Lecture 7: Machine Learning Regularización

Big Data and Machine Learning en el Mercado Inmobiliario Educación Continua

Ignacio Sarmiento-Barbieri

Universidad de los Andes

October 25, 2022

Agenda

Objetivos

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

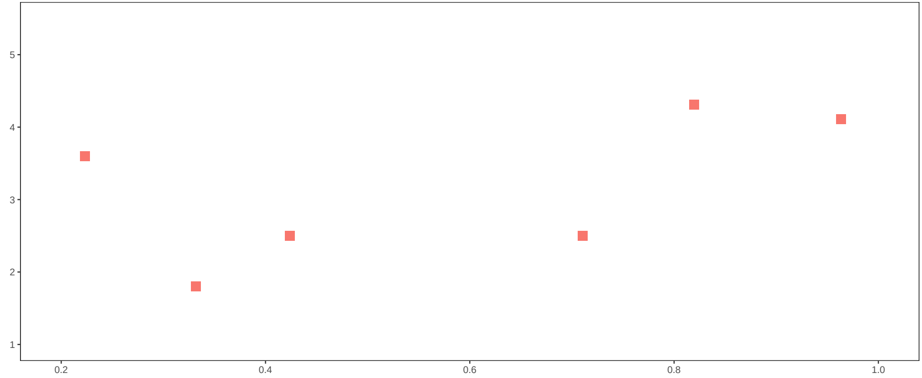
- Asumimos que el link entre y and X esta dado por:

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

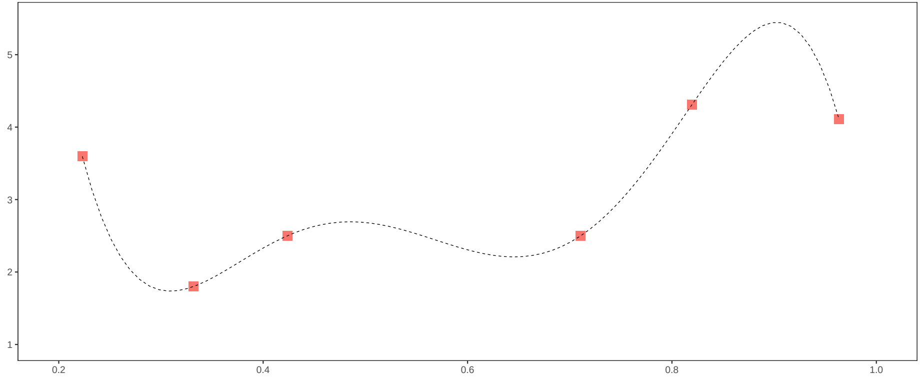
- donde $f(X)$ es cualquier función,

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

Overfit



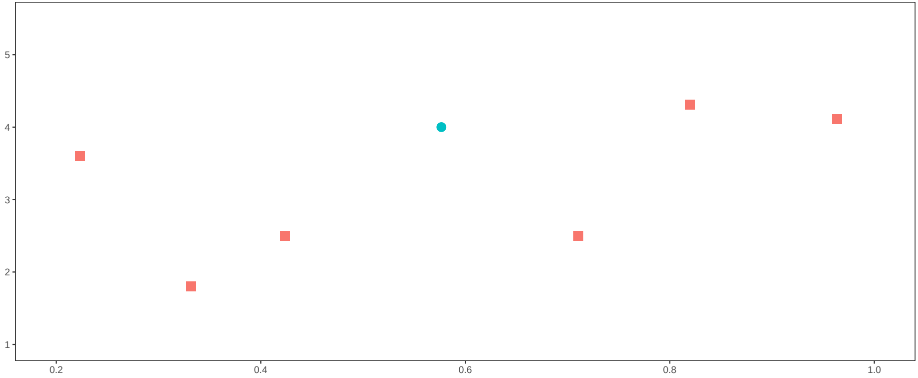
Overfit



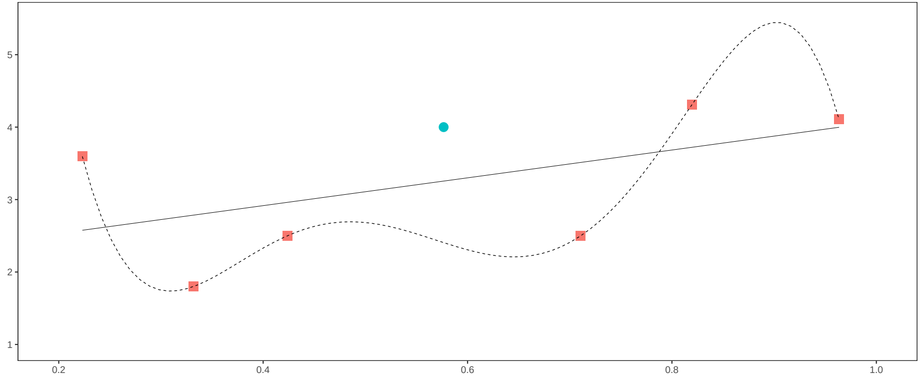
Overfit y Predicción fuera de Muestra

- 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)

Modelos Lineales

- El problema es:

$$y = f(X) + u \tag{2}$$

Modelos Lineales

- El problema es:

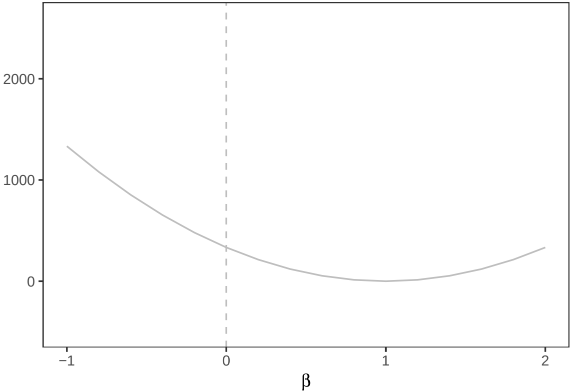
$$y = f(X) + u \tag{2}$$

- proponemos :

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{3}$$

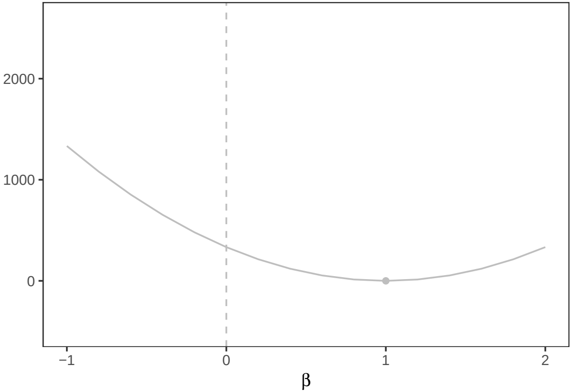
Intuición en 1 Dimension

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 \tag{4}$$



Intuición en 1 Dimension

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



Modelos Lineales

$$Precio = \beta_0 + \beta_1 Habitaciones + \epsilon \tag{6}$$

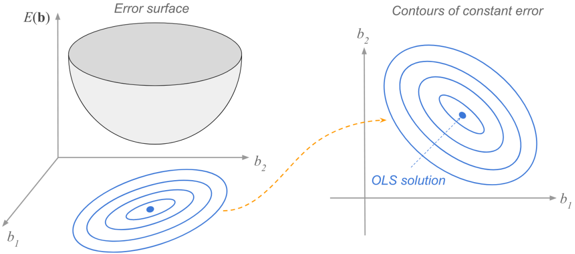
Modelos Lineales

$$Precio = \beta_0 + \beta_1 Habitaciones + \epsilon \tag{6}$$

```
with(test,mean((price-specification2)^2)):4.56335e+17
```

Intuición en 2 Dimensiones (OLS)

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \qquad (7)$$



Fuente: https://allmodelsarewrong.github.io

Modelos Lineales

$$Precio = \beta_0 + \beta_1 Habitaciones + \beta_2 Superficie + \epsilon \tag{8}$$

Modelos Lineales

$$Precio = \beta_0 + \beta_1 Habitaciones + \beta_2 Superficie + \epsilon \tag{8}$$

```
with(test,mean((price-specification3)^2):  4.512972e+17
```

$$Precio = \beta_0 + \beta_1 Habitaciones + \beta_2 Superficie + \beta_3 Dormitorios + \epsilon \tag{9}$$

```
with(test,mean((price-specification4)^2): 3.401889e+17
```

Lasso

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{10}$$

Lasso

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

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{10}$$

  - "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
  - $L(\beta)$ es no differentiable

Lasso Intuición en 1 Dimension

- Lasso Intuición

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

- Un solo predictor, un solo coeficiente
- Si $\lambda = 0$

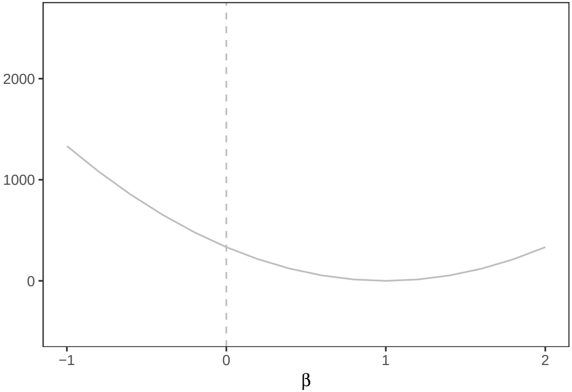
$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 \quad (12)$$

- y la solución es

$$\hat{\beta}_{OLS} \quad (13)$$

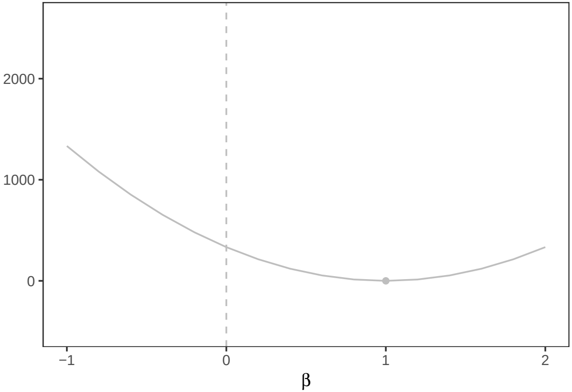
Intuición en 1 Dimension

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



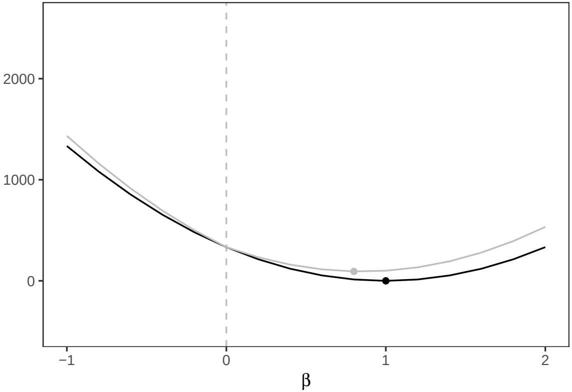
Intuición en 1 Dimension

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



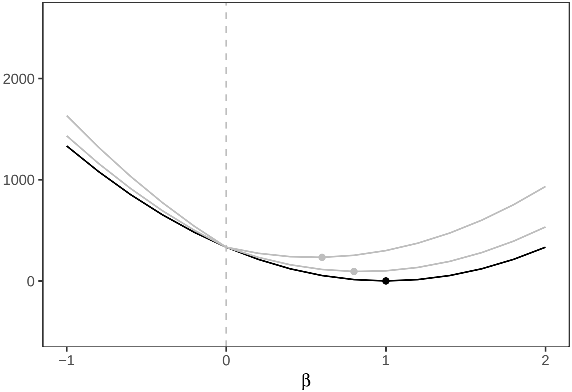
Intuición en 1 Dimension

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



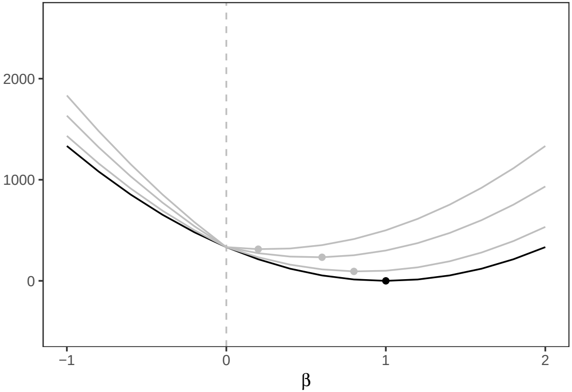
Intuición en 1 Dimension

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



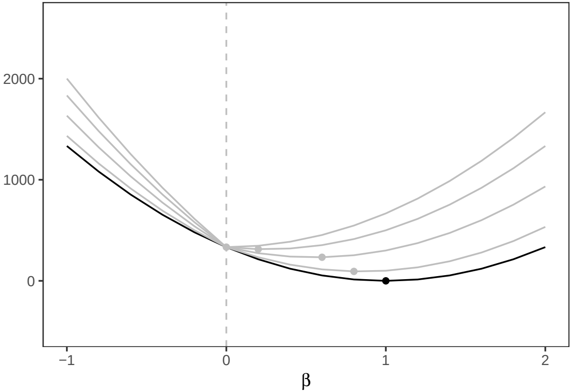
Intuición en 1 Dimension

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



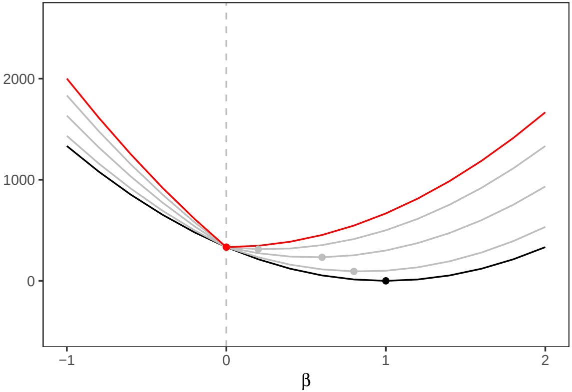
Intuición en 1 Dimension

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda|\beta| \tag{19}$$



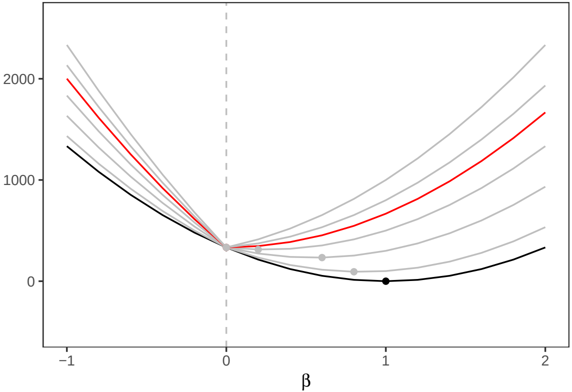
Intuición en 1 Dimension

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{20}$$



Intuición en 1 Dimension

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \quad (21)$$



Intuición en 1 Dimension

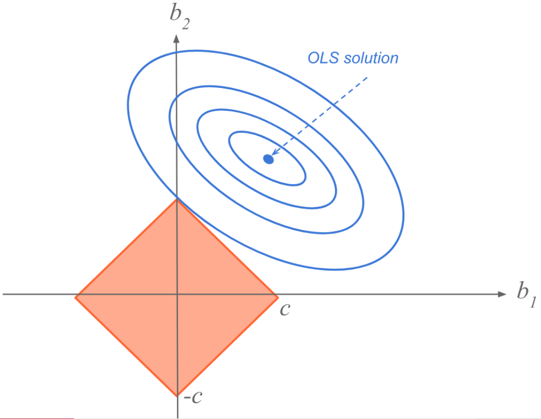
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{22}$$

la solución analítica es

$$\hat{\beta}_{lasso} = \begin{cases} 0 & \text{si } \lambda \geq \lambda^* \\ \hat{\beta}_{OLS} - \frac{\lambda}{2} & \text{si } \lambda < \lambda^* \end{cases} \tag{23}$$

Intuición en 2 Dimensiones (Lasso)

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



Ridge

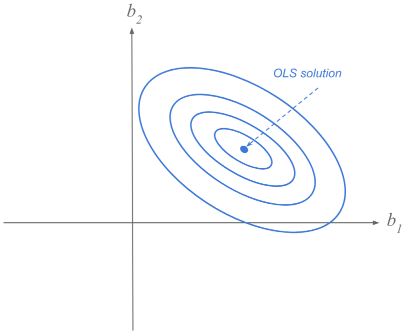
- Para un $\lambda \geq 0$ dado, consideremos ahora el siguiente problema de optimización

$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda\sum_{j=1}^{p}(\beta_j)^2 \qquad (25)$$

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

Intuición en 2 Dimensiones (OLS)

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \tag{26}$$



Fuente: https://allmodelsarewrong.github.io

Intuición en 2 Dimensiones (Ridge)

- Al problema

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda\sum_{j=1}^{p}(\beta_j)^2 \tag{27}$$

- podemos escribirlo como

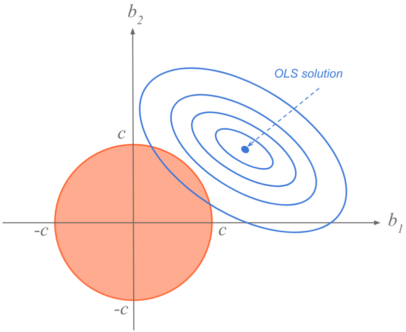
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \tag{28}$$

sujeto a

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

Intuición en 2 Dimensiones (Ridge)

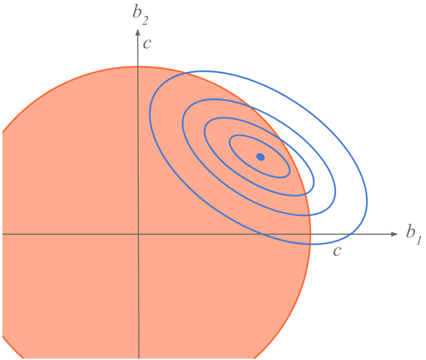
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } ((\beta_1)^2 + (\beta_2)^2) \leq c \tag{29}$$



Fuente: https://allmodelsarewrong.github.io

Intuición en 2 Dimensiones (Ridge)

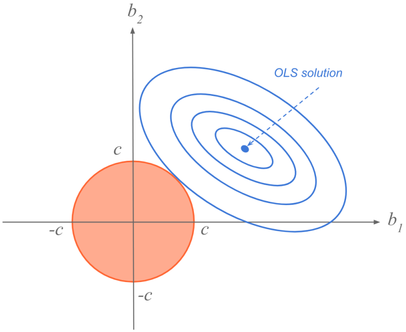
$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{30}$$



Fuente: https://allmodelsarewrong.github.io

Intuición en 2 Dimensiones (Ridge)

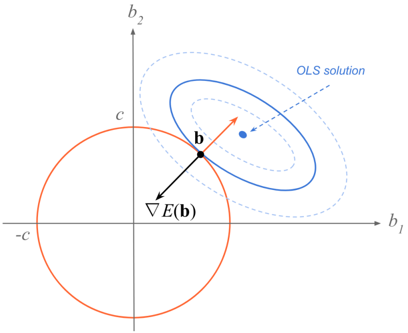
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{31}$$



Fuente: `https://allmodelsarewrong.github.io`

Intuición en 2 Dimensiones (Ridge)

$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{32}$$



Fuente: https://allmodelsarewrong.github.io

Comentarios técnicos

- Lasso y ridge son sesgados, pero las disminuciones en varianza pueden compensar estoy 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?

Comentarios técnicos: selección de λ

- λ es un parámetro y lo elegimos usando validación cruzada
  1. Partimos la muestra de entrenamiento en K Partes: $M_{train} = M_{fold\,1} \cup M_{fold\,2} \cdots \cup M_{fold\,K}$
  2. Cada conjunto $M_{fold\,K}$ va a jugar el rol de una muestra de evaluación $M_{eval\,k}$. Entonces para cada muestra
     - $M_{train-1} = M_{train} - M_{fold\,1}$
     - $\vdots$
     - $M_{train-k} = M_{train} - M_{fold\,k}$
  3. Luego hacemos el siguiente loop
     1. Para $\lambda_i = 0, 0.001, 0.002, \ldots, \lambda_{max}$
        - Para $k = 1, \ldots, K$
          - Ajustar el modelo $m_{i,k}$ con λ_i en $M_{train-k}$
          - Calcular y guardar el $MSE(m_{i,k})$ usando M_{eval-k}
        - fin para k
        - Calcular y guardar $MSE_i = \frac{1}{K} MSE(m_{i,k})$
     2. fin para λ
  4. Encontrar el menor MSE_i y usar ese $\lambda_i = \lambda^*$

Elastic net

- Elastic Net es un happy medium

$$min_{\beta} EL(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \alpha\lambda \sum_{j=1}^{p} |\beta_j| + (1-\alpha)\lambda \sum_{j=1}^{p} (\beta_j)^2 \qquad (33)$$

- Si $\alpha = 1$ Lasso
- Si $\alpha = 0$ Rigdge
- 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)