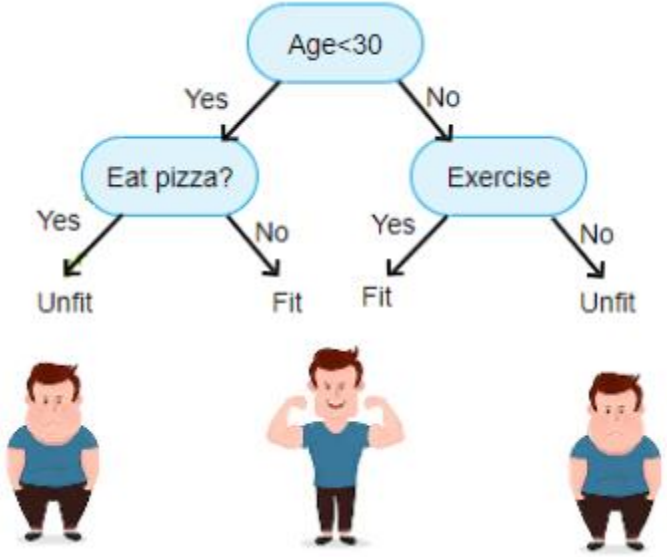
CC 442

Minería de Datos



Árboles de Decisión

Aprendizaje esperado

El alumno aprenderá y entenderá que los árboles de decisión son una herramienta visual que ayudan a personas y empresas a tomar decisiones al visualizar posibles resultados y consecuencias. Permiten a los usuarios sopesar diferentes oportunidades y trazar un camino hacia el resultado deseado.

Contenido

Árboles de decisión

1. Árboles de decisión
2. Selección de modelos
3. Criterios de selección
4. Ganancia de información
5. Algoritmo ID 3
6. Construcción
7. Variables continuas
8. Poda (Prunning)
9. Árboles de regresión
10. Selección de modelos

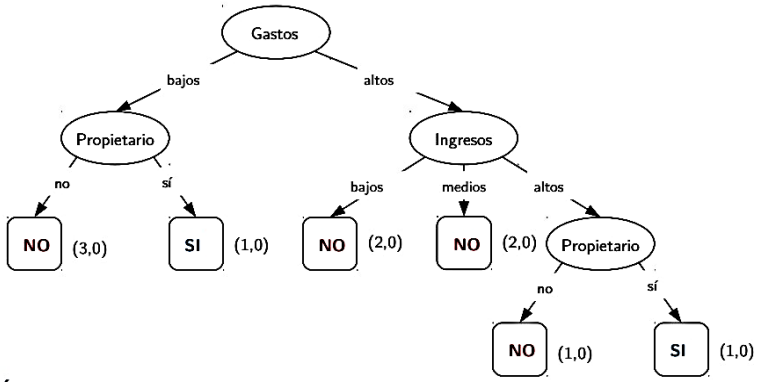
Árboles de decisión

Un **Árbol de decisión** representa la función de hipótesis mediante grafo dirigido tal que:

- ✔ Cada nodo representa una variable de entrada.
- ✔ Cada rama que pende de un nodo representa uno de los posibles valores que puede tomar la variable correspondiente.
- ✔ Las hojas corresponden con valores de la variable de clase.
- ✔ Un ejemplo se clasifica recorriendo el árbol desde la raíz y eligiendo en cada momento la rama que satisface la condición para el valor del atributo correspondiente. La clase elegida sería la asignada a la hoja a la que se llega.

Ejemplo:

ingresos	propietario	gastos	crédito
bajos	no	altos	**NO**
bajos	si	altos	**NO**
medios	si	altos	**NO**
medios	no	altos	**NO**
altos	no	altos	**NO**
altos	si	altos	**SI**
bajos	no	bajos	**NO**
medios	no	bajos	**NO**
altos	no	bajos	**NO**
medios	si	bajos	**SI**

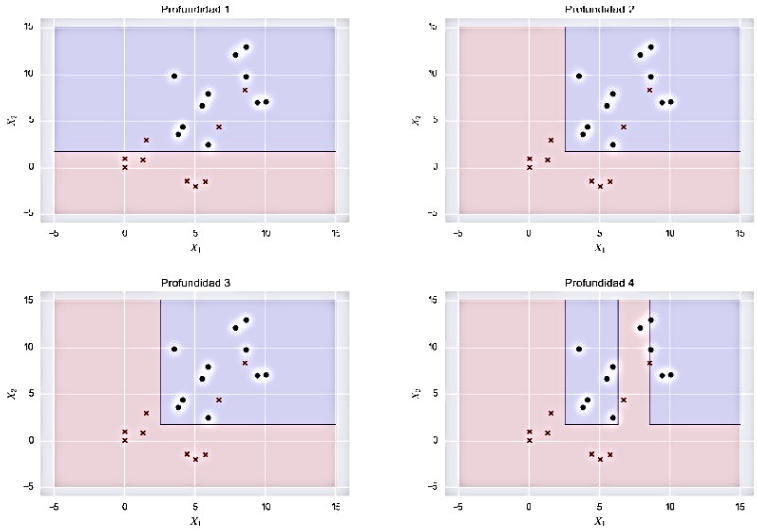


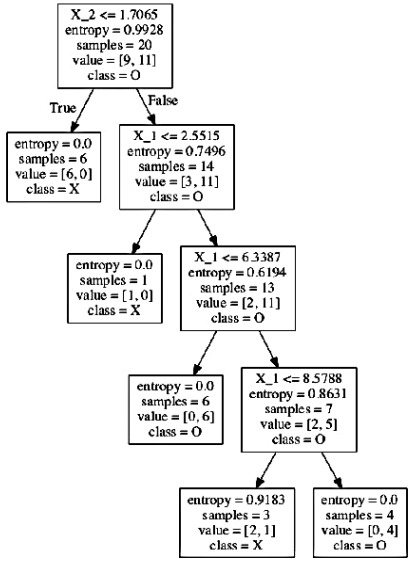
Árboles de decisión

Se construyen particionando el espacio de entrada de manera recursiva. En cada paso se elige la variable que produce la partición óptima. La partición se representa en un árbol.

Un caso de entrada se procesa recorriendo el árbol, eligiendo en cada nodo el correspondiente al valor de la variable, y asignando el valor correspondiente a la hoja alcanzada.

Ejemplo con profundidad máxima 4





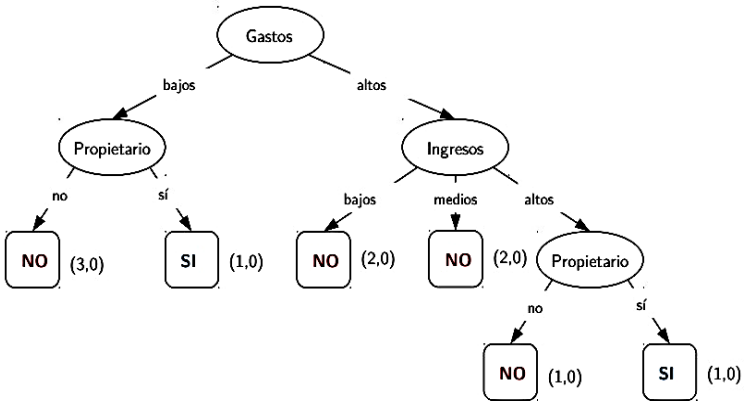
Árboles de decisión

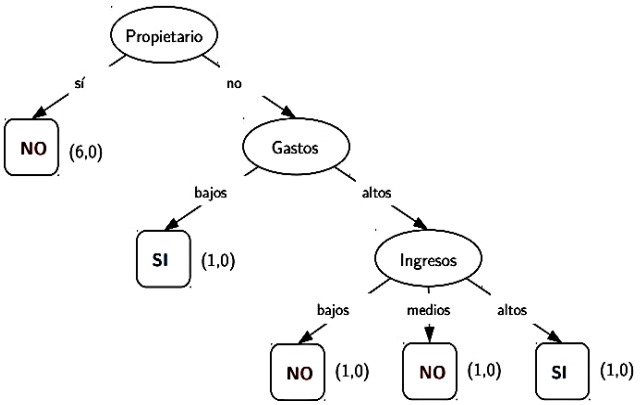
Algoritmo basado en divide y vencerás

- ✔ La construcción se realiza, generalmente, de forma voraz, siguiendo dicho esquema de particionamiento.
- ✔ **TDIDT**: *Top-Down Induction of Decision Trees*
  - ✗ Algoritmo voraz que construye un árbol de decisión aplicando recursivamente y de forma top-down un procedimiento de divide y vencerás.
- ✔ Básicamente:
  - ✗ Un árbol de decisión se construye recursivamente desde la raíz.
  - ✗ El procedimiento recibe un conjunto de datos para crear un nodo.
    - ✓ Si para dichos datos la variable clase tiene muy poca variabilidad o son todos iguales, el proceso se detiene creando una variable hoja.
    - ✓ En otro caso, usando la información contenida en los datos, se selecciona una variable como criterio de decisión.
  - ✗ Usando la variable seleccionada, el conjunto de datos se particiona en una serie de subconjuntos y se invoca el procedimiento recursivamente para cada uno de ellos.

Árboles de decisión: selección de modelos

¿Qué árbol es mejor?





Árboles de decisión: criterios de selección

El objetivo es seleccionar una variable que conduzca a una partición más **pura**/**menos ruidosa** con respecto a la variable objetivo.

- ✔ **Entropía** (ID3, C4.5).

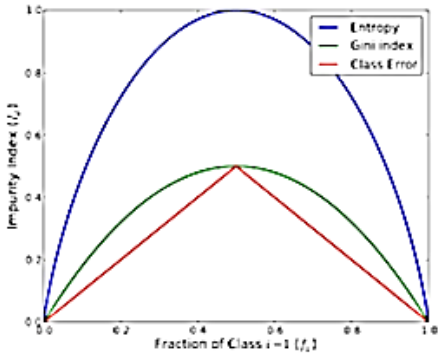
$$H(X) = -\sum_{i=1}^{L} P_i \cdot \log_2(P_i)$$

- ✔ **Indice Gini** (CART).

$$Gini(X) = 1 - \sum_{i=1}^{L} P_i^2$$

- ✔ **Error de clasificación**.

$$Err(X) = 1 - max(P_1, \ldots, P_L)$$



Árboles de decisión: ganancia de información

Ganancia

Para cada variable candidata, calculamos la ganancia con respecto al criterio elegido. Seleccionamos la que maximiza la ganancia.

- ✔ $\mathbf{D}$ se divide en $\mathbf{D}_X^1, \ldots, \mathbf{D}_X^r$ usando la variable X.
- ✔ Calculamos el criterio correspondiente (H, Gini, Err) para la variable clase en el conjunto inicial $C(\mathbf{D})$.
- ✔ Calculamos para todos los subconjuntos resultantes de la partición: $C(\mathbf{D}_X^1), \ldots, C(\mathbf{D}_X^r)$.
- ✔ Calculamos para la partición promediando:

$$C(\mathbf{D}_X^1, \ldots, \mathbf{D}_X^r) = \sum_{i=1}^{r} \frac{|\mathbf{D}_X^i|}{|\mathbf{D}|} \cdot C(\mathbf{D}_X^i)$$

- ✔ Y obtenemos la ganancia:

$$gain(X) = C(\mathbf{D}) - C(\mathbf{D}_X^1, \ldots, \mathbf{D}_X^r)$$

Algoritmo ID 3

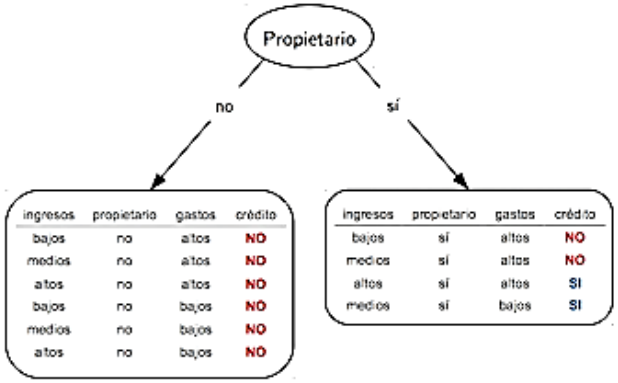
Algoritmo 1 ID3 (Algoritmo de Hunt utilizando la ganancia de información)

A: Lista de atributos
Y: Variable de clase
D: Partición de la base de datos correspondiente a la rama (o raíz)
procedimiento CONSTRUIRÁRBOL(A, Y, D)
    # Si toda la base de datos tiene la misma clase, crea una hoja
    si $y^{(i)} = c_k \quad \forall (x^{(i)}, y^{(i)}) \in D$ **entonces**
        devuelve CREARHOJA(c_k)
    fin si
    # Si no quedan atributos, se devuelve la clase mayoritaria
    si $A = \emptyset$ **entonces**
        $c_k \leftarrow$CLASEMAYORITARIA(D)
        devuelve CREARHOJA(c_k)
    fin si
    # Si no es una hoja, se crea un nodo normal.
    $X_{max} \leftarrow$ MaxGain(A, D) # Selecciona el atributo con mayor ganancia de información.
    Nodo $\leftarrow$ CREARNODO(X_{max}) # Crea un nodo con la variable seleccionada
    para x_l posible valor de X_{max} **hacer**
        AÑADIRRAMA(*nodo*, x_l)
        $D_l \leftarrow \{(x^{(i)}, y^{(i)}) \in D | X^{(i)}_{max} = x_l\}$
        CONSTRUIRÁRBOL($A - X_{max}, Y, D_l$)
    fin para
    devuelve *Nodo*
fin procedimiento

Árboles de decisión: construcción

Construcción de árboles de clasificación: Ejemplo

ingresos	propietario	gastos	crédito
bajos	no	altos	NO
bajos	si	altos	NO
medios	si	altos	NO
medios	no	altos	NO
altos	no	altos	NO
altos	si	altos	SI
bajos	no	bajos	NO
medios	no	bajos	NO
altos	no	bajos	NO
medios	si	bajos	SI



$H(crédito) = 0.72$

Gain(propietario) = 0.32

$Gain(ingresos) = 0.12$

$Gain(gastos) = 0.01$

(propietario = no) $\Longrightarrow$ (crédito = NO)

(propietario = sí) $\Longrightarrow$ (crédito = ?)

Se puede expandir la rama recursivamente

Árboles de decisión: variables continuas

- Las variables discretas solo aparecen una sola vez en el camino desde la raíz.
- Las variables continuas se deben ir particionando en intervalos a diferentes profundidades del árbol.

Ejemplo

Teoría	Prácticas	Aprobado
2.5	regulares	no
3	malas	no
4	regulares	no
5	malas	no
5	buenas	si
6	regulares	si
7.5	buenas	si
7.5	malas	no
9	buenas	si
9.5	regulares	si

$umbral = 3$

$\Longrightarrow$

Teoría	Prácticas	Aprobado
bajo	regulares	no
bajo	malas	no
alto	regulares	no
alto	malas	no
alto	buenas	si
alto	regulares	si
alto	buenas	si
alto	malas	no
alto	buenas	si
alto	regulares	si

La entropía[1] condicionada al atributo teoría, utilizando $umbral = 3$:

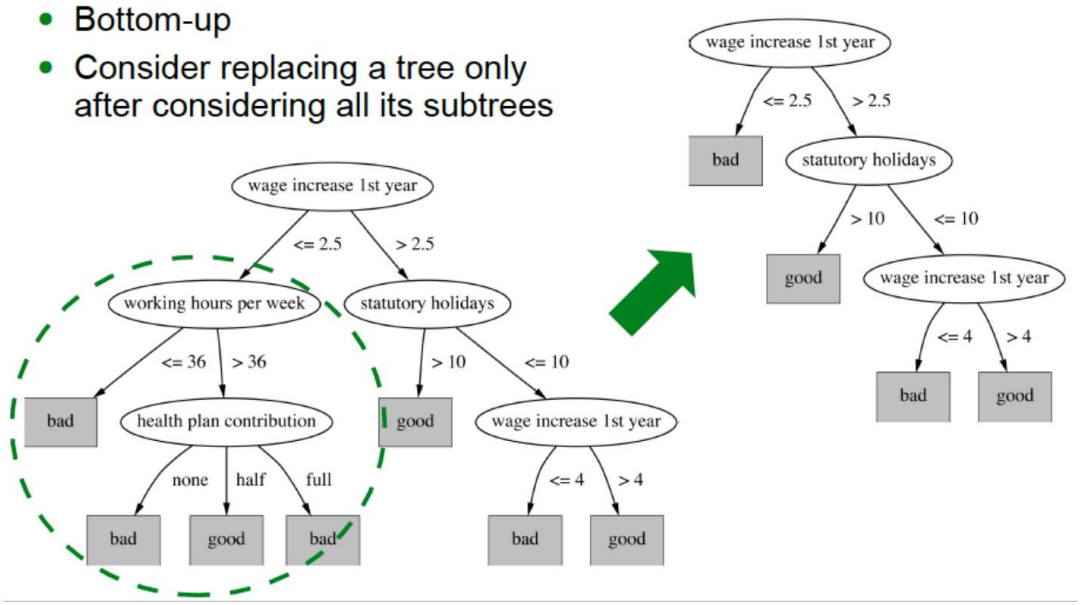
$$H(Aprobado|teoria\,[umbral = 3]) = \frac{2}{10} \cdot H(2,0) + \frac{8}{10} \cdot H(3,5) = 0.76$$

Árboles de decisión: Poda

Poda en c4.5

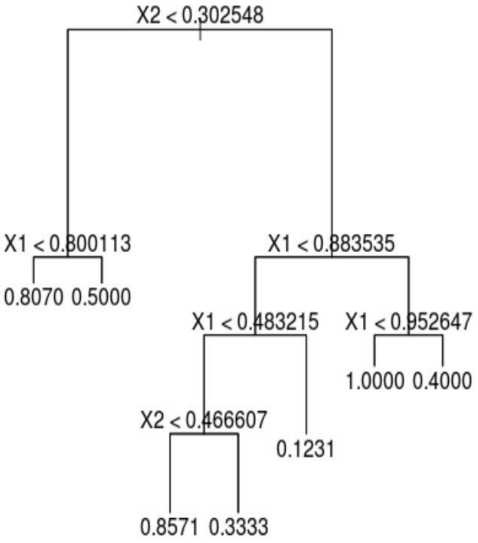
En c4.5 hay dos operaciones de poda:

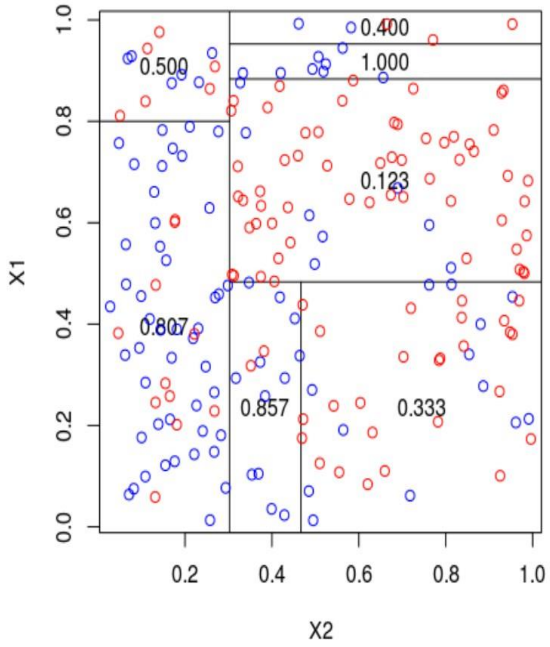
- ✔ **Subtree-replacement**: un subárbol se sustituye por una hoja.
- ✔ **Subtree-raising**: un subárbol sustituye a otro.



Árboles de regresión

- La estructura es idéntica a un árbol de decisión, pero ahora las hojas contienen un valor **numérico**, normalmente la media para la variable objetivo en todos los registros que llegan al hipercubo correspondiente.





Selección de modelos

- ✔ Dividiremos el espacio de soluciones en un conjunto de regiones $\{R_1, R_2, \ldots, R_m\}$, distintas y sin solapamiento t.q. para todo ejemplo que caiga en una región devolveremos el mismo valor, la media de los ejemplos del conjunto de training que caen en dicha región: $\bar{y}_{R_j}$.
- ✔ El objetivo es encontrar el conjunto de regiones $\{R_1, R_2, \ldots, R_m\}$ que minimice:

$$\sum_{j=1}^{m} \sum_{i \in R_j} (y^{(i)} - \bar{y}_{R_j})^2.$$

- ✔ **Diferencia** respecto a clasificación: **criterio para elegir la variable** y el umbral en cada nodo.
  - ✗ Dado un nodo concreto a ramificar, se eligirá la variable X_j y el umbral t en su dominio, t.q. se minimice la siguiente expresión (RSS):

$$\sum_{i:x_j^{(i)} < t} (y^{(i)} - \bar{y}_{R_{<t}})^2 + \sum_{i:x_j^{(i)} \geq t} (y^{(i)} - \bar{y}_{R_{\geq t}})^2$$

  - ✗ Este es el criterio usado en CART (Breiman). Computacionalmente el proceso es similar al de tratar atributos numéricos en C4.5
  - ✗ Otra opción es minimizar la varianza o desviación estándar resultante:

Logros

Al término de la sesión el alumno tendrá una visión global y clara respecto al objetivo final de un árbol de decisión que es minimizar la impureza de Gini de los nodos de las hojas , asegurando que cada nodo represente una categoría única para una clasificación precisa, que se alinea con el principio de máxima parsimonia y apunta a la mayor utilidad

.

Conclusiones

Los árboles de decisiones son modelos más complejos que los controladores unidireccionales y bidireccionales. Amplían la secuencia como los modelos de combinación. La principal diferencia es que los árboles de decisiones admiten el descubrimiento de la interacción entre varios predictores y, por lo tanto, conocimientos más profundos que los controladores.

Referencias

1. Ian h: Witten, Eibe Frank, Mark A. Hall (2011). Data Mining. Third Edition. Chapter 4.
2. Gorunescu Florin (2017). Data Mining: Concepts, Models and Terchniques. Intellligent Systems Reference Library Volume12. Chapter 4.
3. Aurélien Géron (2019). Hands-On Machine Learning with Scikit-Learn and TensorFlow. 2nd Edition. Chapter 6.