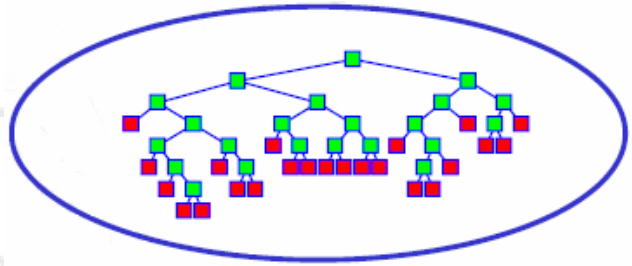


Árboles de decisión

CART – Clasification and Regression Tree

FRANCISCO CISTERNAS

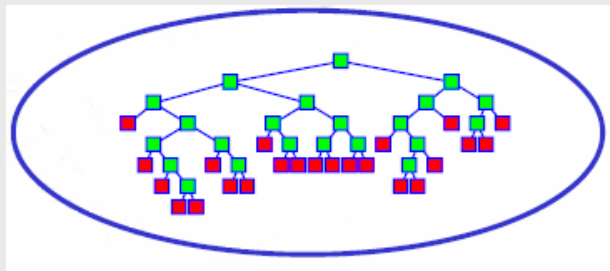
FABIÁN MEDEL

Departamento de Ingeniería Industrial

Universidad de Chile

Clase de hoy

1. Refresco de árboles
   - Indicador de ganancia
2. Sólo CART
   - Construcción del Árboles
   - Poda Costo-Eficiencia
   - Árboles de Regresión
   - Árboles de Clasificación
     - Gini
     - Twoin



Inducción de Árboles de Decisión Top-Down

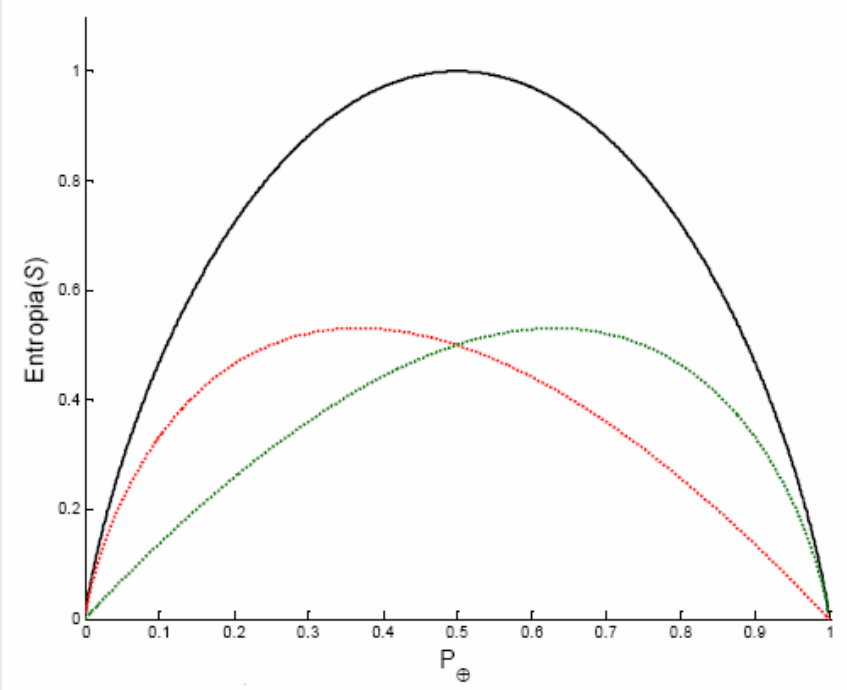
1. El loop es el siguiente:
   - Vemos Cual es el mejor atributo en el nodo actual para dividir el problema, llamémoslo *A*.
   - Asignamos *A* como atributo de decisión para el nodo.
   - Para los distintos valores de *A*, creamos un nuevo nodo hijo o descendiente.
   - Ordenamos los datos de entrenamiento conforme a la partición generada por cada hijo.
   - Si los datos de entrenamiento quedan perfectamente clasificados en el nodo, entonces nos detenemos. Si no seguimos iterando en los otros nodos hijos.

¿Pero cual de los atributos es el mejor para poner en el nodo?

Entropía

"baja Entropía": el atributo es uniforme

"Alta Entropía": el atributo es variado e interesante



- S es un conjunto de muestras de entrenamiento
- *p*(+): es la proporción de ejemplos positivos en *S*
- *p*(-): es la proporción de ejemplos negativos en *S*
- La Entropía mide la "impureza" de *S*.

$$Entropia(S) \equiv -p_{(+)} \cdot \log_2\left(p_{(+)}\right) - p_{(-)} \cdot \log_2\left(p_{(-)}\right)$$

Entropía

1. *Entropia* (*S*) = número esperado de bits necesarios para codificar la clase de una muestra aleatoria de S.
2. ¿Por Qué?
   - Teoría de Información dice: para un mensaje que tiene una probabilidad *p* el largo óptimo del código asignado debe ser –log2*p* bits
3. Entonces el número esperado para codificar en (+) o (-) un miembro aleatorio de *S* es:

$$-p_{(+)} \cdot \log_2\left(p_{(+)}\right) - p_{(-)} \cdot \log_2\left(p_{(-)}\right)$$

Entropía condicional a un ejemplo

- La Definición

X	Y
Matemáticas	SI
Historia	NO
Ciencias Nat	SI
Matemáticas	NO
Matemáticas	NO
Ciencias Nat	SI
Historia	NO
Matemáticas	SI

$Entropia(\ Y \mid X = v\ )$ = La entropía de Y dentro de los registros en los cuales X tienen el valor v

= Número esperado de bits para transmitir Y si a ambos lados se conoce el valor de X

$$= \sum_j P(X = v_j)\, Entropia(Y \mid X = v_j)$$

v_j	$P(X = v_j)$	$Entropia(Y \mid X = v_j)$
Matemáticas	0.5	1
Historia	0.25	0
Ciencias Nat	0.25	0

$$Entropia\ (Y \mid X = v\ ) = 0.5 \times 1 + 0.25 \times 0 + 0.25 \times 0 = 0.5$$

Ganancia de Información

1. *Gain*(*S* | *A*) = reducción de entropía debido a que se condiciona según el atributo *A*

$$Gain(S \mid A) = Entropia(S) - \sum_{r \in Valores(A)} \frac{|S_r|}{|S|} \times Entropia(S_r)$$

X	Y
Matemáticas	SI
Historia	NO
Ciencias Nat	SI
Matemáticas	NO
Matemáticas	NO
Ciencias Nat	SI
Historia	NO
Matemáticas	SI

- En el Ejemplo Anterior

$$Gain(Y \mid X) = Entropia(Y) - Entropia(Y \mid X = v)$$

$$Gain(Y \mid X) = 1 - 0.5 = 0.5$$

Principio de Okham

… En inglés Occam's Razor dijo que la Naturaleza:

"Nunquam poneda est pluralitas sin necesitate"

- Es decir, la naturaleza no crea complejidad sin necesidad
- Por esto es preferible un árbol pequeño, es decir, hipótesis cortas, frente a uno largas y complejas

Aspectos importantes

1. Idea Tras la Creación del Árbol:
   - Minimizar el error de asociado al reemplazo de un conjunto de valores por un valor único (resusbtitution error)
2. Criterios de Discriminación entre Atributos:
   - GINI
   - TWOIN

Aspectos importantes

3. Criterios de Detención
   - ❑ Número Mínimo de Muestras en el Nodo
4. Criterios de Poda
   - ❑ Poda de Costo-Complejidad
5. Característica distintiva:
   - ❑ Sólo son árboles binarios

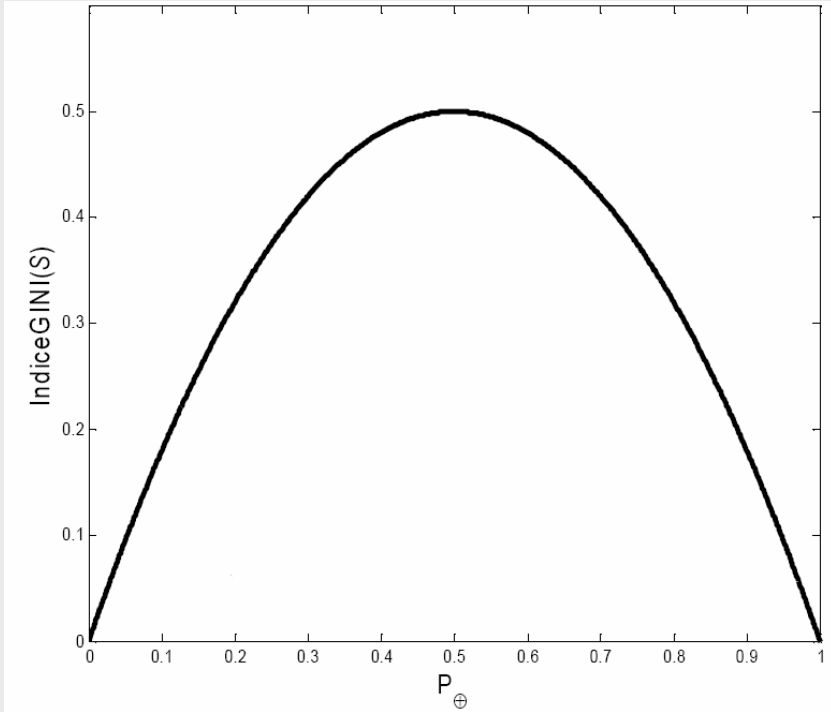
Motivación

1. CART busca minimizar el ‘resusbtitution error’
   - Puede ser interpretado como la probabilidad de equivocarse en la clasificación de una muestra.
   - En un sentido más realista se incluye el costo de cometer un error en la clasificación.
2. Riesgo será la probabilidad de caer un en nodo multiplicado por la probabilidad de error de clasificación del nodo.

CART: El Criterio de Discriminación

1. El Criterio GINI: Pretende medir el grado de impureza de un nodo

$$GINI(S) = 2 \cdot P((+) \mid S) \cdot P((-) \mid S)$$
$$= 2 \cdot P_{(+)} \cdot P_{(-)}$$



2. Para más de dos clases

$$GINI(S) = \sum_{i \neq j} P(j \mid S) P(i \mid S)$$
$$= 1 - \sum_{j} P(j \mid t)^2$$

CART: El Criterio de Discriminación

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

1. El Criterio TWOIN:
   - No es una medida de impureza (o sea no alcanza su máximocuando la impureza es máxima)
   - Es recomendada cuando existen más de dos clases en el atributo objetivo
   - En problemas con dos clases (o en la vecindad) no es posible saber cual es mejor (GINI o TWOIN)
   - La ecuación está estrictamente asociada a un árbol binario:

$$TWOIN(S) = \frac{P(left)P(right)}{4}\left[\sum_j \left|P(j \mid S_{left}) - P(i \mid S_{right})\right|\right]^2$$

CART: El Criterio de Discriminación

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

1. El Criterio de la Varianza:
   - Es utilizado para problemas de regresión.
   - También mide la impureza del nodo.
   - La elección se realiza mediante el atributo que explique mejor la varianza del atributo objetivo. Es decir el que minimice la siguiente expresión:

$$P(left)s^2(S_{left}) + P(rigth)s^2(S_{right})$$

   - Donde

$$s^2(S) = \frac{1}{N(S)} \sum_{x_n \in t} (y_n - \bar{y}(S))^2$$

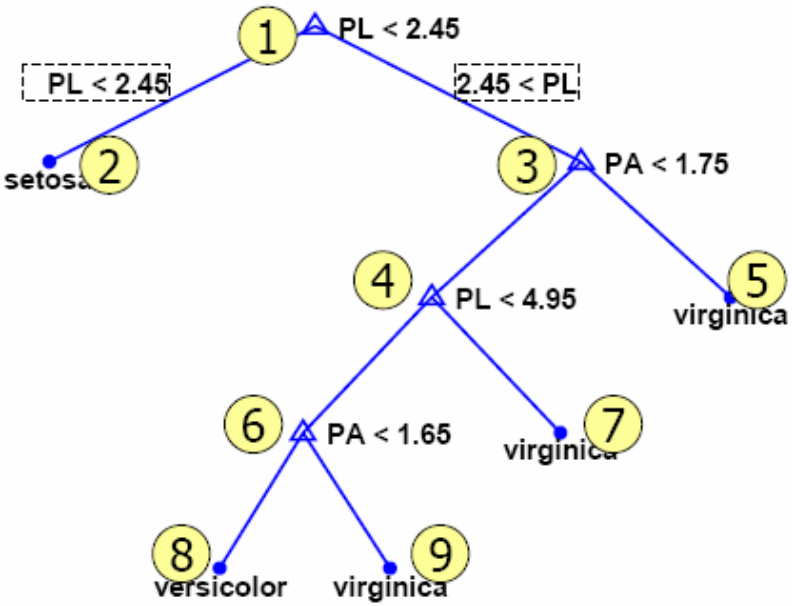
Un Ejemplo Simple

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

1. El ejemplo de los lirios de Fisher (Fisher iris).
   - Base de datos creada en 1936
   - Cada registro corresponde a una planta, hay 50 plantas de cada tipo y hay 3 tipos (Setosa, Versicolour y Virginica). Uno de los tipos es linealmente separable (análisis discriminante) de los otros dos.
   - El atributo a predecir es la clase de lirio que es.
   - Los atributos son 4:
     - Largo del sépalo (cm)
     - Ancho del sépalo (cm)
     - Largo del pétalo (cm)
     - Ancho del pétalo (cm)

Un Ejemplo Simple

1. Los Resultados



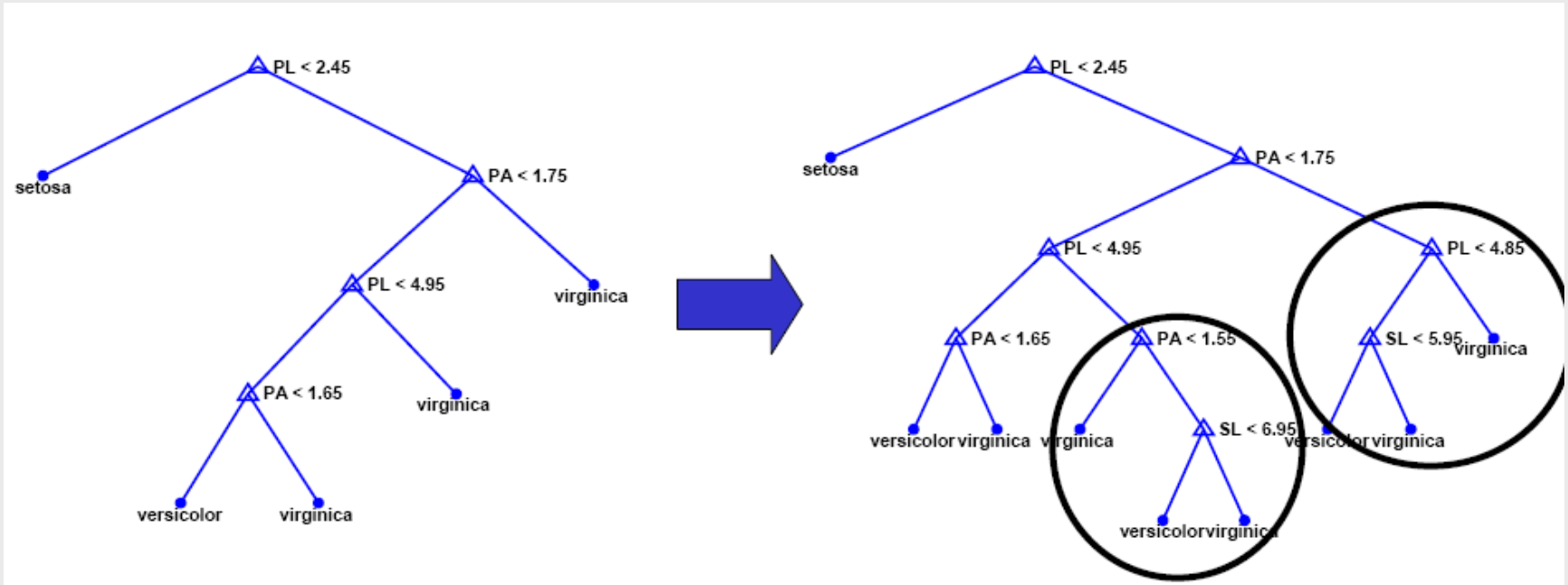
Nodo	Probabilidad de estar en el Nodo	Error de Resubstitution	Riesgo	Tamaño
1	1,000	0,667	0,667	150
2	0,333	0,000	0,000	50
3	0,667	0,500	0,333	100
4	0,360	0,093	0,033	54
5	0,307	0,022	0,007	46
6	0,320	0,021	0,007	48
7	0,040	0,333	0,013	6
8	0,313	0,000	0,000	47
9	0,007	0,000	0,000	1

Más datos:

- Número mínimo de datos para realizar una subdivisión (split min): 10
- Número de Nodos Terminales: 5

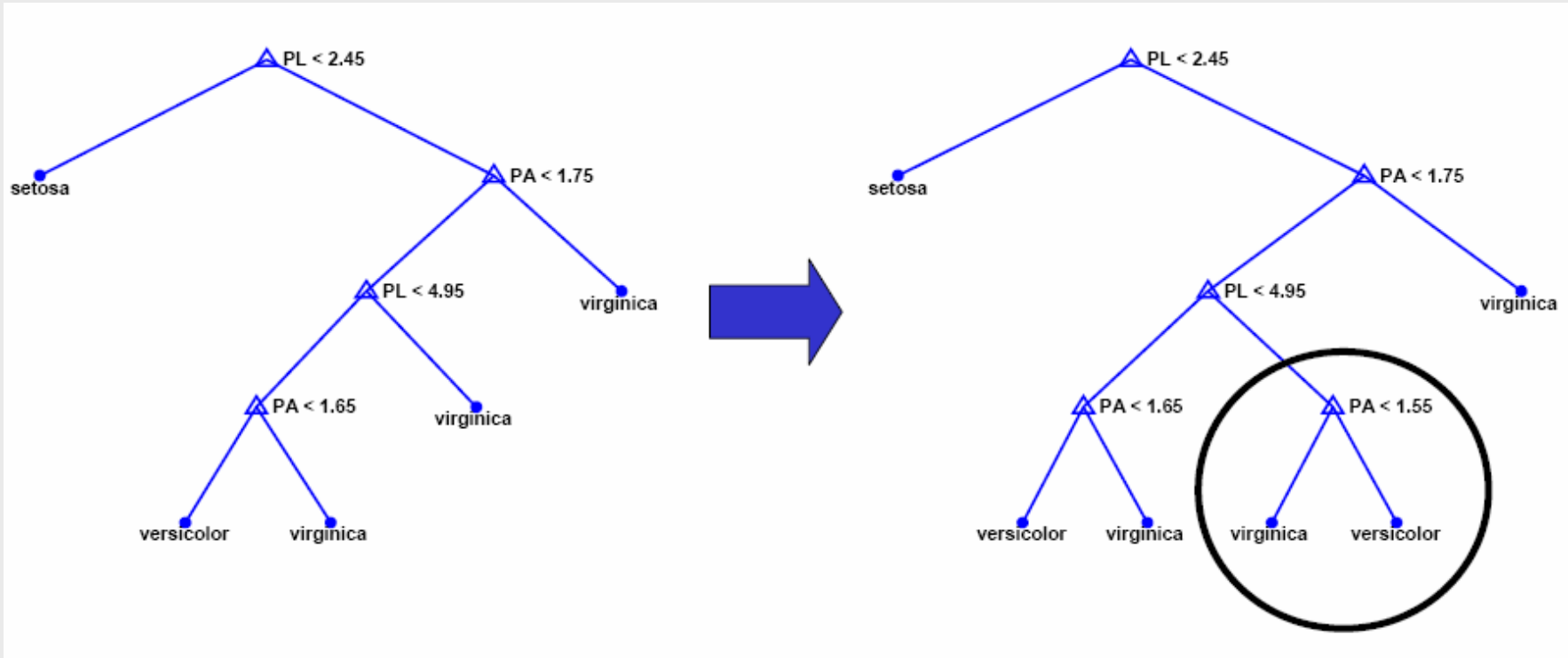
Un Ejemplo Simple

1. Que hubiese pasado si:
   - El splitmin=1



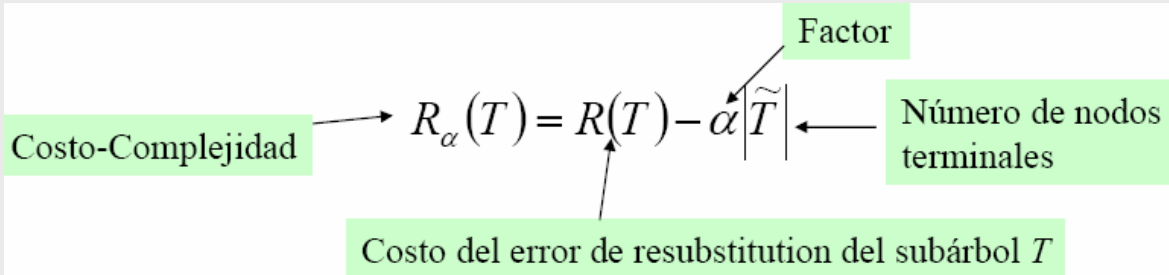
Un Ejemplo Simple

1. Que hubiese pasado si:
   - El splitmin=5

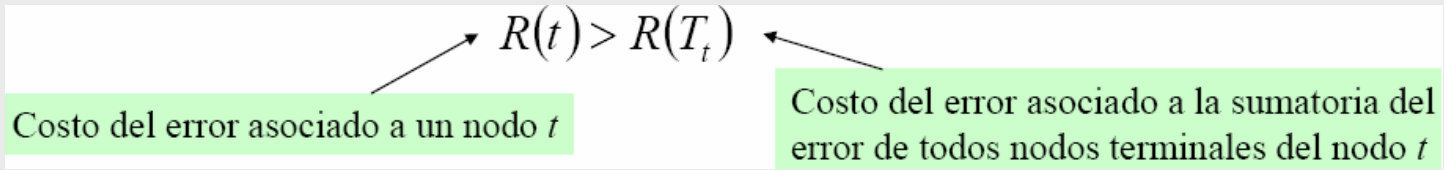


Sobreajuste

1. Tenemos una manera simple de evitar el sobre ajuste,deteniendo el crecimiento, ¿pero bajo que criterio?, ... No se puede decir a priori.
2. Otra manera: Poda costo-complejidad
   - La idea de fondo:



   - Se puede probar sobre el conjunto de entrenamiento que:



Poda Costo-Complejidad

1. Con simples cálculos se puede llegar a que

$$R_\alpha(S) > R_\alpha(T_S)$$

2. Pero a nosotros nos interesa podar, es decir quedarnos con el padre, por lo que nuestro interés está en encontrar un valor de α en que estemos indeferentes en podar o no. O sea:

$$R_\alpha(S) = R_\alpha(T_S)$$

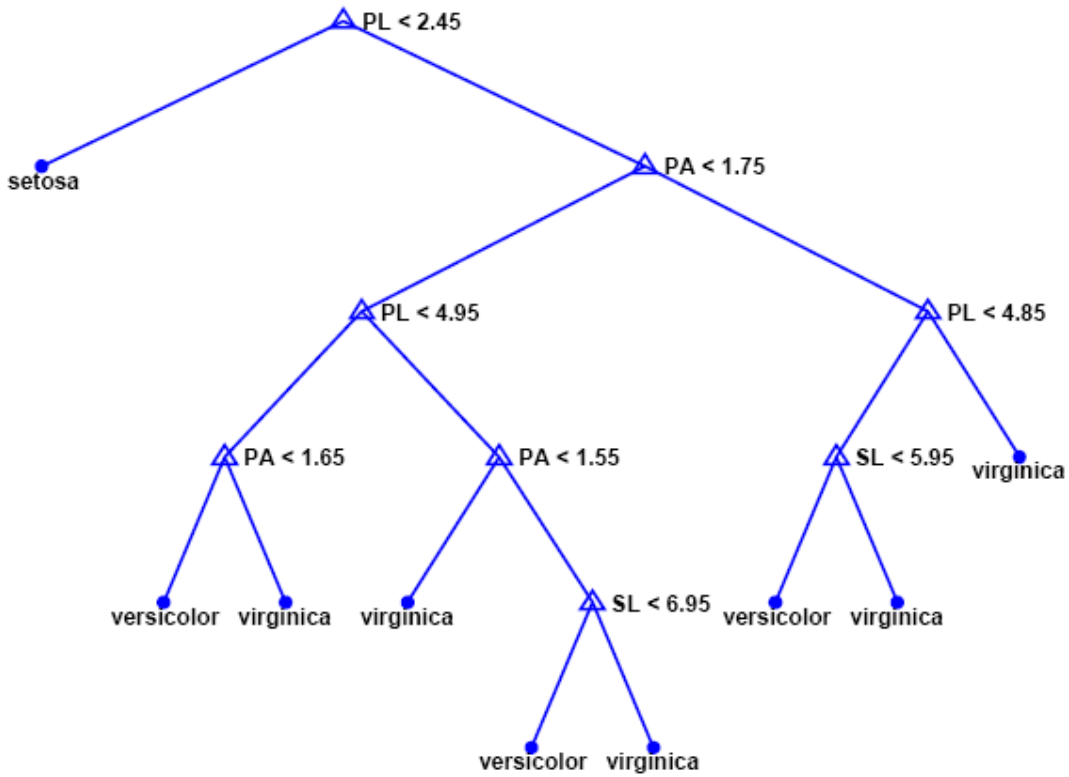
3. La idea de fondo es encontrar los enlaces más débiles del árbol, a mediad que α aumenta el conjunto de enlaces que pasan a la indiferencia aumenta

Ejemplo: Poda Costo-Complejidad

1. Nivel de poda 0

$\alpha = 0$

Número de Nodos Terminales = 9

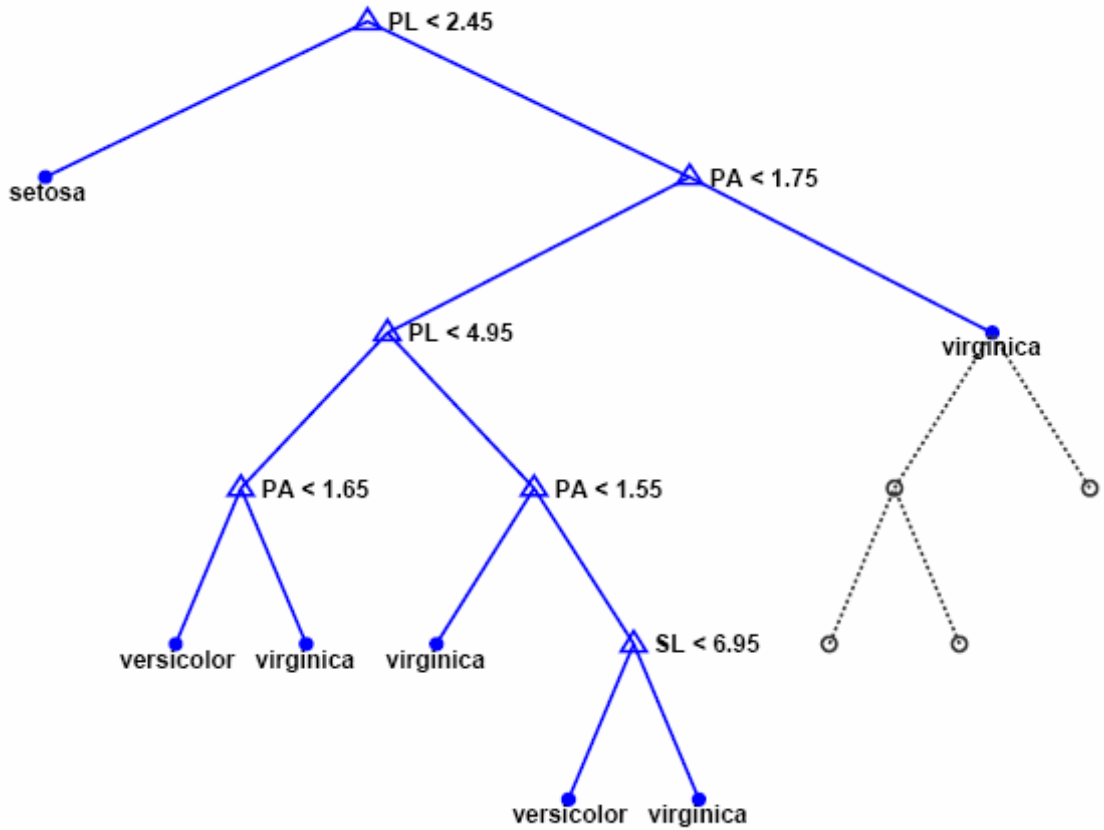


Ejemplo: Poda Costo-Complejidad

1. Nivel de poda 1

$\alpha = 0.0033$

Número de Nodos Terminales = 7

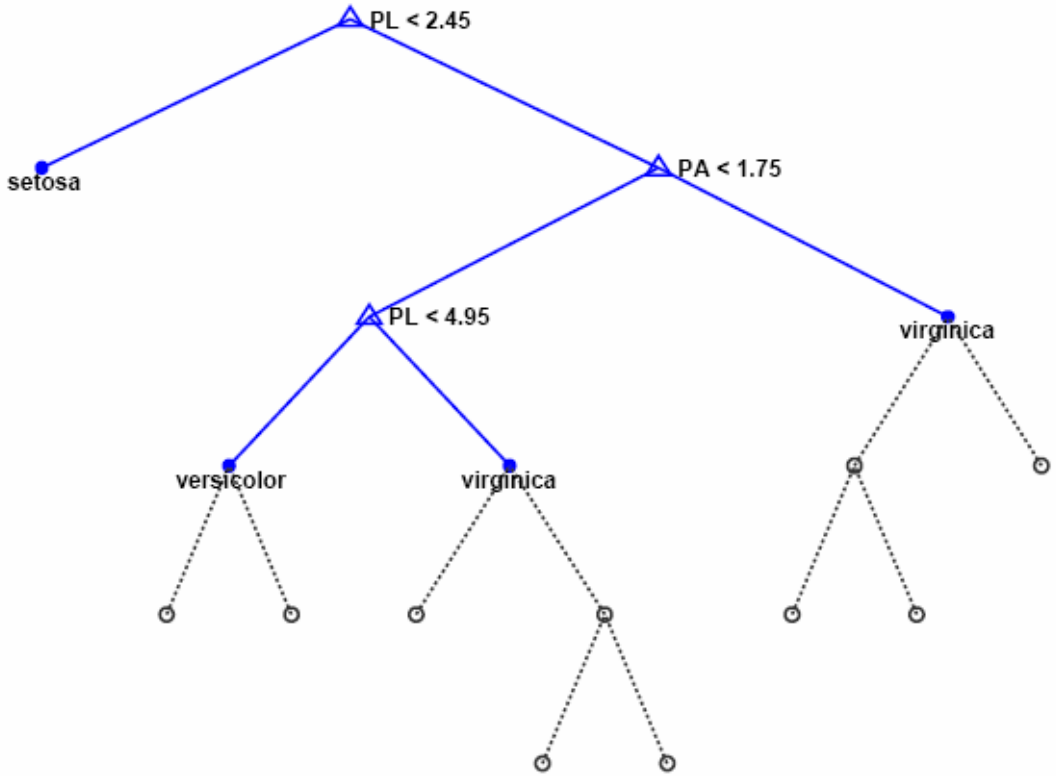


Ejemplo: Poda Costo-Complejidad

1. Nivel de poda 2

$\alpha = 0.0066$

Número de Nodos Terminales = 4

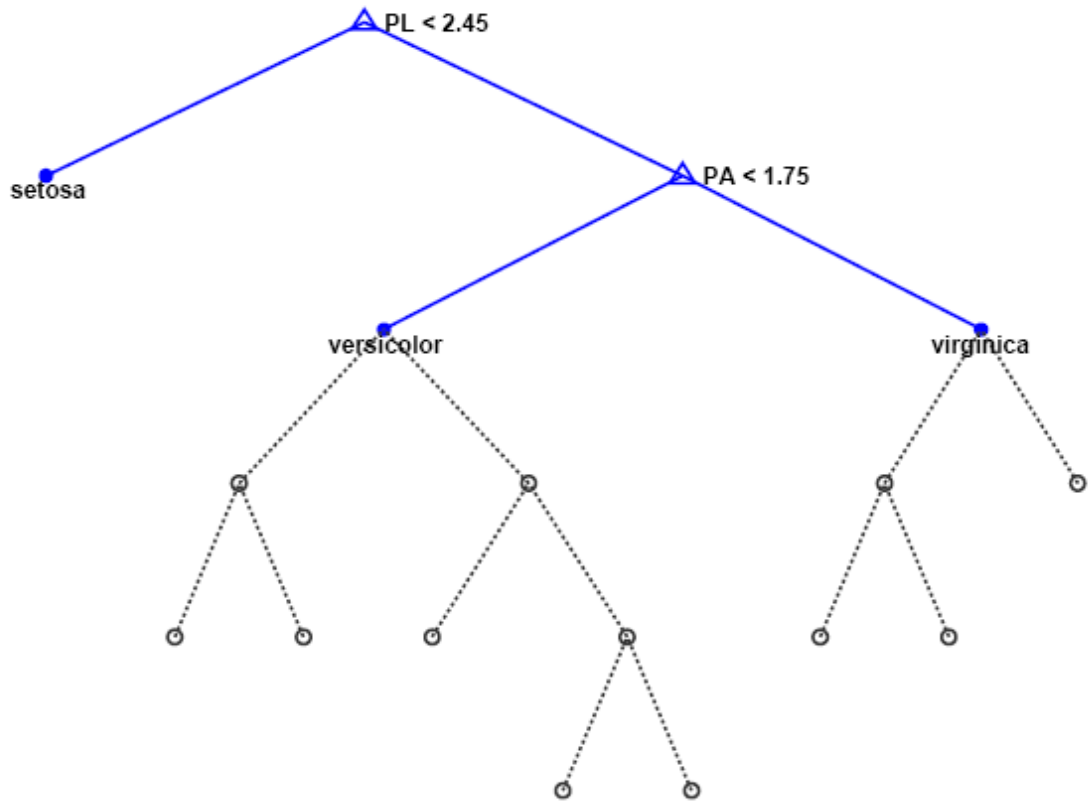


Ejemplo: Poda Costo-Complejidad

1. Nivel de poda 3

$\alpha = 0.0133$

Número de Nodos Terminales = 3

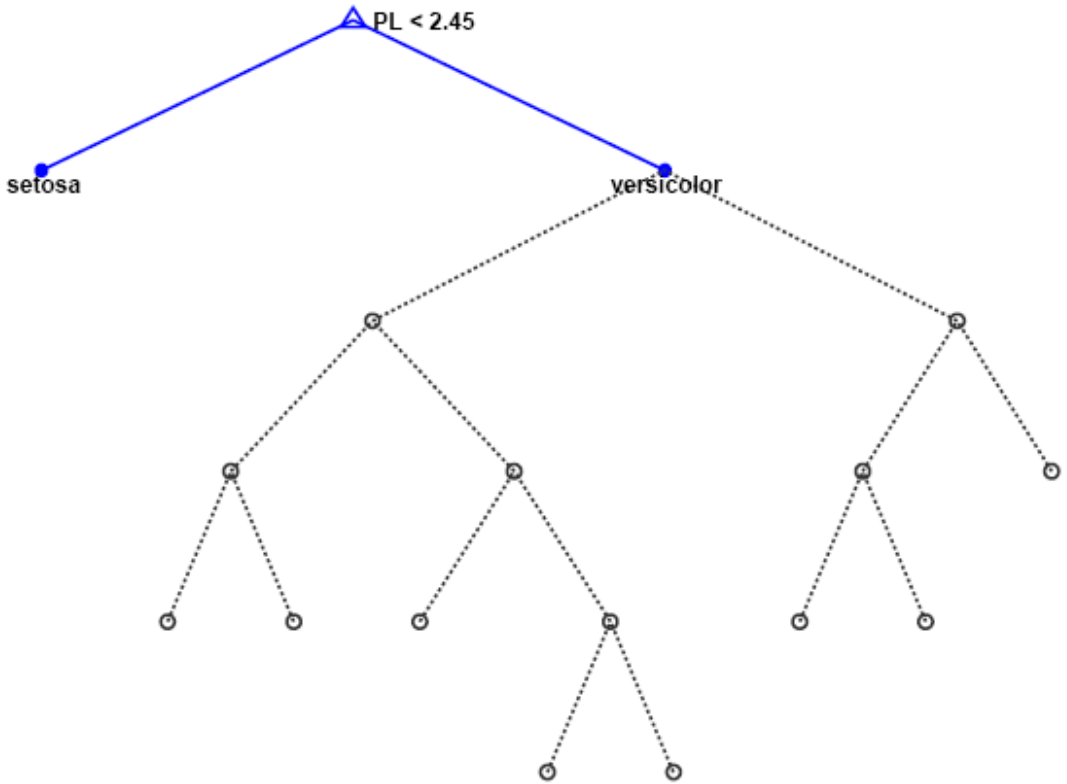


Ejemplo: Poda Costo-Complejidad

1. Nivel de poda 4

$\alpha = 0.2933$

Número de Nodos Terminales = 2

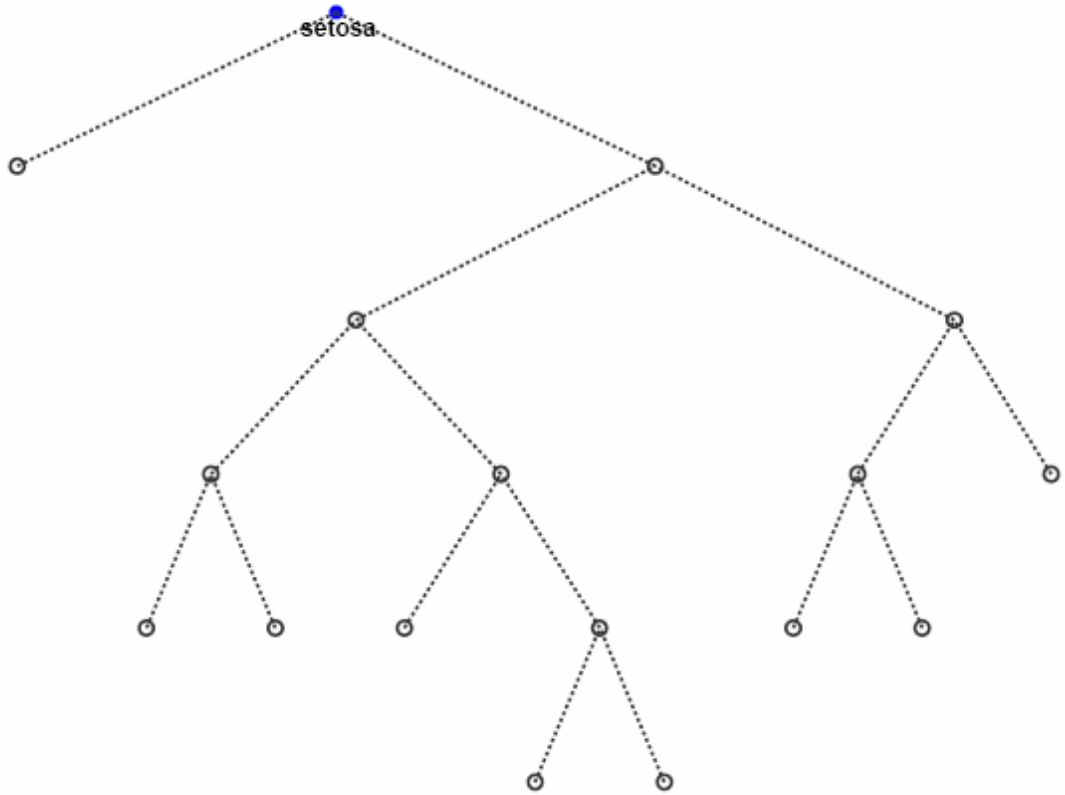


Ejemplo: Poda Costo-Complejidad

1. Nivel de poda 5

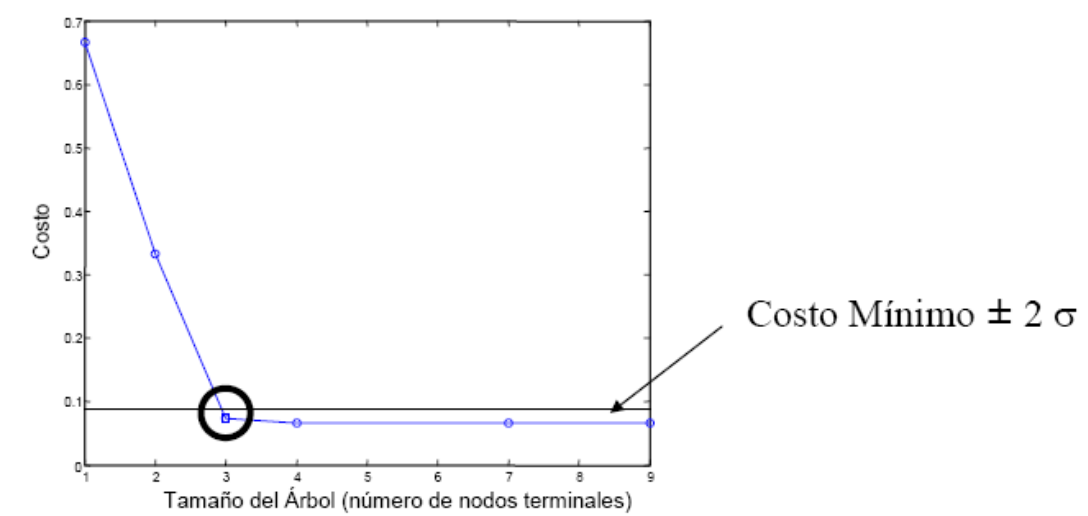
$\alpha = 0.3333$

Número de Nodos Terminales = 1



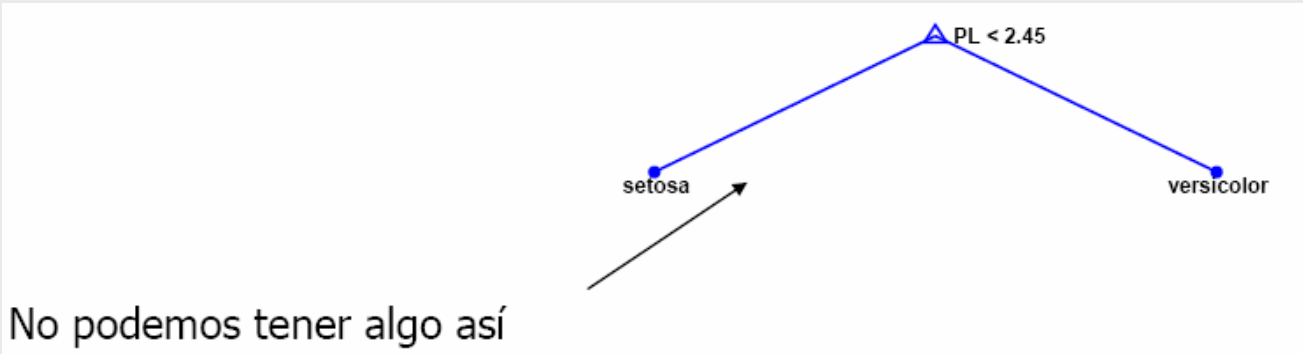
¿No será mucha poda?

1. El Nivel Óptimo de Poda, ¿Cómo determinarlo?
   - Utilizar el error de resubstitution ⇨ No sirve.
   - Utilizar un conjunto de prueba, para determinar su valor óptimo ⇨ es bueno pero se pierden muchos datos.
   - Utilizar validación cruzada ⇨ Es buena y no se pierden muchos
2. datos.



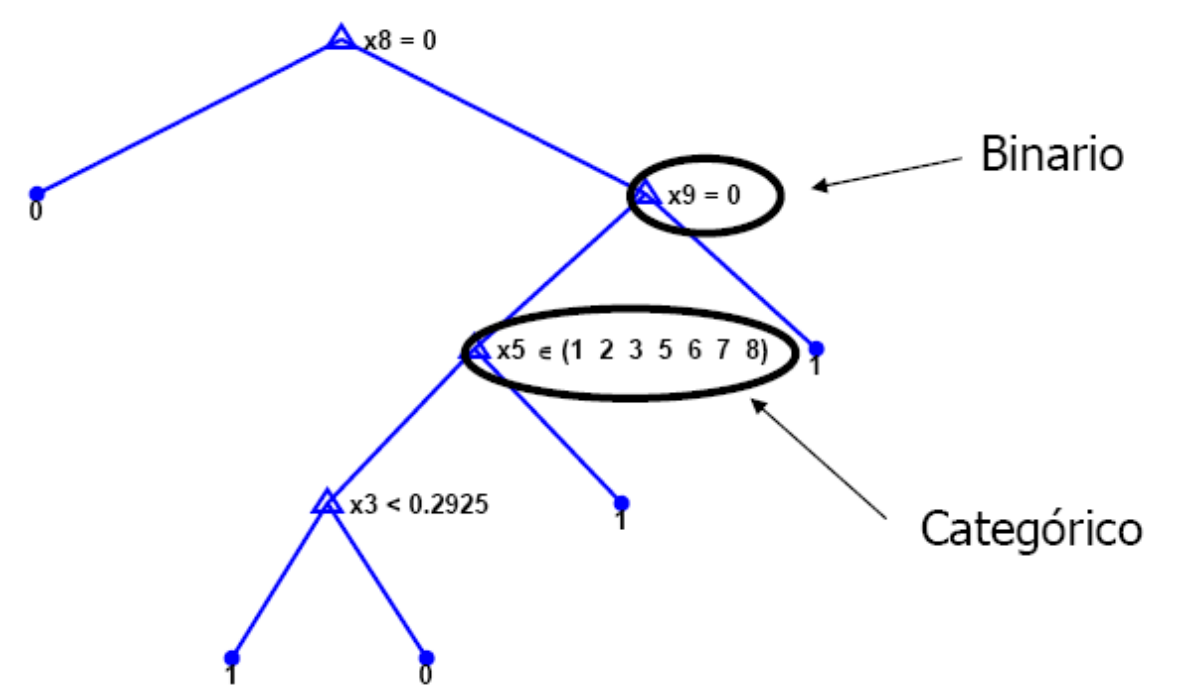
Otro Ejemplo

1. Base de Datos para la Asignación de créditos Bancarios.
2. 690 instancias 383 muestras de clase 0 y 307 de clase 1.
3. 6 atributos numéricos y 8 categóricos, más una clase, que es si es que se le asigna crédito o no.
4. Acá tenemos atributos categóricos también, que pasa si no siguen un orden



La solución

1. Agruparlos en subconjuntos que minimicen el valor de criterio de división que estemos utilizando , ..., pero eso se puede demorar mucho ($2n-1$ Conjuntos), por suerte los criterios de división de CART hacen que sea necesario revisar sólo $2n$ conjuntos.

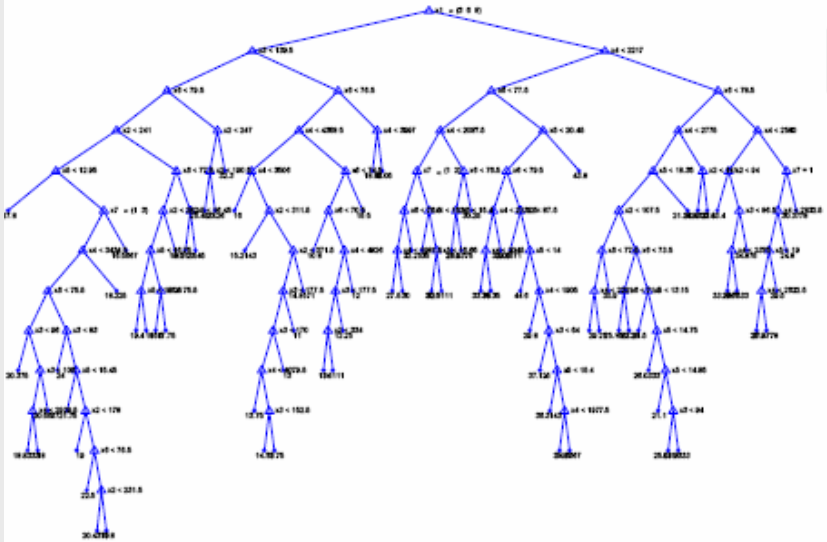


Otro Ejemplo

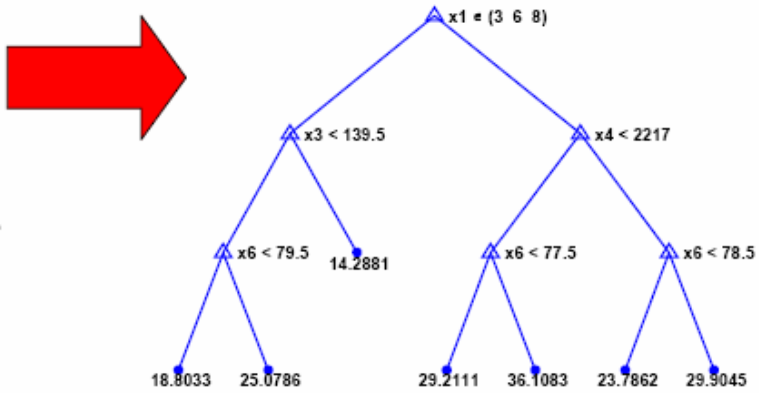
1. El siempre amigo MPG
2. 7 atributos más el valor de las millas por galón a obtener.
3. Nuevo problema las MPG es una clasificación si no un valor continuo.
4. Cómo se puede solucionar…..
5. Con Árboles de Regresión, probemos

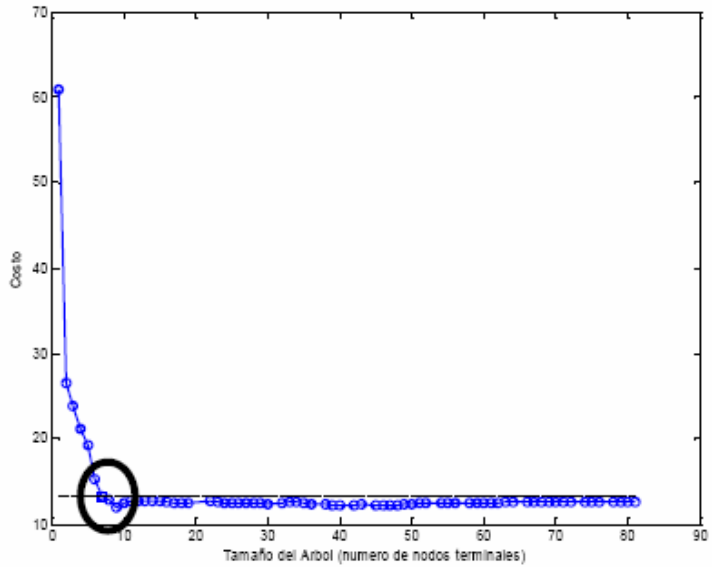
Otro Ejemplo

Sin podar

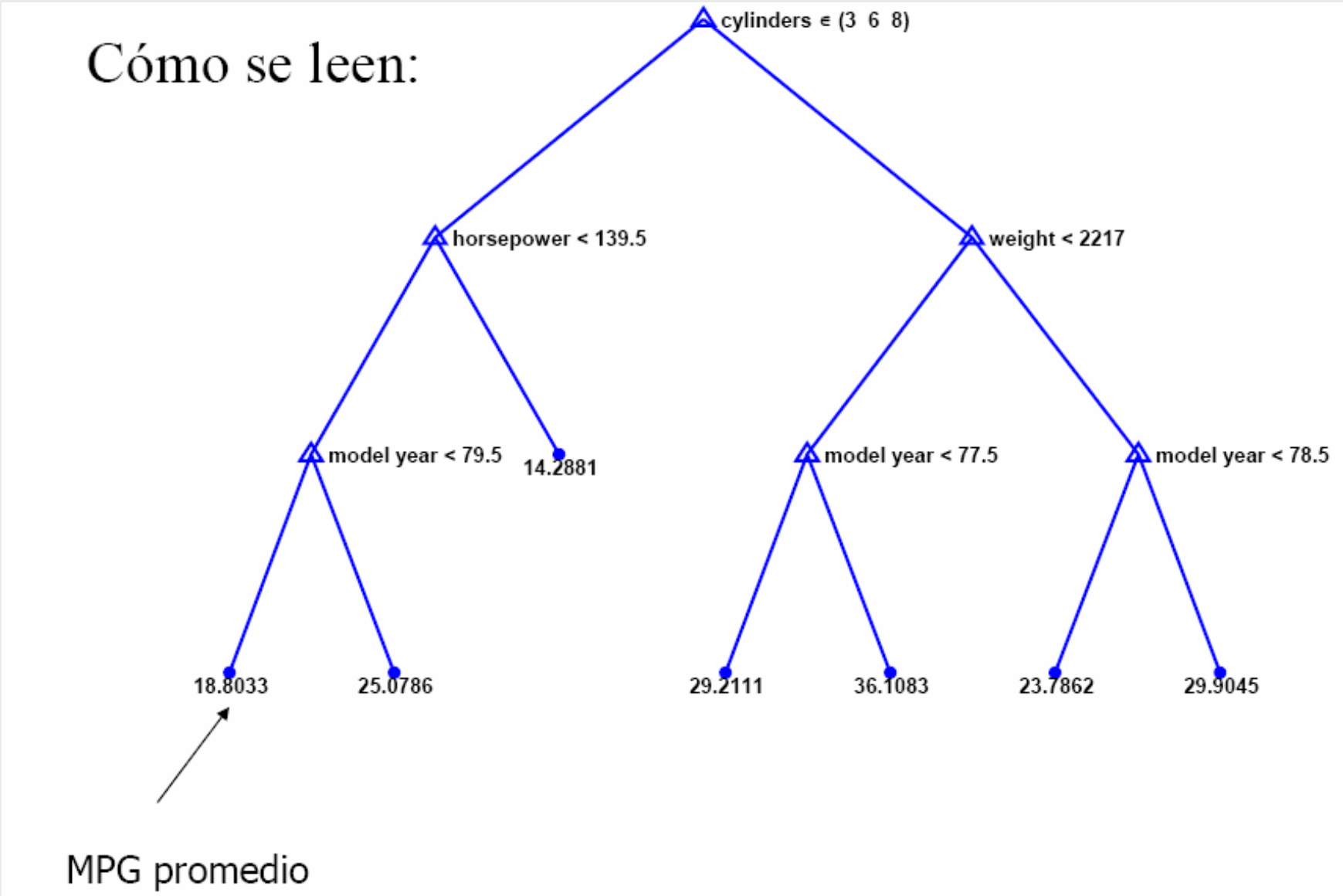


Con podar



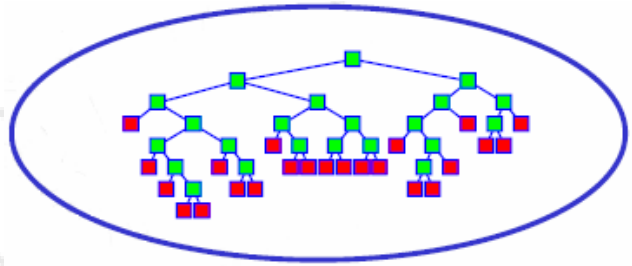


Árboles de Regresión



Comentarios

1. ¿Por que la poda es importante?
   - Distintos criterios de crecimiento de árbol combinado con el mismo criterio de poda entregan resultados similares en acierto
   - Distintos criterios de poda entregan resultados distintos bajo iguales criterios de crecimiento de árboles.
2. ¿Por qué CART es importante?
   - Es lejos el método de construcción de árboles de decisión más utilizado. Ej., CART 5.0, SPSS Answertree, SAS Enterprice Miner, Matlab 6.5, etc..
   - También es uno de los menos difundidos en las publicaciones, debido a que el código es cerrado.



Árboles de decisión

CART – Clasification and Regression Tree

FRANCISCO CISTERNAS

FABIÁN MEDEL

Departamento de Ingeniería Industrial

Universidad de Chile