérica A y una variable binaria de la clase C . Supongamos que consideramos separadores binarios de la forma $(A \leq v, A > v)$, donde v es un valor de $\text{dom}(A)$ que aparece en los datos.*

(i) *Escriba un algoritmo que realice una única lectura (iteración) de D y calcule el GiniSplit de cada separador siguiendo la siguiente estrategia: para cada separador defina cuatro contadores, luego en la iteración sobre los datos en D actualice los contadores, a medida que vaya accediendo a cada dato.*

(ii) *¿Cuántos contadores y operaciones de incremento de éstos se necesitan para computar los GiniSplits en su algoritmo ?*

(iii) *Escriba un algoritmo eficiente que calcule los GiniSplits de los separadores, ahora suponiendo que D está ordenado por A y que se conoce el número de datos en cada clase.*

(iv) *¿Cuántos contadores y operaciones de incremento de éstos se necesitan para el algoritmo con los datos ordenados?*

2. Poda de Árboles de Decisión

Problema 4 (i) *Corra J48 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. (ii) Sobre el árbol obtenido anteriormente diga si podar o no usando test de Chi-cuadrado.*

Problema 5 *El costo de codificar un conjunto de datos D con un modelo M está dado por: $\text{Costo}(M, D) = \text{Costo}(D | M) + \text{Costo}(M)$. Diga dos formas distintas de medir estos costos para árboles de decisión. Entregue para cada caso las fórmulas exactas.*

Problema 6 Supongamos que tenemos un conjunto D de $n = 2^k$ datos, donde k es un entero positivo. La variable de la clase de los datos es binaria. Además, tenemos un árbol de decisión M con separadores binarios y balanceados (es decir, cada separador divide los datos en partes iguales). Cada hoja del árbol tienen un único dato.

Demuestre que, de acuerdo al principio de descripción mínima, conviene podar M completamente.

Hint 1: Un árbol binario completo de L niveles tiene $\approx 2^{L+1}$ nodos.

Hint 2: Recuerde que el costo de descripción de un conjunto de datos D es

$$\text{Costo}(D) = N_D \text{Entropy}(D) + \log \binom{N_D+m-1}{m-1},$$

donde N_D es el número de datos y m el número de clases. El costo de descripción de un árbol A es:

$$\text{Costo}(A) = \text{Costo Separadores} + N_A,$$

donde N_A es la cantidad de nodos del árbol.

Problema 7 Considere un conjunto de n datos D , con variable de clase binaria, cuya entropía con respecto a la clase es máxima. A partir de los datos, construimos un árbol M de L niveles con splits binarios y balanceados de la forma $X \leq v, X > v$, donde v puede tomar hasta n valores distintos. Diga cuánto debe ser la entropía promedio de las hojas en función de n y L , para que no convenga podar el árbol completo.

3. Resolución de Problemas

Resolución Problema 6. Demostraremos que

$$C(D) \leq C(D, M)$$

El costo de descripción de los datos, es decir, si podemos completamente el árbol es:

$$C(D) = n \text{Entropy}(D) + \log(n+1) \leq n + \log(n+1)$$

El costo de descripción de los datos y el árbol es:

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

Como en cada hoja tenemos un único dato, que pertenece a una única clase, el costo de los datos D_h de una hoja h es $C(D_h) = 0 + \log 2 = 1$. Luego, sumamos los costos de todas las hojas y tenemos $C(D|M) = n$.

De acuerdo a la fórmula dada en el hint, el árbol tiene $2^{\log n} 2$ nodos, es decir $2n$ nodos. Luego, el costo del modelo $C(M) = 2n + n \text{split}$. Luego tenemos:

$$C(M) \geq 2n, \text{ luego tenemos:}$$

$$C(D, M) \geq 3n.$$

Bastaría verificar que

$$n + \log(n+1) \leq 3n, \text{ es decir}$$

$$\log(n+1) \leq 2n, \text{ lo que se cumple para } n \geq 1.$$