

# Árboles de Regresión

(*Regression Trees*)

Fernando Arias-Rodríguez

Banco Central de Bolivia

30 de agosto de 2024



- ① Procedimientos de selección automática de variables
- ② Estructura del modelo
- ③ Ventajas y desventajas
- ④ Inferencia

# 1 Procedimientos de selección automática de variables

## Introducción

## 2 Estructura del modelo

## 3 Ventajas y desventajas

## 4 Inferencia

- 1 Procedimientos de selección automática de variables  
Introducción
- 2 Estructura del modelo
- 3 Ventajas y desventajas
- 4 Inferencia

- Este tipo de metodologías son no paramétricas, lo que implica que se sabe muy poco del proceso generador de datos.
- Su lógica es similar a la de árboles de decisión, donde una estructura jerárquica le permite al usuario elaborar una serie de decisiones secuenciales para llegar a la solución.
- Así, las decisiones pueden asociarse con los insumos  $x$  y la solución como el resultado  $y$ .
- Hay dos tipos de árboles: clasificación y regresión (de ahí que se hable de *Classification and Regression Trees* - CART). Ambos tienen una estructura similar y solo se diferencian en la naturaleza de la variable solución,  $y$ .
- Solo se abordan los árboles de regresión.

Los árboles de regresión han ganado popularidad por:

- su enfoque basado en reglas de decisión,  $x'$ s, para llegar a un resultado.
- La estructura de un árbol suele ser fácil de entender, interpretar y visualizar.
- En casi todos los casos, los datos no necesitan ser largamente preprocesados para implementar la metodología.

## 1 Procedimientos de selección automática de variables

## 2 Estructura del modelo

*Recursive binary splitting*

Estructurando un árbol

Ejemplo conceptual de un árbol

## 3 Ventajas y desventajas

## 4 Inferencia

- El proceso de elaborar un árbol arranca con particionar recursivamente el espacio de los predictores  $X_i$  para predecir  $y_i$ .
- Como resultado, se tienen  $J$  regiones  $R_1, R_2, \dots, R_J$  distintas y que no se superponen.
- Para todas las observaciones que caen en la región  $R_j$ , la misma predicción aplica para todas. En el caso de árboles de regresión, esta no es sino la media de  $y$  para dicha región.
- El proceso de partición se realiza a partir de un acercamiento conocido como partición binaria recursiva o *recursive binary splitting*.



## 1 Procedimientos de selección automática de variables

## 2 Estructura del modelo

*Recursive binary splitting*

Estructurando un árbol

Ejemplo conceptual de un árbol

## 3 Ventajas y desventajas

## 4 Inferencia

# Recursive binary splitting

- 1 Se selecciona una variable  $X_j$  y un valor  $s$ , correspondiente a un umbral que determinará la partición.
- 2 Usando el valor  $s$ , se parte el espacio del regresor en **dos** regiones:  $\{X|X_j < s\}$  y  $\{X|X_j \geq s\}$ .
- 3 Se calcula un error, medido como la distancia entre los valores observados y las predicciones de  $y$  hechas por el modelo con esta partición.
- 4 Los anteriores pasos se repiten hasta que se consiga la reducción más grande posible en el error.
- 5 Los pasos 1 a 4 se repiten para todos los posibles umbrales y variables involucradas en el ejercicio hasta que se llegue a algún criterio de interrupción del algoritmo.

## Recursive binary splitting

- Un criterio de interrupción puede ser, por ejemplo, que haya siempre un mínimo de observaciones en cada  $R_j$ .
- En el caso de árboles de regresión, el criterio que se utiliza es la suma de residuos al cuadrado (RSS), calculado de la siguiente manera:

$$RSS = \sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \quad (1)$$

donde  $\hat{y}_{R_j}$  es la media de  $y$  en la región  $R_j$ .

- Minimizar RSS implica que todas las desviaciones cuadráticas para una región dada se minimizan a lo largo de todas las regiones.

## 1 Procedimientos de selección automática de variables

## 2 Estructura del modelo

*Recursive binary splitting*

Estructurando un árbol

Ejemplo conceptual de un árbol

## 3 Ventajas y desventajas

## 4 Inferencia

A partir del procedimiento descrito atrás, el proceso de estructurar un árbol es, como sigue:

- 1 Se parte de la muestra completa de datos. Se dividen variables  $X_j$  en el umbral  $s$  para obtener las primeras dos regiones:

$$R_1(j, s) = \{X | X_j \leq s\} \quad R_2(j, s) = \{X | X_j > s\} \quad (2)$$

- 2 Se evalúa la siguiente expresión para determinar si la variable dividida  $X_j$  y el umbral  $s$  minimizan el error:

$$\min_{j,s} \left[ \sum_{x_i \in R_1(j,s)} (y_i - \bar{y}_{R_1})^2 + \sum_{x_i \in R_2(j,s)} (y_i - \bar{y}_{R_2})^2 \right] \quad (3)$$

donde  $\bar{y}_{R_m}$  es el promedio de  $y$  en la región  $m$ .

- ③ Una vez se encuentra la primera división, se repite el mismo proceso en las dos regiones resultantes ( $R_1$  y  $R_2$ ).
- ④ Parar si hay un número mínimo de observaciones en una región. Por ejemplo, ninguna región debe tener menos de 5 observaciones.

Nótese que, para un número grande de variables, el proceso antes descrito producirá un modelo sobreajustado.

Para un árbol, una medida que capture tanto precisión como el tamaño del árbol se define como:

$$RSS + \alpha |T| \quad (4)$$

donde  $\alpha$  es un parámetro de penalización (necesita ser calibrado). Dicho parámetro se ajusta con validación cruzada. En este caso, dicho parámetro NO tiene ninguna fundamentación teórica.

## 1 Procedimientos de selección automática de variables

## 2 Estructura del modelo

*Recursive binary splitting*

Estructurando un árbol

Ejemplo conceptual de un árbol

## 3 Ventajas y desventajas

## 4 Inferencia

- La punta más alta del árbol, donde se implementa la primera bifurcación, se conoce como el "nodo raíz" (*root node*).
- Una rama (*branch*) es una región resultante de la partición de una variable.
- Si un nodo no tiene ramas, este se conoce como nodo terminal (*terminal node*) u hoja (*leaf*).
- Crecer (*growing*) un árbol es aumentar el número de ramas, mientras que disminuir el número de estas se conoce como podar el árbol (*pruning*).



- Considere el modelo en el que la variable respuesta depende de dos variables,  $x_1$  y  $x_2$ .
- Sean  $\wedge$  y  $\vee$  los operadores "y" & "o", respectivamente y sea  $I(.)$  una función indicadora que toma el valor de 1 cuando el argumento es cierto, cero en otro caso.
- Entonces,

$$y_i = \beta_1 I(x_{1i} < c_1 \wedge x_{2i} < c_2) + \beta_2 I(x_{1i} < c_1 \wedge x_{2i} \geq c_2) \\ + \beta_3 I(x_{1i} \geq c_1 \wedge x_{2i} < c_2) + \beta_4 I(x_{1i} \geq c_1 \wedge x_{2i} \geq c_2) \quad (5)$$

- En este caso, se divide el espacio de información  $(x_1, x_2)$  en 4 regiones (*branches*).

- En cada región, la predicción de la variable respuesta es la misma para todos los valores de  $x_1$  y  $x_2$  que pertenecen a esa región. Si suponemos que  $c_1 = 2$  y  $c_2 = 3$ , las 4 regiones son:

$$\{(x_1, x_2) : x_1 < 2 \wedge x_2 < 3\}$$

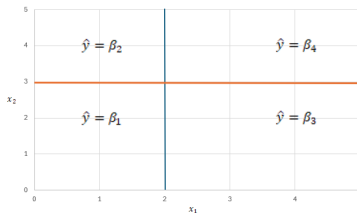
$$\{(x_1, x_2) : x_1 < 2 \wedge x_2 \geq 3\}$$

$$\{(x_1, x_2) : x_1 \geq 2 \wedge x_2 < 3\}$$

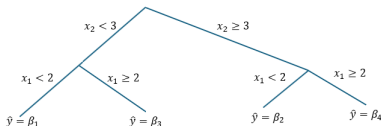
$$\{(x_1, x_2) : x_1 \geq 2 \wedge x_2 \geq 3\}$$

- Con esta configuración, se pueden hallar los parámetros que gobiernan el pronóstico de  $\hat{y}$ .

- Gráficamente, el espacio creado por  $x_1, x_2$  para estimar y se ve gráficamente así:



- Además, el árbol luce de la siguiente manera:



- 1 Procedimientos de selección automática de variables
- 2 Estructura del modelo
- 3 Ventajas y desventajas**
- 4 Inferencia

Entre las ventajas podemos mencionar:

- Los modelos pueden mostrarse gráficamente.
- Permite explorar las variables en orden, con el fin de capturar su habilidad para particionar (importancia de la variable).
- Pueden tratar variables cualitativas directamente, sin crear variables dicótomas.
- Son relativamente robustos a datos atípicos.
- Si se tienen muchos datos, es posible estimar medias no lineales e interacciones sin necesidad de especificarlos previamente.
- Puede ser robusto a heteroscedasticidad.

Entre las desventajas se tienen:

- Los árboles no producen coeficientes de regresión, por lo que no es posible cuantificar la relación entre  $y$  y las variables  $X$ .
- A medida que crece el número de variables, se hace mucho más demorado estimar el modelo.
- Tiene tendencia a sobreajustar, por lo que su poder de predicción es menor a la de métodos alternativos.
- Tienden a ser inestables, es decir, pequeños cambios en la muestra de entrenamiento llevan a cambios drásticos en las predicciones.

- 1 Procedimientos de selección automática de variables
- 2 Estructura del modelo
- 3 Ventajas y desventajas
- 4 Inferencia

- En primera instancia, es posible ver visualmente el papel de cada variable en producir particiones o, en otras palabras, su importancia dentro del árbol.
- Sin embargo, cuando el número de variables es grande, la visualización es muy difícil. En este caso, se evalúa la importancia de la variable mediante la medición en el RSS si la variable es excluida del árbol.
- El promedio del RSS para todos los árboles que contengan alguna variable  $x_j$  se conoce como *score*. Un valor alto del *score* indica que remover dicha variable conlleva a, en promedio, incrementos en RSS, por lo que se considera que ella es "importante".



- Técnicamente, para evaluar la importancia de una variable en el árbol se usa el indicador de importancia relativa:

$$I_j^2 = \sum_{t=1}^{T-1} e_t^2 I(v(t) = j) \quad (6)$$

con  $T$  siendo el número de nodos internos (no hojas),  $v(t)$  es la variable seleccionada en el nodo  $t$ ,  $e_t$  es la mejora en el error antes y después de particionar el espacio vía la variable  $v(t)$ .

- en el caso de *Regression Trees*, se puede utilizar la ganancia de información ( $Gain(D)$ ).

- La ganancia de información ( $Gain(D)$ ) se define como:

$$Gain(D) = \text{Variación}(S) - \text{Variación}_D(S)$$

con  $\text{Variación}(S) = \sum_{i=1}^N (y_i - \bar{y})^2$  y

$\text{Variación}_D(S) = \frac{1}{v} \sum_{j=1}^v \text{Variación}_j(S)$  y  $v$  igual al número de sub-espacios seaparados por  $D$ . Ejemplo: en partición binaria,  $v = 2$ .

- Interpretación de la Ecuación 6:
  - Si la variable  $j$  es muy importante, la mejora en el error es sustancial, por lo que  $I(v(t) = j)$  será usualmente 1 e  $I_j^2$  será grande.
  - Si la variable no es muy importante,  $I_j^2$  será pequeño.

**Advertencia:** no hay propiedades teóricas conocidas de que este método de selección de variables sea válido. De allí que los *scores* deban tomarse con precaución.

De allí que el método de importancia de variables sea de uso limitado en términos de inferencia en árboles.