Lecture 8: Machine Learning
Mas allá de la linealidad

Big Data and Machine Learning en el Mercado Inmobiliario
Educación Continua

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 9, 2021

Agenda

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:

$$Y = f(X) + u \tag{1}$$

- donde $f(X)$ es la función de interés

- u una variable aleatoria no observable $E(u) = 0$ and $V(u) = \sigma^2$

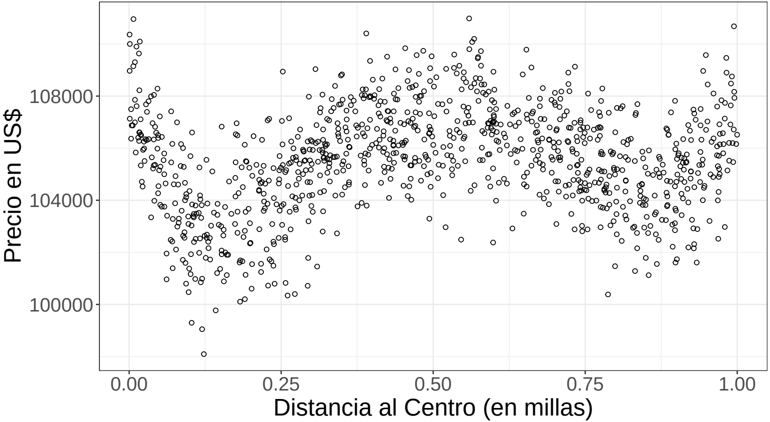
Más allá de la linealidad

- El supuesto de linealidad es bueno en muchos problemas de machine learning (aprendizaje automático).

$$y_i = \alpha + \beta x_i + u_i \quad i = 1, \ldots, n \tag{2}$$

- Sin embargo, existen otros métodos que ofrecen mucha flexibilidad, sin perder la facilidad e interpretabilidad de los modelos lineales:
  - Regresión polinomial
  - Funciones escalonadas
  - Splines de regresión
  - Regresión local
  - CARTs

Regresión polinomial



Regresión polinomial

- Reemplace el modelo lineal estándar

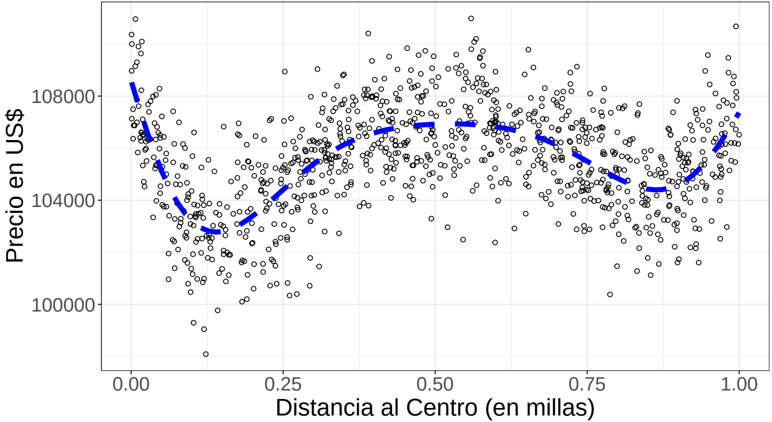
$$y_i = \alpha + \beta x_i + u_i \;\; i = 1, \ldots, n \tag{3}$$

- con una función polinomial:

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + u_i \;\; i = 1, \ldots, n \tag{4}$$

- Para un grado d suficientemente grande , una regresión polinomial nos permite producir una curva extremadamente no lineal.
- Hacemos esto creando nuevas variables $x_1 = x$, $x_2 = x^2$, etc. y luego las tratamos como regresión lineal múltiple MCO.
- Como elegir d? $\rightarrow$ validación cruzada

Regresión polinomial



Funciones escalonadas

- El uso de funciones polinomiales de las características como predictor en un modelo lineal impone una estructura global a la función no lineal de X.
- Para evitar imponer una estructura tan global, podemos crear transformaciones de una variable cortando la variable en distintas regiones.
- En particular, usamos indicadores para dividir X en regiones, y ajustamos una constante diferente en cada región.

Funciones escalonadas

- Esto equivale a convertir una variable continua en una variable categórica ordenada .

- Creamos puntos de corte (o nudos) $C_1, C_2, \ldots, C_K$, en el rango de X y luego construimos $K+1$ nuevas variables: donde $I(.)$ es una función indicadora que devuelve un 1 si la condición es verdadera y 0 en caso contrario.

$$
\begin{aligned}
C_0(X) &= I(X < c_1) \\
C_1(X) &= I(c_1 \leq X \leq c_2) \\
C_2(X) &= I(c_2 \leq X \leq c_3) \\
&\vdots \\
C_{K-1}(X) &= I(c_{K-1} \leq X \leq c_K) \\
C_K(X) &= I(c_K \leq X)
\end{aligned} \tag{5}
$$

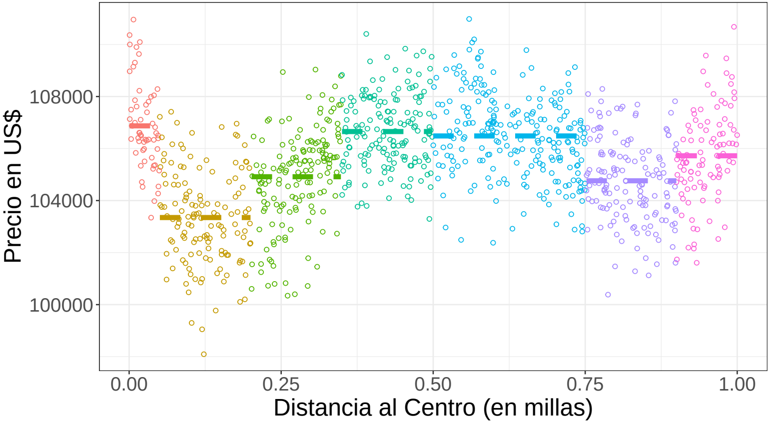
Funciones escalonadas

- Luego usamos la estimación MCO para ajustar un modelo lineal usando estas nuevas variables $K+1$:

$$y_i = \alpha + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \beta_3 C_3(x_i) + \cdots + \beta_K C_K(x_i) + u_i \;\; i = 1, \ldots, n \tag{6}$$

- Note que cuando $X < c_1$, todos los predictores restantes son cero, por lo que β_0 es el valor promedio de Y cuando $X < c_1$

Funciones escalonadas



- A menos que haya puntos de interrupción naturales en los predictores, las funciones escalonadas pueden ”miss the action”.

Funciones base

- Los modelos de regresión polinomial y escalonado son casos especiales de un enfoque de función base.
- La idea es tener a mano una familia de funciones o transformaciones que se puedan aplicar a una variable X : $b_1(X), \ldots, b_K(X)$
- En lugar de ajustar un modelo lineal en X , ajustamos el siguiente modelo:

$$y_i = \alpha + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \cdots + \beta_K b_K(x_i) + u_i \quad i = 1, \ldots, n \tag{7}$$

- Tenga en cuenta que las funciones base $b_1(x_i)), \ldots, B_K(x_i))$ son fijas y conocidas.
  - Para la regresión polinomial, las funciones base son $b_j(x_i) = x_i^j$
  - Para la regresión escalonada, las funciones base son $b_j(x_i) = I(c_j \leq x_i < c_{j+1})$

Funciones base

- Entonces las funciones polinomiales y escalonadas son ejemplos de funciones bases.
- Que otros?

Funciones base

- Entonces las funciones polinomiales y escalonadas son ejemplos de funciones bases.
- Que otros?
- Fourier series, wavelets, etc
- Una muy popular son *Splines*

Splines de regresión

- Los splines de regresión son una clase flexible de funciones base que se extienden sobre las regresiones polinomiales y los enfoques de regresión escalonada (constante por partes).
- Implican dividir el rango de X en K regiones distintas; dentro de cada región, una función polinomial se ajusta a los datos.
- Estos polinomios están restringidos para que se unan suavemente en los límites de la región (o nudos).
- Siempre que el intervalo se divida en suficientes regiones, esto puede producir un fit extremadamente flexible.

Splines de regresión

- En lugar de ajustar un polinomio de alto grado en todo el rango de X ,
- Splines implica ajustar polinomios de bajo grado separadas más de diferentes regiones de X .
- Aquí, los coeficientes beta difieren en diferentes partes del rango de X ; los puntos donde cambian los coeficientes se llaman nudos.
- Ejemplo : un polinomio cúbico a trozos con un solo nudo en un punto c toma la siguiente forma:

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i \;\; i = 1, \ldots, n \tag{8}$$

$$y_i = \begin{cases} \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i & \textit{if } x_i < c \\ \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i & \textit{if } x_i \geq c \end{cases} \tag{9}$$

Splines de regresión

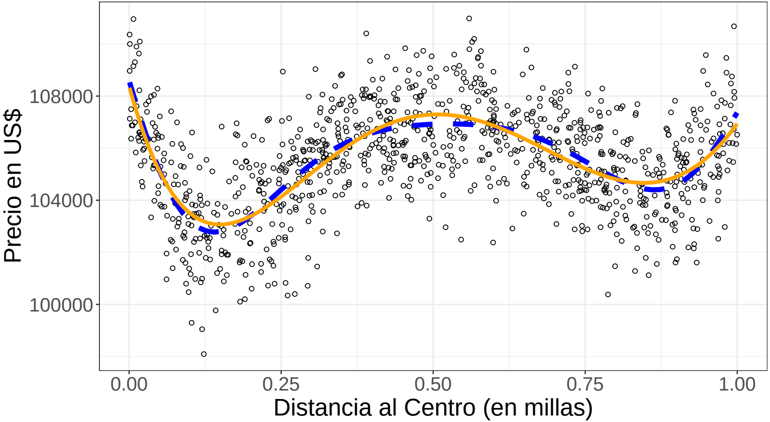
- Cada una de las funciones polinomiales se puede ajustar utilizando MCO aplicado a funciones simples del predictor original.
- El uso de más nudos conduce a un polinomio por partes más flexible.
- Si es general, si colocamos K nudos diferentes en el rango de X , terminamos ajustando $K + 1$ polinomios diferentes.

Splines de regresión

Elegir la ubicación de los nudos

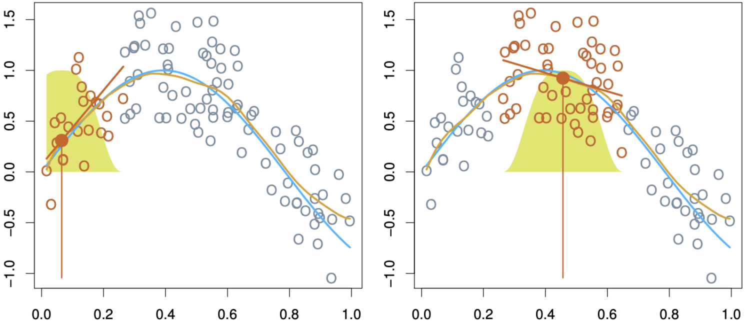
- El spline de regresión es más flexible en regiones que contienen muchos nudos, porque en esas regiones los coeficientes polinomiales pueden cambiar rápidamente.
- Una opción es colocar más nudos en los lugares donde creemos que la función puede variar más rápidamente y colocar menos nudos donde parece más estable.
- En la práctica, es común colocar los nudos de manera uniforme. Por ejemplo, una estrategia es decidir K , el número de nudos, y luego colocarlos en los cuantiles apropiados del X observado .
- Un enfoque más objetivo es utilizar la validación cruzada.

Splines de regresión



Regresión local

- La regresión local es un enfoque diferente para ajustar funciones no lineales flexibles,
- Implica calcular el ajuste en un punto objetivo utilizando solo las observaciones de entrenamiento cercanas.



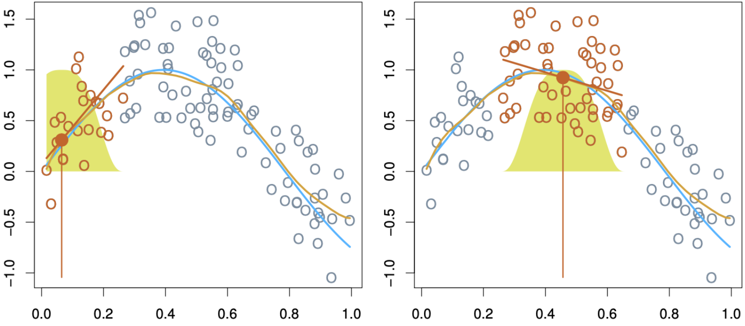
Regresión local

- Algoritmo de regresión local en $X = x_0$

  1. Obtener la fracción $s = k/n$, de puntos de la muestra de entrenamiento cuyos x_i son los mas cercanos a x_0
  2. Asignar el peso $K_{i0} = K(x_i, x_0)$ a cada punto en el barrio cercano, 0 si no esta en el vecindario
  3. Fit mínimos cuadrados ponderados de y en x usando los pesos obtenidos en el paso anterior
  4. Obtener el valor predicho $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1$

Regresión local

- Cuanto menor sea el valor del intervalo s , más local y ondulado será nuestro ajuste.
- Un valor muy grande de s conducirá a un ajuste global a los datos usando todas las observaciones de entrenamiento.
- Podemos usar la validación cruzada para elegir s o especificarlo directamente.
- Otras elecciones a considerar: la función de ponderación $K(x_i, x_0)$, y si se ajusta a una regresión lineal, constante o cuadrática.

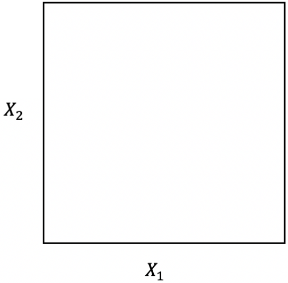


Árboles: Motivacion

- El modelo que queremos es $y = f(x) + \epsilon$ para mejorar la predicción
  - Hasta ahora vimos modelos lineables o linealizables.
    - Regresión lineal
    - Regresión polinomial
    - Funciones escalonadas
    - Splines de regresión
    - Regresión local
- Arboles (CARTs)
  - Modelo flexible e interpretable para la relación entre Y y X.
  - Para que? No-linealidades, interacciones.

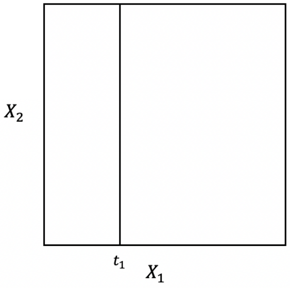
Árboles: que hacen?

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).



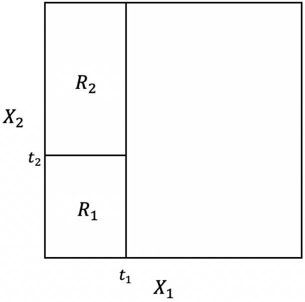
Trees: Background

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).



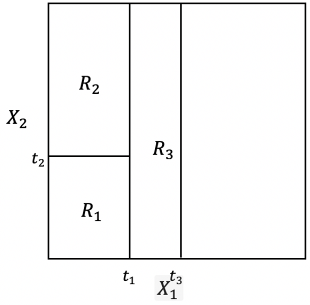
Trees: Background

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



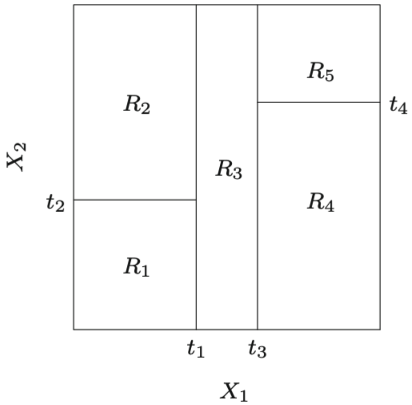
Trees: Background

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



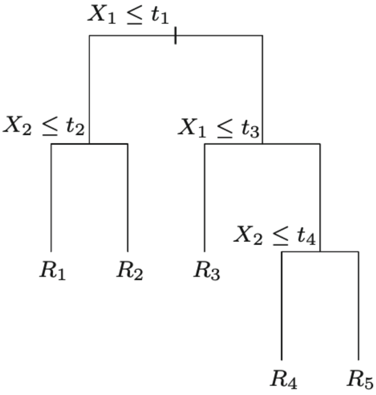
Trees: Background

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

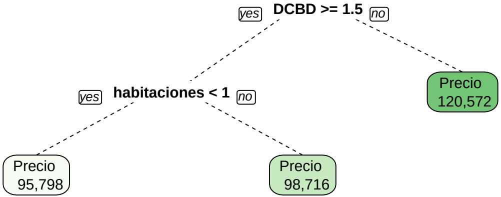


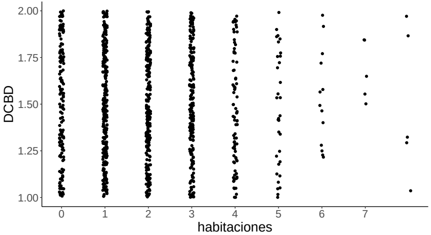
Trees: Background

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



Trees

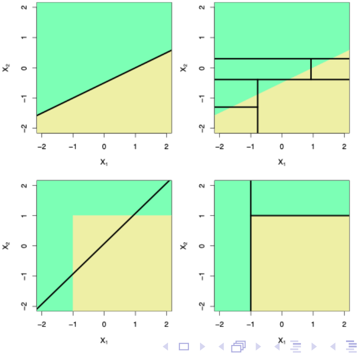

DCBD >= 1.5
yes
no
habitaciones < 1
Precio 120,572
Precio 95,798
Precio 98,716


DCBD
2.00
1.75
1.50
1.25
1.00
0
1
2
3
4
5
6
7
habitaciones

Árboles vs. Modelos Lineales

- Cuál modelo es mejor?
  - Si la relación entre los predictores y la respuesta es lineal, los modelos lineales clásicos, como la regresión lineal, superan a los árboles de regresión.
  - Por otro lado, si la relación entre los predictores no es lineal, los árboles de decisión superarían a los enfoques clásicos.

- Arriba: el límite es lineal
  - Izquierda: modelo lineal (bueno)
  - Derecha: árbol
- Abajo: el límite es no-lineal
  - Izquierda: linear model
  - Derecha: arbol (good)



Ventajas y Desventajas de los Árboles

- Pros:
  - Los árboles son muy fáciles de explicar a las personas (probablemente incluso más fáciles que la regresión lineal)
  - Los árboles se pueden trazar gráficamente y son fácilmente interpretados incluso por no expertos. Variables más importantes en la parte superior
  - Funcionan bien en problemas de clasificación y regresión.
- Cons:
  - Los árboles no son muy precisos o robustos (ensamblados, bosques aleatorios y boosting al rescate)
  - Si la estructura es lineal, CART no funciona bien

Further Readings

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Volvemos en 5 min con `Python`