# Lecture 9: Machine Learning
# Árboles y Bosques

## Big Data and Machine Learning en el Mercado Inmobiliario
## Educación Continua

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 13, 2021

# Agenda

# Más allá de la linealidad

- El objetivo es predecir $Y$ dadas otras variables $X$. Ej: precio vivienda dadas las características

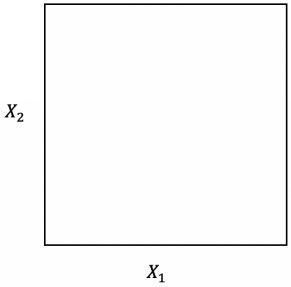
- Asumimos que el link entre $Y$ and $X$ esta dado por el modelo:

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

- Hasta ahora vimos modelos lineables o linealizables.
  - Regresión lineal, polinomial, escalonadas, splines, regresión local
- Árboles (CARTs)
  - Modelo flexible e interpretable para la relación entre Y y X.
  - Para que? No-linealidades, interacciones.

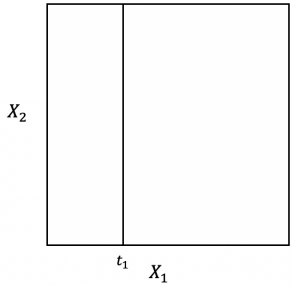
# Árboles: que hacen?

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).



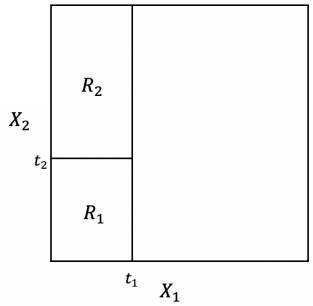
# Árboles: que hacen?

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).



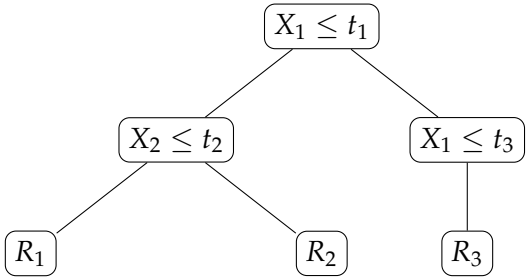
# Árboles: que hacen?

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).

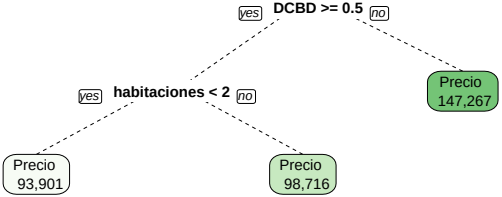


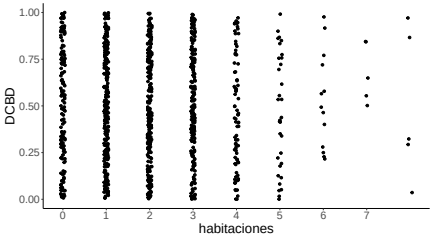
# Árboles: que hacen?

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

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

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

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

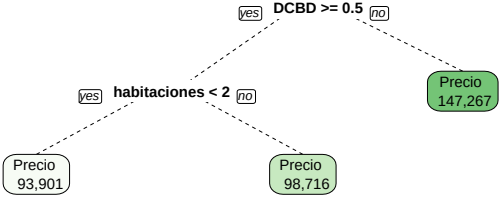
# Árboles: cómo lo hacen?

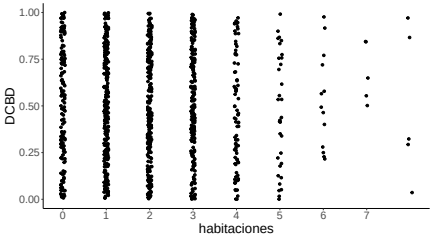
- Para cada variable y punto, la minimización interna es la media

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

- El proceso se repite para todas las regiones

# Árboles: cómo lo hacen?



# Árboles: cómo lo hacen?

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

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

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

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

# Árboles: cómo lo hacen?

- El árbol creció, como lo paramos?
- Si el árbol es muy grade, tenemos overfit
- Un árbol mas chico, puede tener menos regiones. Esto puede llevar a una varianza menor y mejor interpretación al costo de un poco sesgo.
- Solución: Pruning (poda)
  - Dejar crecer un árbol muy grande $T_0$
  - Cortarlo te quedas con un sub-árbol (*subtree*)
  - Como determinamos la mejor forma de cortarlo? $\rightarrow$ menor error de predicción usando cross-validation

# Árboles: cómo lo hacen?

- Desventaja, calcular el error de predicción usando cross-validation para cada sub-árbol posible es demasiado (muchos sub-árboles posbiles)
- Solución: *Cost complexity pruning (cortar las ramas mas débiles)*
  - Indexamos los arboles con $T$.
  - Un sub-árbol $T \in T_0$ es un árbol que se obtuvo colapsando los nodos terminales de otro árbol cortando ramas.
  - $[T]$ = número de nodos terminales del arbol 3 $T$

# Árboles: cómo lo hacen?

- Cost complexity del árbol $T$

$$C_\alpha(T) = \sum_{m=1}^{[T]} n_m Q_m(T) + \alpha[T] \tag{6}$$

  - donde $Q_m(T) = \frac{1}{n_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$ para los árboles de regresión
  - $Q_m(T)$ penaliza la heterogeneidad dentro de la regresión y el número de regiones
  - Objetivo: para un dado $\alpha$, encontrar el pruning optimo que minimice $C_\alpha(T)$

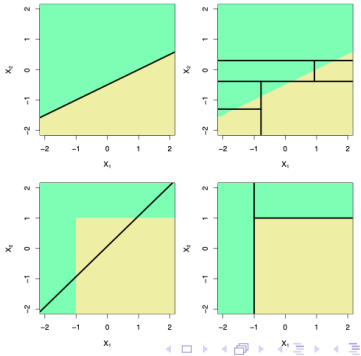
# Árboles: cómo lo hacen?

- Mecanismo de búsqueda para $T_\alpha$ ( pruning optimo dado $\alpha$).
  - Resultado: para cada $\alpha$ hay un sub-árbol único $T_\alpha$ que minimiza C$\alpha$ (T).
  - Ramas mas débiles: eliminar sucesivamente las ramas que producen un aumento mínimo en $\sum_{m=1}^{[T]} n_m Q_m(T)$
  - Idea: remover ramas es colapsar, esto aumenta la heterogenidad, ergo, colapsamos las particiones menos necesarias.
  - Esto eventualmente colapsa hasta el nodo inicial (stump) pero va a través de una sucesión de árboles, que va del mas grande al mas pequeño cortando las ramas mas débiles.
  - Breiman et al. (1984): $T_\alpha$ pertenece a esta sequencia.
  - Uno puede enfocar la búsqueda en esta sucesión de sub-árboles.
  - Elección de $\alpha$: cross validation.

# 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

# Bagging

- Problema con CART: varianza alta.
- Podemos mejorar mucho el rendimiento mediante la agregación
- Bagging:
  - Obtenga repetidamente muestras aleatorias $(X_i^b, Y_i^b)_{i=1}^N$ de la muestra observada.
  - Para cada muestra de arranque, ajuste un árbol de regresión $\hat{f}^b(x)$
  - Promedie las muestras de bootstrap

$$\hat{f}_{bag} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x) \tag{7}$$

- Básicamente estamos suavizando las predicciones.
- Idea: la varianza del promedio es menor que la de una sola predicción.

# Random Forests

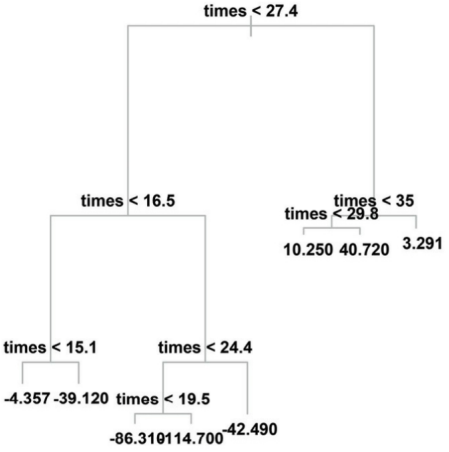
- Problema con el bagging: si hay un predictor fuerte, diferentes árboles son muy similares entre sí. Si hay alta correlación, ¿está realmente reduciendo la varianza?
- Bosques (forests): reduzca la correlación entre los árboles en el boostrap.
- Si hay $p$ predictores, en cada partición use solo $m < p$ predictores, elegidos al azar.
- Bagging es forests con $m = p$ (usando todo los predictores en cada partición).
- Tipicamente $m = \sqrt{(p)}$

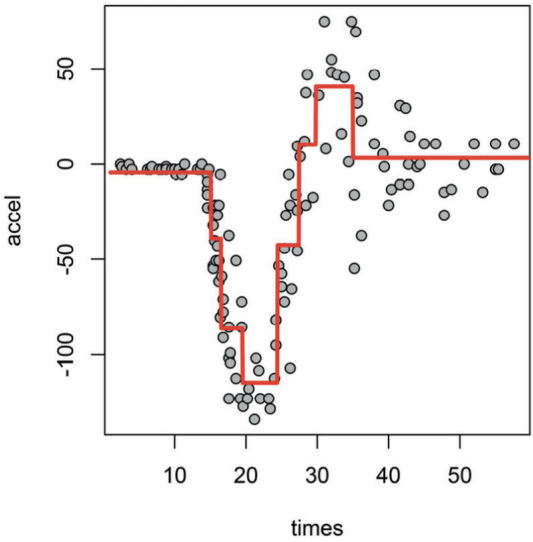
# Random Forests

Trees:

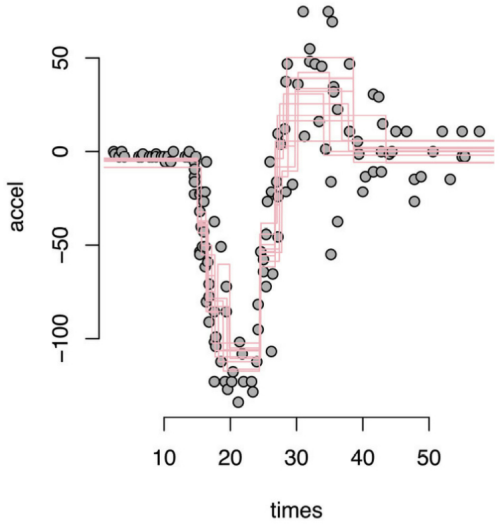
Random Forests:

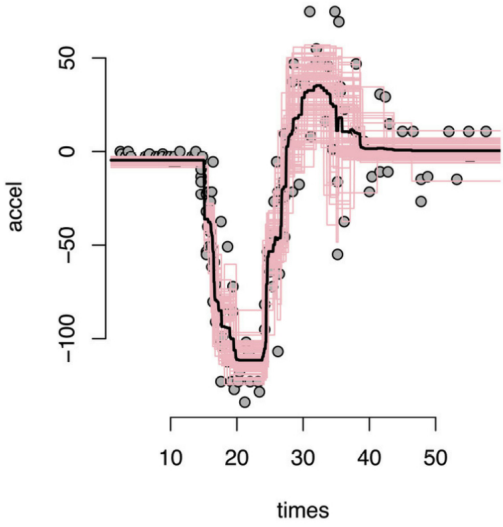
# Random Forests



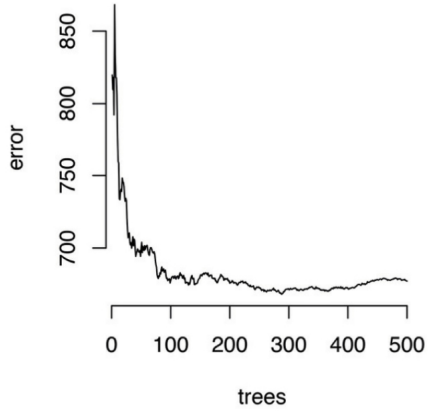



# Random Forests

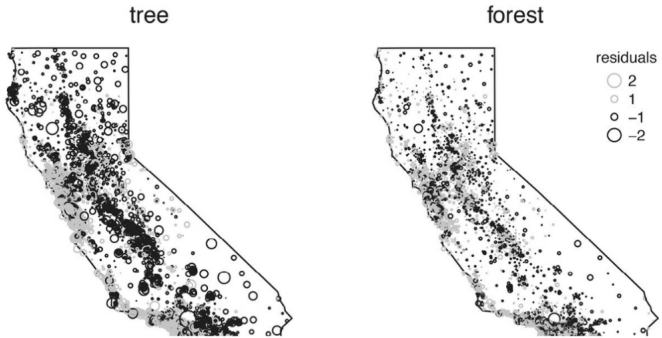


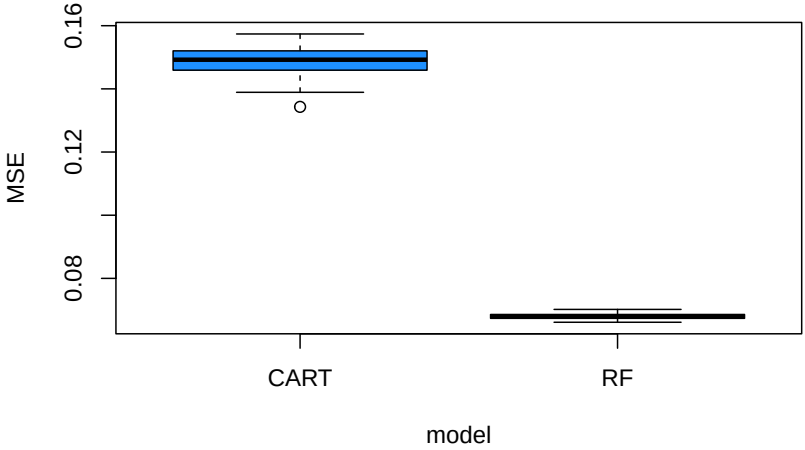
# Random Forests



# Residuales en muestra



# MSE Fuera de Muestra



# Repaso & Próxima Clase

- Árboles
- Bagging y Bosques (Random Forests)
- Próxima Clase: Boosting

# Para seguir leyendo


- Breiman, L. (2001). “Random Forests”. In: Machine Learning. ISSN: 1098-6596. DOI: 10.1017/CBO9781107415324.004. eprint: arXiv:1011.1669v3.
- 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.
- Kasy M. (2019). Trees, forests, and causal trees. Mimeo.
- Taddy, M. (2019). Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. McGraw Hill Professional.

# Volvemos en 5 min con Python