Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ciencias de la Computación

Semestre Primavera 2008
Inteligencia Artificial
CC52A

Profesores: Carlos Hurtado, Pablo Barceló
Auxiliar: Gonzalo Ríos
Fecha: 10 de Septiembre

Auxiliar 5: Sobreajuste y Poda

1 Materia

1. **Error de predicción** de un modelo se debe a tres causas:
   (a) **Varianza**: datos de entrenamiento no son representativos de datos objetivos.
   (b) **Sesgo**: modelo es limitado y aunque lo entrenemos con datos de prueba representativo igual tendremos error.
   (c) **Ruido**: Nuestro espacio de variables no separa adecuadamente las clases.
2. **Sobreajuste**: Dado un espacio de modelos M, un modelo m en M es sobreajustado si existe otro modelo m' en M tal que:
   (a) m tiene menor error que m' en datos de entrenamiento
   (b) m' tiene menor error que m en datos objetivo
3. **Poda**: Mecanismo para obtener árboles con menor error de predicción
   (a) **Pre-poda**: Parar la construcción del árbol en algunas nodos
   (b) **Post-poda:** Construir un árbol complejo (posiblemente sobreajustado) y podarlo después.
4. **Criterios de poda**
   (a) No expandir si GiniSplit < k
   (b) No expandir si el nodo tiene menos de k datos
   (c) No expandir si test de chi-cuadrado no rechaza independencia
5. **Poda Basada en reducción del error**
   (a) Podamos si el error de predicción del árbol podado disminuye
   (b) Recorremos el árbol podando de abajo hacia arriba.
   (c) El podar el árbol de esta forma garantiza que obtenemos el subárbol de menor error.
6. **Límit superior de un intervalo de Wilson**
   (a) p taza de error
   (b) n número de datos
   (c) C nivel de confianza
   (d) z el valor tal que Pr(-z$\leq N(0,1) \leq z) = C$
   (e) $\pi = \frac{p+\frac{z^2}{2n}+z\sqrt{\frac{p}{n}-\frac{p^2}{n}+\frac{z^2}{4n^2}}}{1+\frac{z^2}{n}}$

   Si C=75% $\Longrightarrow$ $z = 0.69$

7. **Poda Basada en Principio de Descripción Mínima**

   (a) El mejor modelo es aquel que minimiza el costo de codificar el modelo C(M) más el costo de codificar los datos usando el modelo C(D|M).

   (b) C(D|M) se puede ver como el error.

   (c) El PDM integra error y complejidad del modelo en una medida única y en bits.

   (d) Costo(M): Costo de codificación del árbol (como grafo) más costo de codificar los splits

   (e) Cost(D|M): se estima como el costo de codificar los errores o usando entropía.

   (f) Costo(M,D)=Costo(M)+Costo(D|M)

8. **Costo de Descripción de un Conjunto de Datos:** Para transmitir D debemos transmitir la distribución de las clases P_D, y luego las clases de los datos.

$$Costo(D) = N_D \times Entropia(P_D) + CostoCod(P_D)$$

   Donde $CostoCod(P_D) = \log \binom{N_D + m - 1}{m - 1}$, donde N_D es el número de datos, y m el número de clases.

9. **Costo de descripción Conjunto de Datos más Split:**

   (a) $Costo(D, Split) = Costo(Split) + Costo(D|Split)$

   (b) $Costo(D|Split) = Costo(D_1) + Costo(D_2)$

   (c) $Costo(Split) = \log A + \log B$, donde A es el número de variables y B es el número de posibles splits usando una variable.

10. **Costo de descripción Conjunto de Datos más Árbol:**

    (a) $Costo(D, Arbol) = Costo(Arbol) + Costo(D|Arbol)$

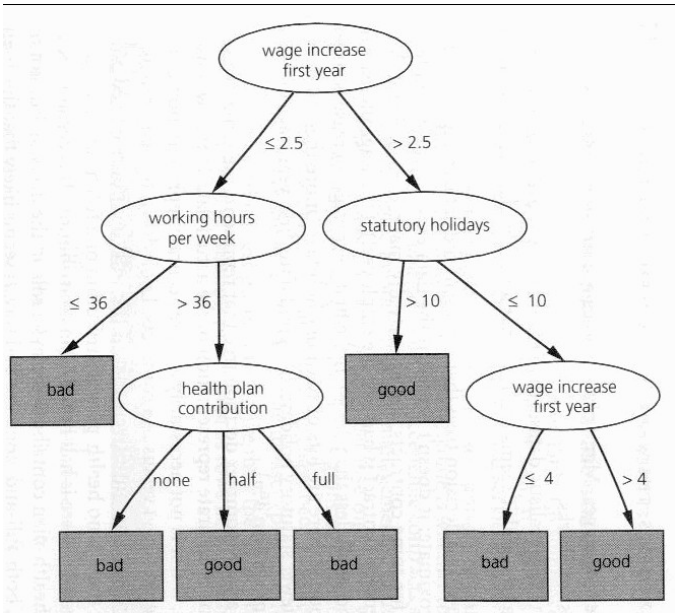
    (b) $Costo(D|Arbol) = \sum_i Costo(D_i)$, donde D_i son los datos de una hoja del árbol

    (c) $Costo(Arbol) = \sum_i Costo(Split_i) + P$, donde P es el número de nodos

11. **Forma alternativa del costo de descripción Conjunto de Datos más Árbol:**
    $Costo(D, Arbol) = Costo(Split) + \sum_i \frac{|D_i|}{|D|} Costo(D_i, Arbol_i)$

2 Poda Basada en reducción del error

Tenemos el árbol de decisión



Y los datos de poda

WI	WH	HP	Clase
2	38	none	Good
1	40	none	Good
1.5	40	none	Bad
2	40	none	Bad
2.2	38	none	Bad
1	40	none	Bad
1.5	40	half	Good
2	38	half	Bad
2	38	full	Good
1	40	full	Good
1	40	full	Bad
1	38	full	Bad
2	38	full	Bad
2	38	full	Bad

Luego, la Tabla de frecuencias es

Nodo	Clase	Num Good	Num Bad	Errores
c1	Bad	2	4	2
c2	Good	1	1	1
c3	Bad	2	4	2

Luego, estimando los errores obtenemos

Nodo	Num. Errores	Num Datos	p	Error Est.
c1	2	6	0.33	0.47
c2	1	2	0.5	0.72
c3	2	6	0.33	0.47

Luego, el error del nodo sin podar es

$$A = \frac{6}{14} \times 0.47 + \frac{2}{14} \times 0.72 + \frac{6}{14} \times 0.47 = 0.505\,71$$

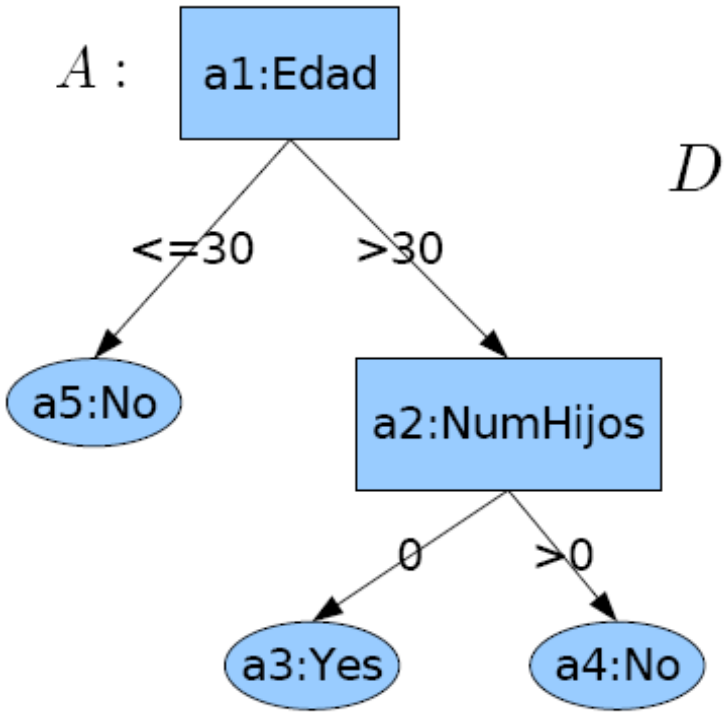
Luego, si podamos el nodo, hay 9 datos de la clase BAD y 5 datos de la clase GOOD, luego el nodo se etiqueta con la clase BAD y nuestro error es:

GOOD	BAD	ERROR	p	Error. est.
5	9	5	$\frac{5}{14} = 0.357\,14$	B=0.46

Como A>B, podamos el nodo.

3 Poda Basada en principio de descripción mínima

Tenemos el árbol de decisión



D :

Nodo	NumYes	Num No
a3	20	5
a4	4	30
a5	4	30

El costo de codificar el árbol está dado por

$$Costo(A) = Costo(Split_1) + Costo(Split_2) + 5$$

- Para el Split_1, tomaremos la edad máxima como 90, luego $\text{Costo}(\text{Split}_1) = \log_2 90 + \log_2 2 = 7.491\,9$
- Para el Split_2, tomaremos el número de hijos máxima como 10, luego $\text{Costo}(\text{Split}_2) = \log_2 10 + \log_2 2 = 4.321\,9$
- Luego, Costo(A)$=5 + 7.49 + 4.32 = 16.81$

El costo de codificar los datos, dados el árbol está dado por

$$Costo(D|A) = Costo(D_3) + Costo(D_4) + Costo(D_5)$$

- $Costo(D_3) = 25 \times Entropia(\frac{20}{25}, \frac{5}{25}) + \log_2 \binom{25+2-1}{2-1}$

  $25 \times \left(-\frac{20}{25}\log_2(\frac{20}{25}) - \frac{5}{25}\log_2(\frac{5}{25})\right) + \log_2 26 = 22.749$
- $Costo(D_4) = 34 \times Entropia(\frac{30}{34}, \frac{4}{34}) + \log_2 \binom{34+2-1}{2-1}$

  $34 \times \left(-\frac{30}{34}\log_2(\frac{30}{34}) - \frac{4}{34}\log_2(\frac{4}{34})\right) + \log_2 35 = 22.896$
- $Costo(D_5) = 34 \times Entropia(\frac{30}{34}, \frac{4}{34}) + \log_2 \binom{34+2-1}{2-1} = 22.896$
- $Costo(D|A) = 22.75 + 22.9 + 22.9 = 68.55$

Luego, el Costo de descripción Conjunto de Datos más el Árbol está dado por

$$Costo(D, A) = Costo(A) + Costo(D|A) = 16.81 + 68.55 = 85.36$$

Para ver si conviene o no podar el nodo a1, debemos calcular el nuevo $Costo(D, A)$, pero

- $Costo(A) = 1$
- Costo(D)$=93 \times Entropia(\frac{28}{93}, \frac{65}{93}) + \log_2 \binom{93+2-1}{2-1}$

  $93 \times \left(-\frac{28}{93}\log_2(\frac{28}{93}) - \frac{65}{93}\log_2(\frac{65}{93})\right) + \log_2 94 = 88.64$

Luego, como el costo del árbol podado es mayor que el costo del árbol sin podar, entonces se decide no podar.