

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

May 18, 2023

Agenda

- 1 Recap: Regularización
- 2 Árboles
- 3 Bagging and Random Forests
- 4 Break

Recap: Regularization

- ▶ Para $\lambda \geq 0$ dado, consideremos el siguiente problema de optimización
- ▶ Lasso:

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (1)$$

- ▶ Ridge:

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^p (\beta_j)^2 \quad (2)$$

Recap: Regularization

- ▶ Elastic Net es un happy medium

$$\min_{\beta} EL(\beta) = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 + \alpha\lambda \sum_{j=1}^p |\beta_j| + (1 - \alpha)\lambda \sum_{j=1}^p (\beta_j)^2 \quad (3)$$

- ▶ Si $\alpha = 1$ Lasso
- ▶ Si $\alpha = 0$ Ridge
- ▶ Como elegir (α, λ) ? \rightarrow Validación Cruzada Bidimensional

Más allá de la linealidad

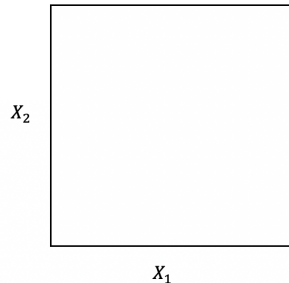
- ▶ El objetivo es predecir Y dadas otras variables X . Ej: precio vivienda dadas las características
- ▶ Asumimos que el link entre Y and X esta dado por el modelo:

$$Y = f(X) + u \quad (4)$$

- ▶ 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.

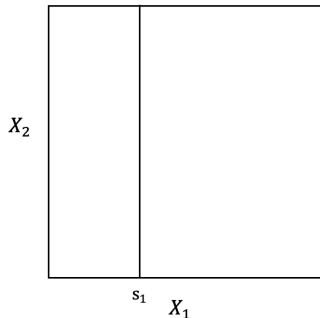
Árboles: que hacen?

- 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).



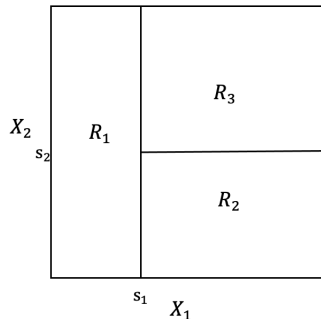
Árboles: que hacen?

- 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).



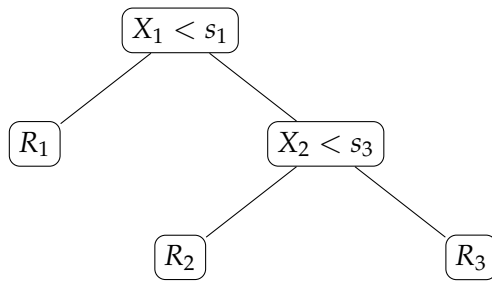
Árboles: que hacen?

- 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).

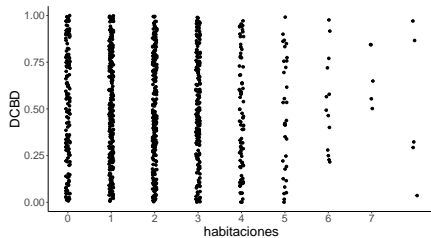
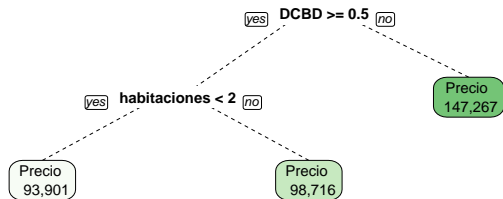


Árboles: que hacen?

- 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

$$R_1(j, s) = \{X | X_j \leq s\} \ \& \ R_2(j, s) = \{X | X_j > s\} \quad (5)$$

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

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y - c_2)^2 \right] \quad (6)$$

Árboles: cómo lo hacen?

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

$$\hat{c}_m = \frac{1}{n_m} \sum (y_i | x_i \in R_m) \quad (7)$$

- ▶ El proceso se repite para todas las regiones

Árboles: cómo lo hacen?

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

$$\hat{c}_m = \frac{1}{n_m} \sum (y_i | x_i \in R_m) \quad (7)$$

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

$$\hat{f}(x) = \sum_{m=1}^M \hat{c}_m I(x \in R_m) \quad (8)$$

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: *Cost complexity pruning* (cortar las ramas mas débiles)

$$C_\lambda(T) = \sum_{m=1}^{[T]} n_m Q_m(T) + \lambda[T] \quad (9)$$

- ▶ 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 λ , encontrar el pruning óptimo que minimice $C_\lambda(T)$

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 aleatoria, ajuste un árbol de regresión $\hat{f}^b(x)$
 - ▶ Promedie las muestras de bootstrap

$$\hat{f}_{bag} = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(x) \quad (10)$$

- ▶ 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 bootstrap.
- ▶ 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 $m = \sqrt{p}$

Árboles vs Random Forests

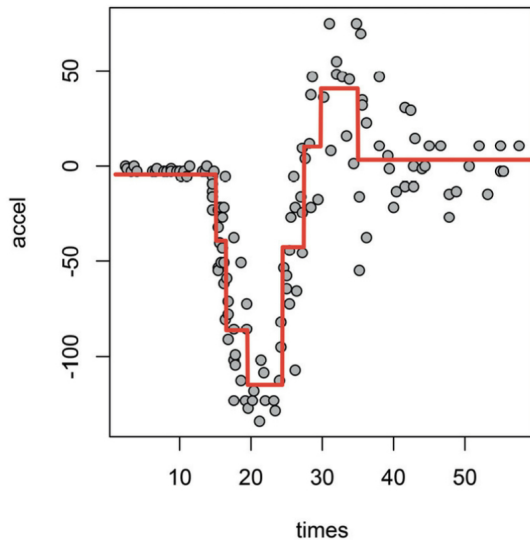
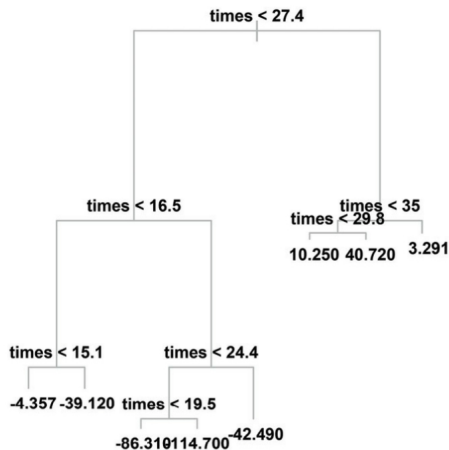
Trees:



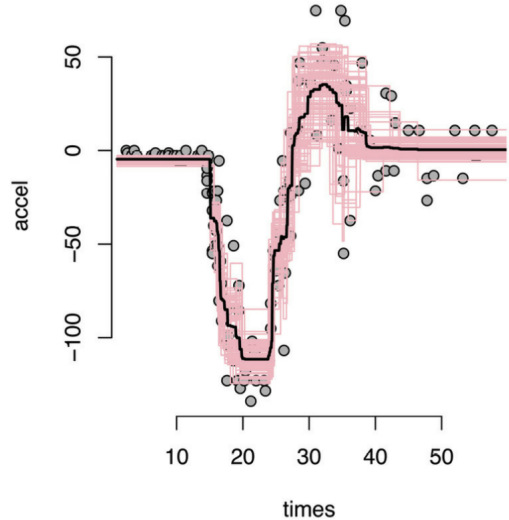
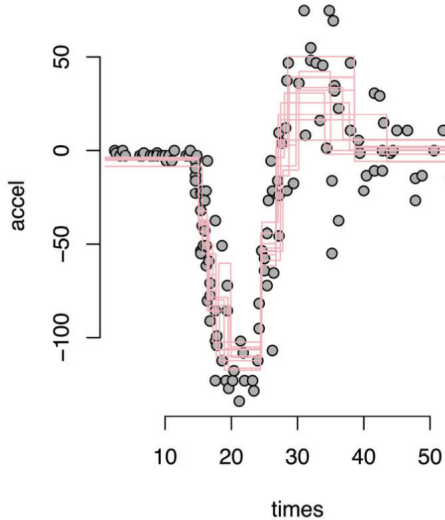
Random Forests:



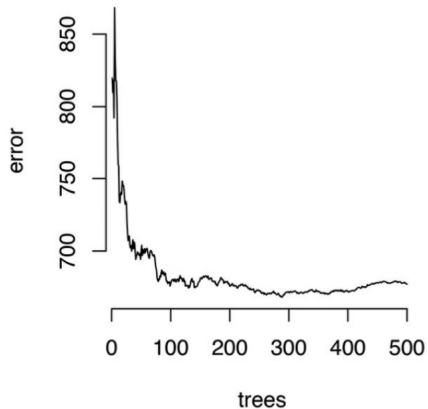
Árboles vs Random Forests



Árboles vs Random Forests



Árboles vs Random Forests



Break