

Arboles de Decisión (IV)

Carlos Hurtado L.

Depto de Ciencias de la
Computación, Universidad de
Chile

Post-poda basada en Reglas

- Es una variante de post-poda basada en reducción del error.
- La idea es transformar el árbol en un conjunto de reglas y podar cada regla por separado.
- Esto permite una poda más fina del árbol.
- La poda de cada regla se realiza con la idea de minimizar el error estimado

Post-poda basada en Reglas

- Convertir el árbol en un conjunto equivalente de reglas
- Podar cada regla: intentar eliminar condiciones de las reglas para reducir el error estimado
- El resultado es un conjunto de reglas (no necesariamente un árbol) ordenadas de menor a mayor error estimado

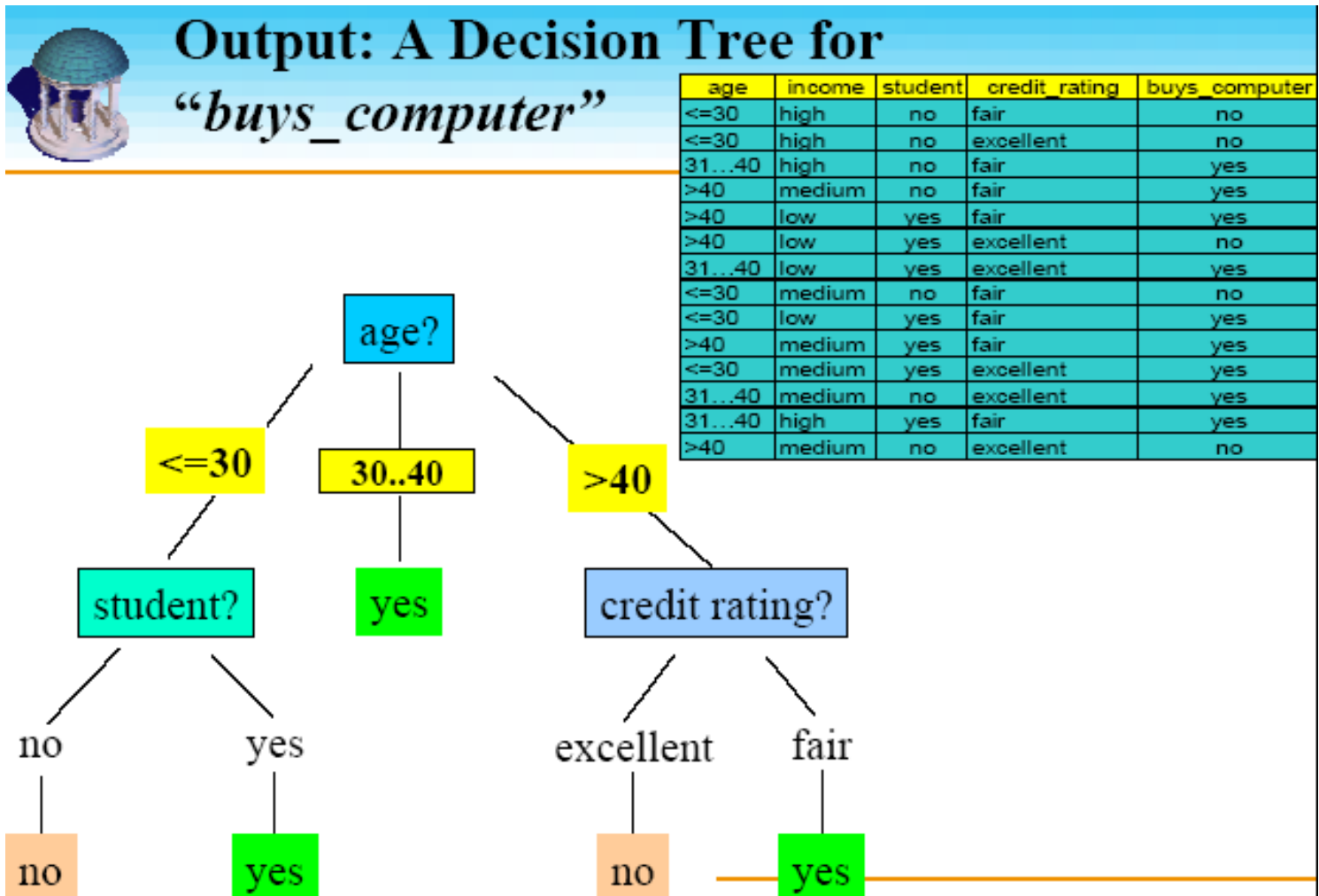
Post-poda basada en reglas

- ¿Por qué convertir el árbol en reglas para podar?
- El objetivo de la poda es minimizar el error estimado
- Permite más operaciones para reducir el árbol. Ejemplos:
 - Podemos podar un nodo interior que causa sobreajuste
 - Podemos podar sólo una condición de un split que causa sobreajuste .

Transformación de un Arbol a Reglas

- Representamos el árbol como un conjunto de reglas **IF-THEN**
- Se crea una regla por cada hoja del árbol.

Ejemplo



Ejemplo (cont.)

IF *age* = "<=30" AND *student* = "no" THEN *buys_computer* = "no"

IF *age* = "<=30" AND *student* = "yes" THEN *buys_computer* = "yes"

IF *age* = "31...40" THEN *buys_computer* = "yes"

IF *age* = ">40" AND *credit_rating* = "excellent" THEN *buys_computer* = "yes"

IF *age* = "<=30" AND *credit_rating* = "fair" THEN *buys_computer* = "no"

Reglas vs. Árboles

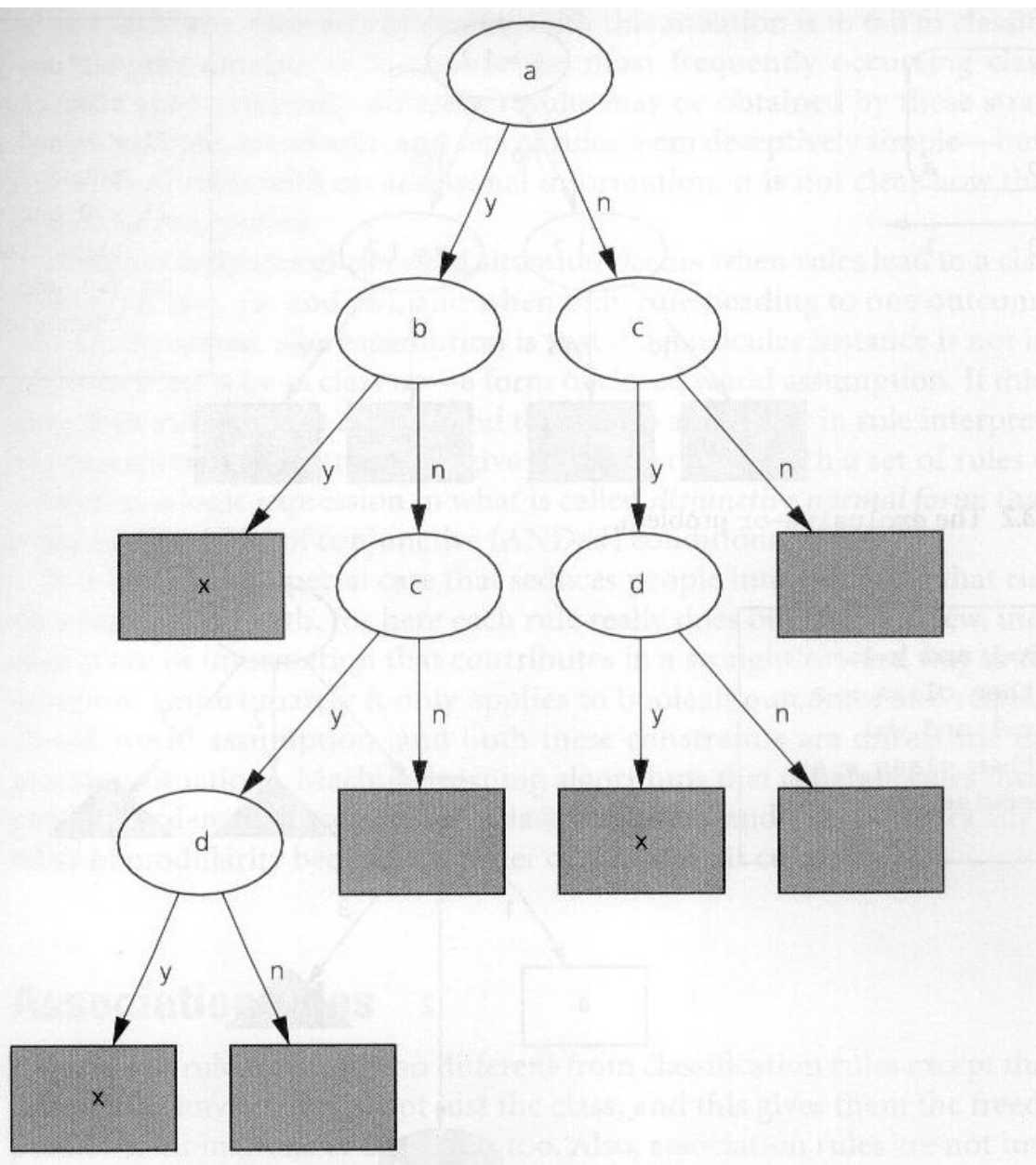
- Transformar un árbol en reglas es sencillo.
- El inverso puede ser mucho más complejo. Problema: sub-árboles replicados
- Por esto existen algoritmos que construyen reglas directamente (sin construcción previa de un árbol).
- Estas reglas se denominan *reglas de clasificación*.

Ejemplo

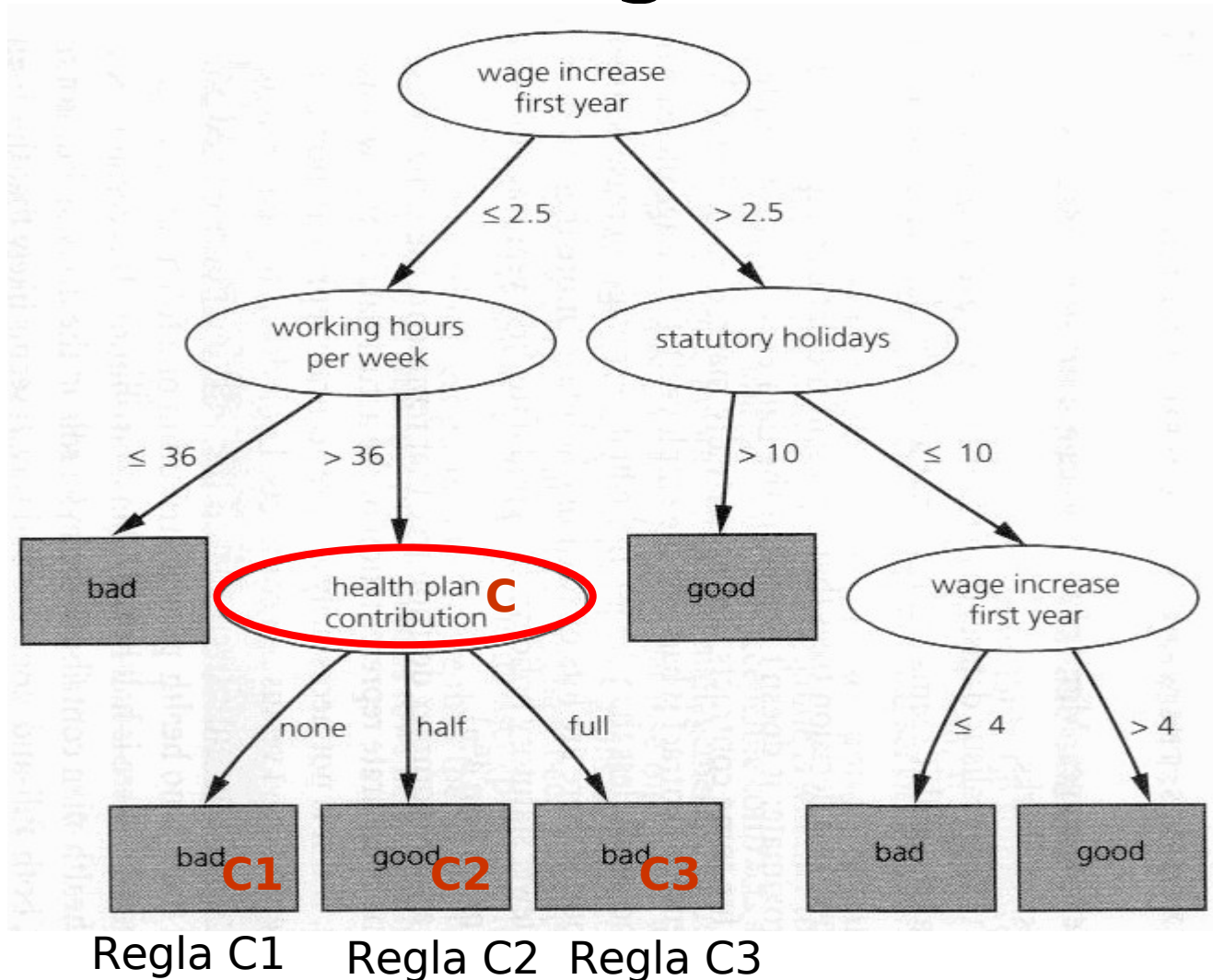
- Construir un árbol para las siguientes reglas:

If a and b then x

If c and d then x



Ejemplo: Poda Basada en Reglas



Ejemplo

- Consideremos las siguientes reglas obtenidas del árbol anterior
 - c1: $WI < 2.5$ AND $WH > 36$ AND $HP = \text{none}$ THEN class = bad
 - c2: $WI < 2.5$ AND $WH > 36$ AND $HP = \text{half}$ THEN class = good
 - c3: $WI < 2.5$ AND $WH > 36$ AND $HP = \text{full}$ THEN class = bad

Ejemplo (cont.)

- Supongamos que tenemos el sgte conjunto de 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

Ejemplo: Poda de Regla C1

- Tomamos la regla C1 y vemos si conviene podar la condición $HP = \text{none}$.
- Debemos comparar el error de las reglas:
 - c1: $WI < 2.5 \text{ AND } WH > 36 \text{ AND } HP = \text{none} \text{ THEN class} = \text{bad}$
 - c1P: $WI < 2.5 \text{ AND } WH > 36 \text{ THEN class} = \text{bad}$

Ejemplo: Poda de Regla C1

- Corremos C1 y C1P sobre los datos de poda obteniendo:

Regla	Num Good	Num Bad	Num Errores	n	p	Error Est.
c1	2	4	2	6	0.33	0.47
c1P	5	9	5	14	0.36	0.46

En este ejemplo, estimamos el error como el rango superior de un intervalo con confianza 50%, $z=0.69$

- Luego, conviene podar C1 y reemplazarla por C1P.

Ejemplo: Poda de Regla C2

- Tomamos la regla C2 y vemos si conviene podar la condición $HP = \text{none}$.
- Debemos comparar el error de las reglas:
 - c2: $WI < 2.5 \text{ AND } WH > 36 \text{ AND } HP = \text{none} \text{ THEN class} = \text{good}$
 - c2P: $WI < 2.5 \text{ AND } WH > 36 \text{ THEN class} = \text{good}$

Ejemplo: Poda de Regla C2

- Corremos C2 y C2P sobre los datos de poda obteniendo:

Regla	Num Good	Num Bad	Num. Errores	N	P	Error Est-
c2	1	1	1	2	0.5	0.72
c2P	5	9	9	14	0.64	0.73

- Luego no conviene podar C2.

Ejemplo: Poda de Regla C3

- Tomamos la regla C3 y vemos si conviene podar la condición $HP = \text{full}$.
- Debemos comparar el error de las reglas:
 - c3: $WI < 2.5 \text{ AND } WH > 36 \text{ AND } HP = \text{full}$
THEN class = bad
 - c3P: $WI < 2.5 \text{ AND } WH > 36$ THEN class = bad

Ejemplo: Poda de Regla C3

- Corremos C3 y C3P sobre los datos de poda obteniendo:

Regla	Num Good	Num Bad	Num Errores	N	p	Error Est.
c3	2	4	2	6	0.33	0.47
c3P	5	9	5	14	0.36	0.46

En este ejemplo, el error se estima con confianza 50% y $z=0.69$

- Luego, conviene podar C3 y reemplazarla por C3P.

Reglas Resultantes

- 1.c1P: $WI < 2.5$ AND $WH > 36$ THEN class = bad (error est. = 0.47)
 - 2.c3P: $WI < 2.5$ AND $WH > 36$ THEN class = bad (error est. = 0.47)
 - 3.c2: $WI < 2.5$ AND $WH > 36$ AND $HP = \text{none}$ THEN class = good (error est. = 0.72)
- Las reglas se ejecutan en este orden.
 - Reglas c1P y c3P son equivalentes y cubren a la regla c2. Luego, podemos eliminar c3p y c2.

Complejidad de Poda Basada en Reglas

- En el ejemplo anterior sólo intentamos podar la última condición de cada regla.
- Obtenemos el mismo árbol que el que obtuvimos usando poda basada en reducción del error sin reglas.
- Para aprovechar la potencialidad de la poda basada en regla debemos considerar podar otras condiciones.

Complejidad de Poda basada en Reglas

- Podemos intentar podar varias condiciones.
- De una regla con c condiciones, tenemos 2^c posibles reglas podadas y cualquiera de ellas puede ser la mejor.
- Esta poda exhaustiva es costosa.
- La poda basada en reglas es costosa; en la práctica se implementan versiones restringidas de ella.