

Lecture 8: Machine Learning

Árboles y Bosques

Big Data and Machine Learning en el Mercado Inmobiliario
Educación Continua

Ignacio Sarmiento-Barbieri

Universidad de los Andes

November 2, 2023

Agenda

1 Árboles

2 Break

Más allá de la linealidad

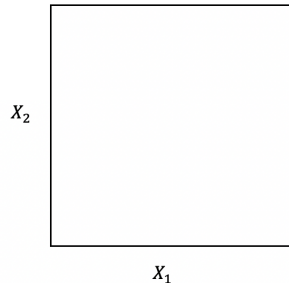
- ▶ El objetivo es predecir Y dadas otras variables X . Ej: precio vivienda dadas las características
- ▶ Asumimos que el link entre Y and X esta dado por el modelo:

$$Y = f(X) + u \quad (1)$$

- ▶ Hasta ahora vimos modelos lineables o linealizables.
 - ▶ Regresión lineal, lasso, ridge, elastic net
- ▶ Árboles (CARTs)
 - ▶ Modelo flexible e interpretable para la relación entre Y y X .
 - ▶ Para que? No-linealidades, interacciones.

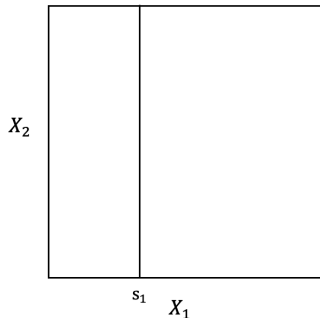
Árboles: que hacen?

- 1 Y es la variable a predecir, los insumos son X_1 y X_2
- 2 Partimos el espacio (X_1, X_2) en dos regiones, en base a una sola variable (particion horizontal o vertical).



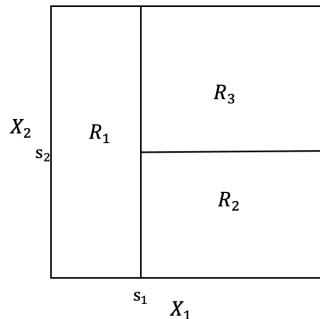
Árboles: que hacen?

- 1 Y es la variable a predecir, los insumos son X_1 y X_2
- 2 Partimos el espacio (X_1, X_2) en dos regiones, en base a una sola variable .
- 3 Dentro de cada región proponemos como predicción la media muestral de Y en cada región.
- 4 Punto: elegir la variable y el punto de partición de manera optima (mejor ajuste global).



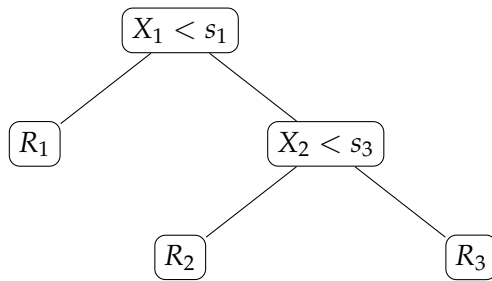
Árboles: que hacen?

- 1 Y es la variable a predecir, los insumos son X_1 y X_2
- 2 Partimos el espacio (X_1, X_2) en dos regiones, en base a una sola variable .
- 3 Dentro de cada región proponemos como predicción la media muestral de Y en cada región.
- 4 Punto: elegir la variable y el punto de partición de manera optima (mejor ajuste global).

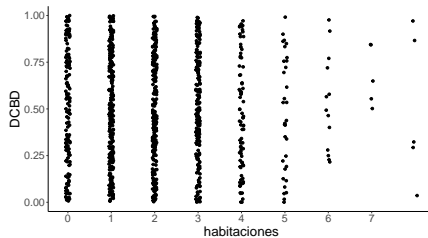
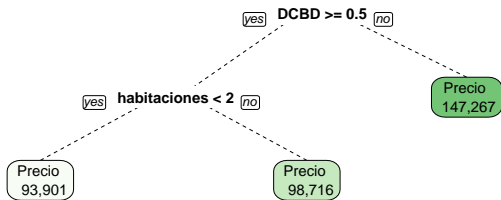


Árboles: que hacen?

- 1 Y es la variable a predecir, los insumos son X_1 y X_2
- 2 Partimos el espacio (X_1, X_2) en dos regiones, en base a una sola variable (partición horizontal o vertical).
- 3 Dentro de cada región proponemos como predicción la media muestral de Y en cada región.
- 4 Punto: elegir la variable y el punto de partición de manera optima (mejor ajuste global).
- 5 Continuamos partiendo



Árboles: que hacen?



Árboles: cómo lo hacen?

- ▶ Tenemos datos $y_{n \times 1}$ (precio) y $X_{n \times p}$ (características)
- ▶ Definiciones
 - ▶ j es la variable que parte el espacio y s es el punto de partición
 - ▶ Defina los siguientes semiplanos

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

- ▶ El problema se reduce a buscar la variable de partición X_j y el punto s de forma tal que

$$\min_{j, s} \left[\min_{c_1} \sum_{x_i \in R_1(j, s)} (y - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j, s)} (y - c_2)^2 \right] \quad (3)$$

Árboles: cómo lo hacen?

- ▶ Para cada variable y punto, la minimización interna es la media

$$\hat{c}_m = \frac{1}{n_m} \sum (y_i | x_i \in R_m) \quad (4)$$

- ▶ El proceso se repite para todas las regiones

Árboles: cómo lo hacen?

- ▶ Para cada variable y punto, la minimización interna es la media

$$\hat{c}_m = \frac{1}{n_m} \sum (y_i | x_i \in R_m) \quad (4)$$

- ▶ El proceso se repite para todas las regiones
- ▶ El árbol final tiene M regiones

$$\hat{f}(x) = \sum_{m=1}^M \hat{c}_m I(x \in R_m) \quad (5)$$

Break