

Lecture 7: Machine Learning

Regularización

Big Data and Machine Learning en el Mercado Inmobiliario
Educación Continua

Ignacio Sarmiento-Barbieri

Universidad de los Andes

May 16, 2023

Agenda

- 1 Recap
- 2 Modelos Lineales
- 3 Regularización
 - Lasso
 - Ridge
 - Elastic net
- 4 Break
- 5 Para seguir leyendo

- 1 Recap
- 2 Modelos Lineales
- 3 Regularización
 - Lasso
 - Ridge
 - Elastic net
- 4 Break
- 5 Para seguir leyendo

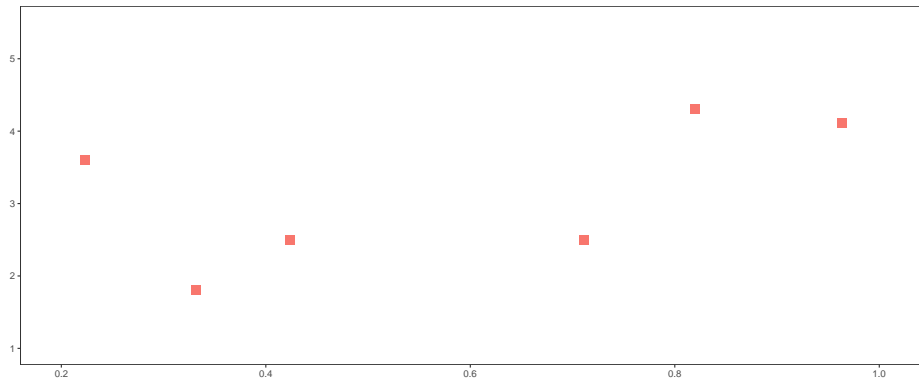
Objetivos

- ▶ 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:

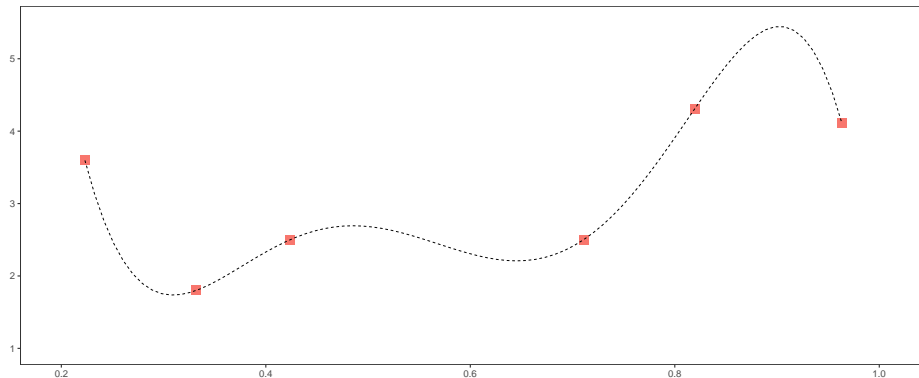
$$y = f(X) + u \tag{1}$$

- ▶ donde $f(X)$ es cualquier función,
- ▶ u una variable aleatoria no observable $E(u) = 0$ and $V(u) = \sigma^2$

Overfit



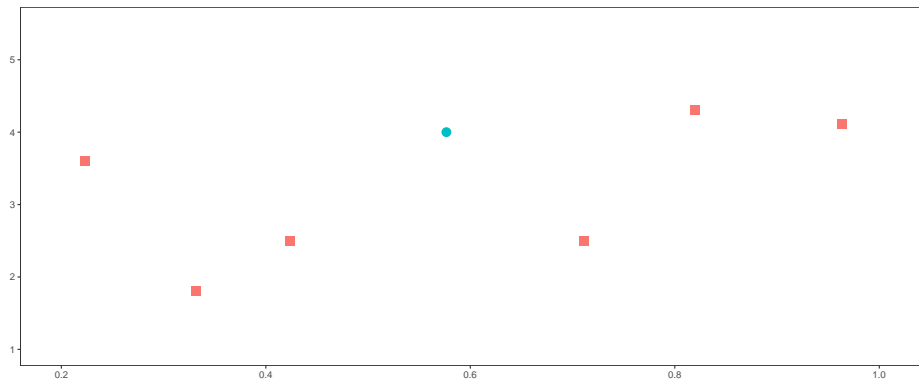
Overfit



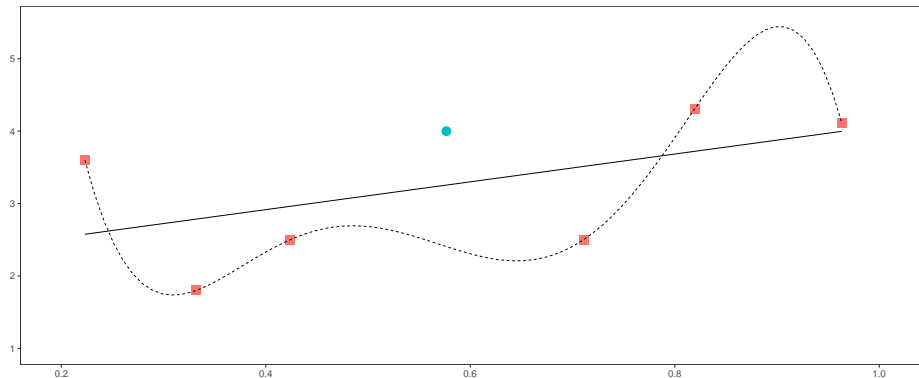
Overfit y Predicción fuera de Muestra

- ▶ ML nos interesa la predicción fuera de muestra
- ▶ Overfit: modelos complejos predicen muy bien dentro de muestra, pero tienden a hacer un trabajo fuera de muestra
- ▶ Hay que elegir el nivel adecuado de complejidad

Overfit



Overfit



Overfit y Predicción fuera de Muestra

Selección de Modelos

- ▶ ML nos interesa la predicción fuera de muestra
- ▶ Overfit: modelos complejos predicen muy bien dentro de muestra, pero tienden a hacer un mal trabajo fuera de muestra
- ▶ Hay que elegir el modelo que “mejor” prediga
 - ▶ Métodos de Remuestreo
 - ▶ Enfoque del conjunto de validación
 - ▶ Loocv
 - ▶ Validación cruzada en K-partes (5 o 10)

- 1 Recap
- 2 Modelos Lineales**
- 3 Regularización
 - Lasso
 - Ridge
 - Elastic net
- 4 Break
- 5 Para seguir leyendo

Modelos Lineales

- El problema es:

$$y = f(X) + u \quad (2)$$

Modelos Lineales

- El problema es:

$$y = f(X) + u \quad (2)$$

- proponemos :

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (3)$$

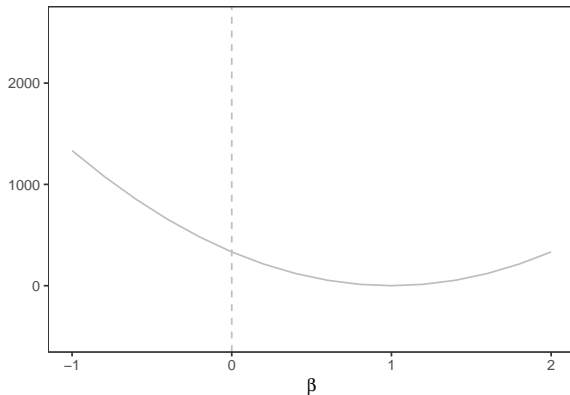
Intuición en 1 Dimension

$$\text{Precio}_i = \beta \text{Habitaciones}_i + \epsilon_i \quad (4)$$

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 \quad (5)$$

Intuición en 1 Dimension

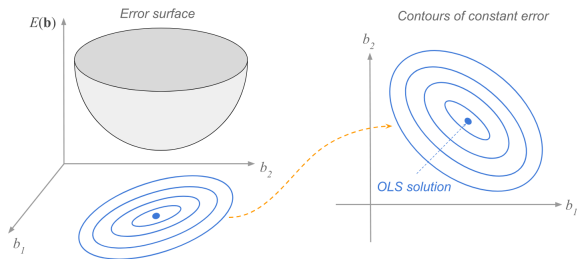
$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 \quad (6)$$



Intuición en 2 Dimensiones (OLS)

$$\text{Precio}_i = \beta_1 \text{Habitaciones}_i + \beta_2 \text{Superficie}_i + \epsilon \quad (7)$$

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \quad (8)$$



Fuente: <https://allmodelsarewrong.github.io>

- 1 Recap
- 2 Modelos Lineales
- 3 Regularización
 - Lasso
 - Ridge
 - Elastic net
- 4 Break
- 5 Para seguir leyendo

Lasso

- Para un $\lambda \geq 0$ dado, consideremos el siguiente problema de optimización

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (9)$$

Lasso

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

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (9)$$

- ▶ “LASSO’s free lunch”: selecciona automáticamente los predictores que van en el modelo ($\beta_j \neq 0$) y los que no ($\beta_j = 0$)
- ▶ Porque? Los coeficientes que no van son soluciones de esquina
- ▶ $E(\beta)$ es no differentiable

Lasso Intuición en 1 Dimensión

► Lasso Intuición

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta| \quad (10)$$

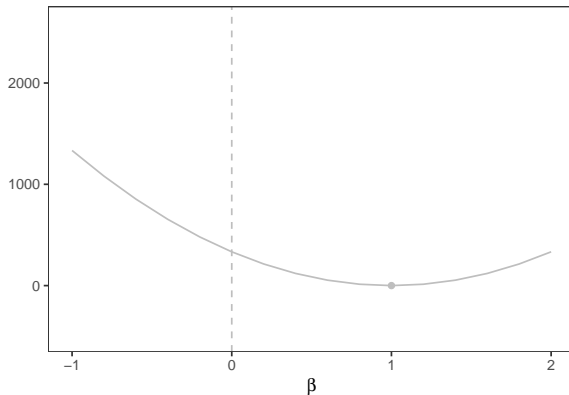
► Un solo predictor, un solo coeficiente

► Si $\lambda = 0$

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 \quad (11)$$

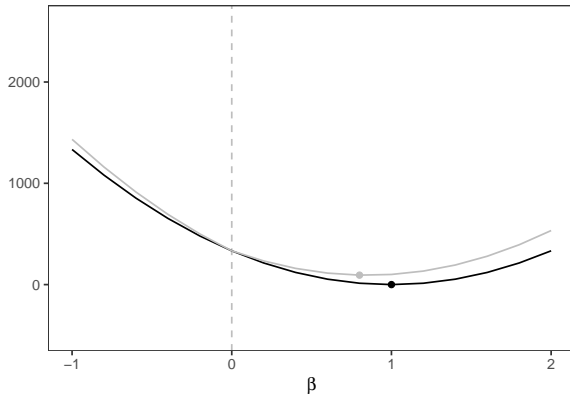
Intuición en 1 Dimension

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta| \quad (12)$$



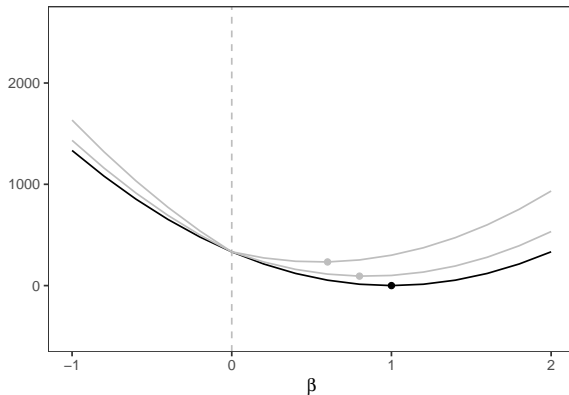
Intuición en 1 Dimension

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta| \quad (13)$$



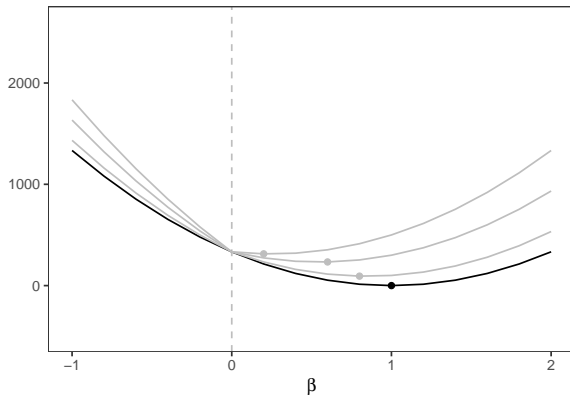
Intuición en 1 Dimension

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta| \quad (14)$$



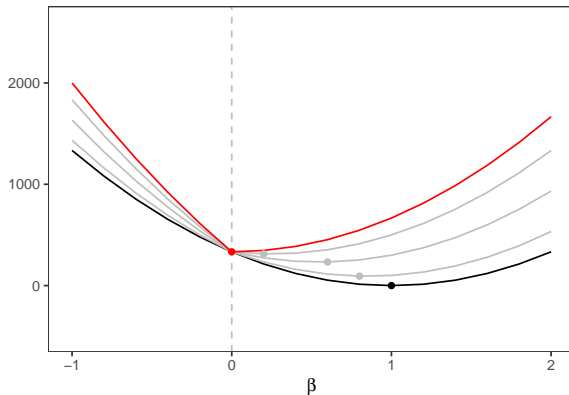
Intuición en 1 Dimension

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta| \quad (15)$$



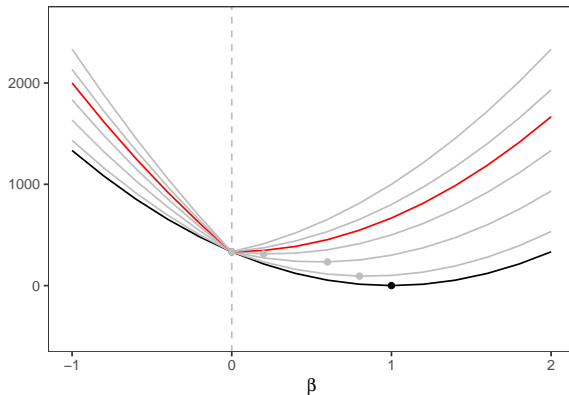
Intuición en 1 Dimension

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta| \quad (16)$$



Intuición en 1 Dimension

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta| \quad (17)$$



Intuición en 1 Dimension

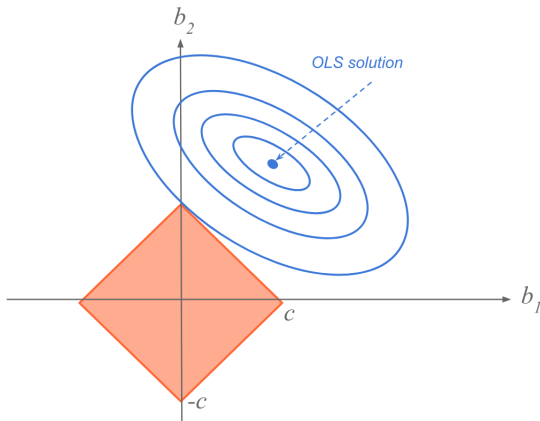
$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta| \quad (18)$$

la solución analítica es

$$\hat{\beta}_{lasso} = \begin{cases} 0 & \text{si } \lambda \geq \lambda^* \\ \hat{\beta}_{OLS} - \frac{\lambda}{2} & \text{si } \lambda < \lambda^* \end{cases} \quad (19)$$

Intuición en 2 Dimensiones (Lasso)

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \quad (20)$$



Ridge

- Para un $\lambda \geq 0$ dado, consideremos ahora el siguiente problema de optimización

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^p (\beta_j)^2 \quad (21)$$

- La intuición es similar a lasso, pero la vamos a extender a 2-Dim

Intuición en 2 Dimensiones (Ridge)

- Al problema

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^p (\beta_j)^2 \quad (22)$$

- podemos escribirlo como

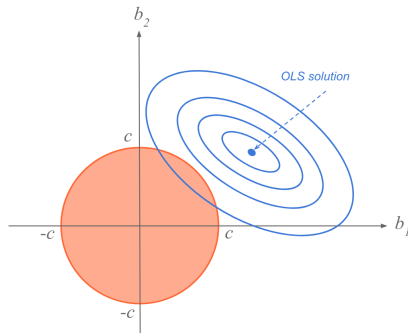
$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \quad (23)$$

sujeto a

$$((\beta_1)^2 + (\beta_2)^2) \leq c$$

Intuición en 2 Dimensiones (Ridge)

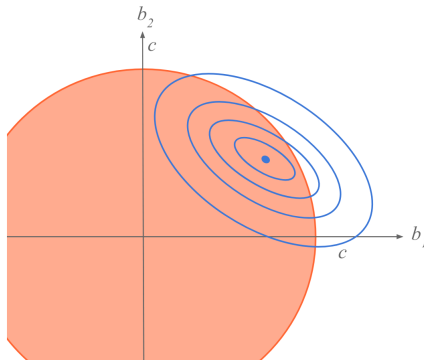
$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \text{ s.a. } ((\beta_1)^2 + (\beta_2)^2) \leq c \quad (24)$$



Fuente: <https://allmodelsarewrong.github.io>

Intuición en 2 Dimensiones (Ridge)

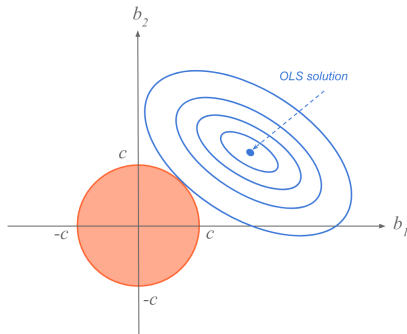
$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \text{ s.a. } ((\beta_1)^2 + (\beta_2)^2) \leq c \quad (25)$$



Fuente: <https://allmodelsarewrong.github.io>

Intuición en 2 Dimensiones (Ridge)

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \text{ s.a } ((\beta_1)^2 + (\beta_2)^2) \leq c \quad (26)$$



Fuente: <https://allmodelsarewrong.github.io>

Comentarios técnicos

- ▶ Lasso y ridge son sesgados, pero las disminuciones en varianza pueden compensar el sesgo y llevar a un MSE menor
- ▶ Lasso encoje a cero, Ridge no tanto
- ▶ Importante para aplicación:
 - ▶ Estandarizar los datos (media 0, y varianza 1)
 - ▶ Como elegimos λ ?
 - ▶ Utilizando validación cruzada

Elastic net

- ▶ Elastic Net es un happy medium

$$\min_{\beta} EL(\beta) = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 + \alpha\lambda \sum_{j=1}^p |\beta_j| + (1 - \alpha)\lambda \sum_{j=1}^p (\beta_j)^2 \quad (27)$$

- ▶ Si $\alpha = 1$ Lasso
- ▶ Si $\alpha = 0$ Ridge
- ▶ How to choose (α, λ) ? \rightarrow Crossvalidation Bidimensional

Break

Para seguir leyendo

- ▶ 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.
- ▶ Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).
- ▶ Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. CRC Press. (Chapters 2 & 6)