Métodos de regularización en regresión lineal

Victor Daniel Alvarado Estrella

19 de mayo de 2021

1. Introducción

En el modelo de regresión lineal, suponemos que hay una variable de respuesta y que puede ser descrita a través de una combinación lineal de variables explicativas $X_1, X_2, \ldots, X_p$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, 2, \ldots, n. \tag{1}$$

Aquí, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ son los (desconocidos) coeficientes de regresión y ϵ_i son errores aleatorios independientes e idénticamente distribuidos con distribución $N(0, \sigma^2)$. La ecuación (1) puede ser reescrita en forma matricial como

$$y = X\beta + \epsilon \tag{2}$$

donde

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

1.1. Mínimos cuadrados ordinarios

Bajo las suposiciones anterior mencionadas, los coeficientes de regresión pueden ser estimados minimizando la suma de los residuos cuadráticos

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip}\right)^2,$$

que en su forma matricial se escribe como

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2. \tag{3}$$

Puede demostrarse que la solución de (3) está dada por

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{4}$$

1.2. Multicolinealidad

Cuando hay presencia de multicolinealidad en los datos, es decir, cuando hay una correlación alta entre las variables explicativas, los estimadores por mínimos cuadrados se vuelven mal condicionados. Cuando esto ocurre, pequeños cambios en los datos pueden producir grandes cambios en los coeficientes estimados. Otro problema con la multicolinealidad es que los coeficientes de regresión tienden a tener varianzas muy grandes, haciendo la estimación pobre. Además, existe el riesgo de *overfitting*, en donde si bien el modelo ajusta bien a los datos, no sirve para predecir datos adicionales.

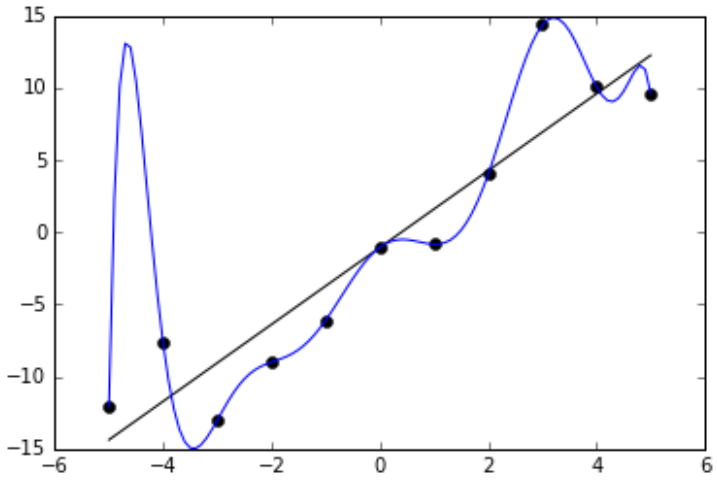

Figura 1: Overfitting.

Existen distintos métodos para detectar si hay presencia de multicolinealidad en los datos. Entre ellos

- Calcular el número de condición de la matriz $X^T X$. Un número de condición alto indica la presencia de multicolinealidad.
- Construir la matriz de correlación de las variables explicativas. Valores cercanos a 1 o a -1 fuera de la diagonal indican una correlación alta entre los regresores.

Cuando esto ocurre, existen métodos de estimación alternativos a mínimos cuadrados.

2. Regresión Ridge

De manera similar a mínimos cuadrados, la regresión Ridge minimiza la norma al cuadrado del vector de residuos. Sin embargo, adicionalmente, Ridge penaliza sumando un múltiplo de la norma al cuadrado del vector de coeficientes

$$\hat{\beta} = \arg\min_{\beta} \left(\|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right). \tag{5}$$

Podemos reescribir la función objetivo como

$$\begin{aligned} f(\beta) &= \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \\ &= (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta \\ &= y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta + \lambda\beta^T\beta. \end{aligned}$$

Calculando el gradiente de la función objetivo obtenemos que

$$\nabla f(\beta) = -2X^Ty + 2X^TX\beta + 2\lambda\beta.$$

Igualando el gradiente a cero obtenemos las ecuaciones normales del estimador Ridge

$$(X^TX + \lambda I)\beta = X^Ty.$$

Por lo tanto, la solución de (5) está dada por

$$\hat{\beta} = (X^TX + \lambda I)^{-1}X^Ty. \tag{6}$$

3. Regresión LASSO

De manera similar a mínimos cuadrados, la regresión LASSO (Least Absolute Shrinkage and Selection Operator) minimiza la norma al cuadrado del vector de residuos. Sin embargo, al igual que Ridge, LASSO penaliza sumando una cierta cantidad. En el caso de LASSO, la penalización es un múltiplo de la norma 1 del vector de coeficientes

$$\hat{\beta} = \arg\min_{\beta} \left(\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right). \tag{7}$$

No obstante, la ecuación (7) no tiene solución analítica, sino que debe resolverse de manera numérica. Además, puesto que la función objetivo no es diferenciable en algunos puntos, concretamente cuando algún $\beta_i = 0$, no podemos utilizar los algoritmos de optimización que requieren el cálculo del gradiente. Una alternativa es utilizar *descenso por coordenadas*.

Descenso por coordenadas

Entrada: Una función objetivo $f : \mathbb{R}^m \to \mathbb{R}$, un punto inicial $x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_m^{(0)}\right)$, una tolerancia τ y un número máximo de iteraciones N.

Para $k = 1, 2, \ldots, N$ hacer:

1. $x^{(k)} = x^{(k-1)}$.
2. Para $i = 1, 2, \ldots, m$ resolver cada problema de minimización unidimensional

$$x_i^{(k)} = \arg\min_{y} f\left(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, y, x_{i+1}^{(k)}, \ldots, x_m^{(k)}\right)$$

3. Si $\left\|x^{(k)} - x^{(k-1)}\right\| < m\tau$, terminar y devolver $x^{(k)}$.

Ahora bien, en el caso de LASSO, podemos tomar como punto inicial el estimador Ridge

$$\beta^{(0)} = (X^T X + \lambda I)^{-1} X^T y. \tag{8}$$

Por otra parte, denotemos por X_{-i} a la matriz X cuando se elimina la i-ésima columna, β_{-i} al vector de coeficientes β cuando se elimina el i-ésimo coeficiente, y X_i la i-ésima columna de la matriz X. La función objetivo puede reescribirse como

$$\begin{aligned} f(\beta) &= \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \\ &= \|y - X_{-i}\beta_{-i} - X_i\beta_i\|_2^2 + \lambda(\|\beta_{-i}\|_1 + |\beta_i|) \\ &= (y - X_{-i}\beta_{-i} - X_i\beta_i)^T (y - X_{-i}\beta_{-i} - X_i\beta_i) + \lambda\|\beta_{-i}\|_1 + \lambda|\beta_i| \\ &= (y - X_{-i}\beta_{-i})^T (y - X_{-i}\beta_{-i}) - 2\beta_i X_i^T (y - X_{-i}\beta_{-i}) + \beta_i^2 X_i^T X_i + \lambda\|\beta_{-i}\|_1 + \lambda|\beta_i|. \end{aligned}$$

Derivando la función objetivo respecto a β_i, cuando $\beta_i \neq 0$, obtenemos que

$$\frac{\partial f}{\partial \beta_i} = -2X_i^T (y - X_{-i}\beta_{-i}) + 2\beta_i X_i^T X_i + \lambda \operatorname{sgn}(\beta_i).$$

Igualando la derivada a cero y despejando β_i obtenemos que

$$\beta_i = \frac{X_i^T (y - X_{-i}\beta_{-i}) - \frac{\lambda}{2}\operatorname{sgn}(\beta_i)}{X_i^T X_i}.$$

Tenemos entonces los siguientes casos. Si $\beta_i > 0$ entonces

$$\beta_i = \frac{X_i^T (y - X_{-i}\beta_{-i}) - \frac{\lambda}{2}}{X_i^T X_i} > 0,$$

de donde $X_i^T (y - X_{-i}\beta_{-i}) > \frac{\lambda}{2}$. Si $\beta_i < 0$ entonces

$$\beta_i = \frac{X_i^T (y - X_{-i}\beta_{-i}) + \frac{\lambda}{2}}{X_i^T X_i} < 0,$$

de donde $X_i^T (y - X_{-i}\beta_{-i}) < -\frac{\lambda}{2}$.

Por último, consideremos el caso en donde $-\frac{\lambda}{2} \leq X_i^T (y - X_{-i}\beta_{-i}) \leq \frac{\lambda}{2}$. Tenemos que

$$\begin{aligned} \frac{\partial f}{\partial \beta_i} &= -2X_i^T (y - X_{-i}\beta_{-i}) + 2\beta_i X_i^T X_i + \lambda \operatorname{sgn}(\beta_i) \\ &= -2\left[X_i^T (y - X_{-i}\beta_{-i}) - \frac{\lambda}{2}\operatorname{sgn}(\beta_i)\right] + 2\beta_i X_i^T X_i, \end{aligned}$$

de donde vemos que $\frac{\partial f}{\partial \beta_i} > 0$, de modo que la función objetivo es creciente, para $\beta_i > 0$; mientras que $\frac{\partial f}{\partial \beta_i} < 0$, de modo que la función objetivo es decreciente, para $\beta_i < 0$. Así pues, la función objetivo alcanza su mínimo sobre la dirección β_i en $\beta_i = 0$.

En resumen, la solución al problema de minimización LASSO unidimensional está dada por

$$\beta_i^{(k)} = \begin{cases} \frac{X_i^T\left(y - X_{-i}\beta_{-i}^{(k)}\right) - \frac{\lambda}{2}}{X_i^T X_i}, & \text{si } X_i^T\left(y - X_{-i}\beta_{-i}^{(k)}\right) > \frac{\lambda}{2}, \\ \frac{X_i^T\left(y - X_{-i}\beta_{-i}^{(k)}\right) + \frac{\lambda}{2}}{X_i^T X_i}, & \text{si } X_i^T\left(y - X_{-i}\beta_{-i}^{(k)}\right) < -\frac{\lambda}{2}, \\ 0, & \text{si } -\frac{\lambda}{2} \leq X_i^T\left(y - X_{-i}\beta_{-i}^{(k)}\right) \leq \frac{\lambda}{2}. \end{cases} \tag{9}$$

Por lo tanto, el algoritmo de descenso por coordenadas para LASSO es el siguiente

Descenso por coordenadas (LASSO)

Entrada: Los datos X, y, una tolerancia τ y un número máximo de iteraciones N.

Inicializar $\beta^{(0)} = (X^T X + \lambda I)^{-1} X^T y$.

Para $k = 1, 2, \ldots, N$ hacer:

1. $\beta^{(k)} = \beta^{(k-1)}$.
2. Para $i = 0, 1, \ldots, p$ hacer:

$$\beta_i^{(k)} = \begin{cases} \frac{X_i^T\left(y - X_{-i}\beta_{-i}^{(k)}\right) - \frac{\lambda}{2}}{X_i^T X_i}, & \text{si } X_i^T\left(y - X_{-i}\beta_{-i}^{(k)}\right) > \frac{\lambda}{2}, \\ \frac{X_i^T\left(y - X_{-i}\beta_{-i}^{(k)}\right) + \frac{\lambda}{2}}{X_i^T X_i}, & \text{si } X_i^T\left(y - X_{-i}\beta_{-i}^{(k)}\right) < -\frac{\lambda}{2}, \\ 0, & \text{si } -\frac{\lambda}{2} \leq X_i^T\left(y - X_{-i}\beta_{-i}^{(k)}\right) \leq \frac{\lambda}{2}. \end{cases}$$

3. Si $\left\|\beta^{(k)} - \beta^{(k-1)}\right\| < p\tau$, terminar y devolver $\beta^{(k)}$.

3.1. Comparación entre Ridge y LASSO

Notemos que las ecuaciones (5) y (7) que definen a los estimadores Ridge y LASSO se obtienen al llevar a su forma de multiplicadores de Lagrange los siguientes problemas de minimización con restricción

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \|y - X\beta\|_2^2 \qquad \text{sujeto a } \|\beta\|_2^2 \leq c^2, \tag{10}$$

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \|y - X\beta\|_2^2 \qquad \text{sujeto a } \|\beta\|_1 \leq c. \tag{11}$$

De esta manera, evitamos que los coeficientes de regresión tomen valores muy grandes. Además, cabe mencionar que por la forma de las funciones de restricción, LASSO logra que algunos coeficientes sean cero, mientras que con Ridge esto por lo general no ocurre. En determinadas ocasiones, esto puede llegar a ser realmente útil, por ejemplo, cuando no todas las variables explicativas son importantes y se desea excluir las menos influyentes. No obstante, los estimadores Ridge son generalmente más estables que los estimadores LASSO.

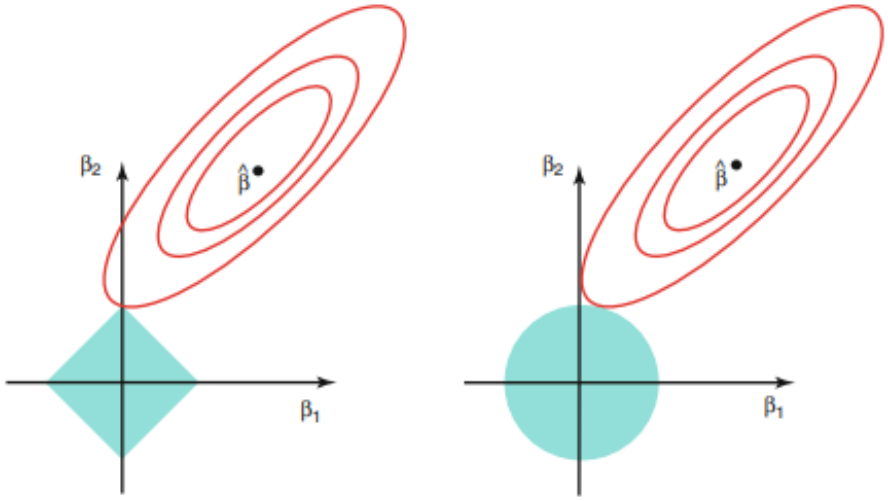

Figura 2: Contornos de la función objetivo y la función de restricción para la regresión LASSO (izquierda) y regresión Ridge (derecha). Las áreas en azul son las regiones de restricción $|\beta_1|+|\beta_2| \leq c$ y $\beta_1^2 + \beta_2^2 \leq c^2$, mientras que las elipses en rojo son los contornos de la suma de residuos cuadráticos.

4. Validación cruzada

Hasta el momento hemos visto como estimar los coeficientes de regresión Ridge y LASSO en función del parámetro de regularización λ. Sin embargo, aún no es claro cómo escoger el valor de λ. Un método para esto es validación cruzada.

La idea de validación cruzada consiste en dividir el conjunto de datos en dos subconjuntos ajenos, ajustar el modelo utilizando alguno de los dos subconjuntos (denominado conjunto de entrenamiento o *training set*), y evaluar el desempeño del modelo utilizando el otro subconjunto (denominado conjunto de validación o *test set*). De esta manera, evitamos el problema de *overfitting*. Existen varios tipos de validación cruzada, entre ellos se encuentra la validación cruzada de K iteraciones.

Validación cruzada de K iteraciones
Entrada: Los datos X, y, un modelo a ajustar m, una función de *score* s y un número de iteraciones K.
Particionar los datos en K subconjuntos del mismo tamaño: $$X = X_1 \cup X_2 \cup \ldots \cup X_K$$ $$y = y_1 \cup y_2 \cup \ldots \cup y_K.$$ Para $i = 1, 2, \ldots, K$ hacer: 1. Definir el i-ésimo subconjunto como conjunto de validación: $$X_{test} = X_i, \quad y_{test} = y_i.$$

2. Definir los restantes $K-1$ subconjuntos como conjunto de entrenamiento:
$$X_{train} = X \setminus X_i, \quad y_{train} = y \setminus y_i.$$
3. Estimar los parámetros del modelo utilizando el conjunto de entrenamiento:
$$\hat{\beta} = m.fit(X_{train}, y_{train}).$$
4. Evaluar el desempeño del modelo utilizando el conjunto de validación:
$$\hat{y}_{test} = m.predict(X_{test}), \quad s_i = s(y_{test}, \hat{y}_{test}).$$
Devolver el promedio de los scores calculados: $\frac{1}{K}\sum_{i=1}^{K} s_i$.

En el caso de regresión, podemos tomar como función de *score* el error cuadrático medio

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . \tag{12}$$

De esta manera, para escoger un valor de λ adecuado, podemos proponer un conjunto de valores candidatos $\lambda_1, \lambda_2, \ldots, \lambda_N$. Posteriormente, para cada λ_i, calculamos el *score* del modelo mediante validación cruzada y utilizando el error cuadrático medio. Al final, seleccionamos el valor de λ que obtuvo el menor *score* de entre todos.

5. Ejemplo (Porcentaje de grasa en puercos)

La medición del porcentaje de grasa (FAT) en puercos es un procedimiento costoso, por lo cual es importante investigar si este porcentaje se puede predecir a partir de otras propiedades del puerco de fácil medición. Estas variables son:

- $AVBF$ es un promedio de tres mediciones del grosor de grasa en la espalda;
- MUS es una puntuación de musculatura para la carcasa. Entre mayor sea este número, hay más músculo y menos grasa;
- LEA es una medición del área del lomo;
- DEP es un promedio de tres mediciones de la profundidad de la grasa frente a la décima costilla;
- LWT es el peso vivo de la carcasa;
- CWT es el peso de la carcasa sacrificada;
- $WTWAT$ es una medida usada para determinar la gravedad especifica;
- $DPSL$ es el promedio de tres determinaciones de la profundidad del vientre;
- $LESL$ es la medida promedio de la delgadez de tres secciones transversales del vientre;
- $BELWT$ es el peso total del vientre.

El número de condición de la matriz X^TX es 1.982×10^7, un número demasiado grande. Por otra parte, si construimos la matriz de correlación de los predictores, notaremos que algunos de ellos están fuertemente correlacionados entre sí. Por lo cual, hay evidencia de la presencia de multicolinealidad en los datos.

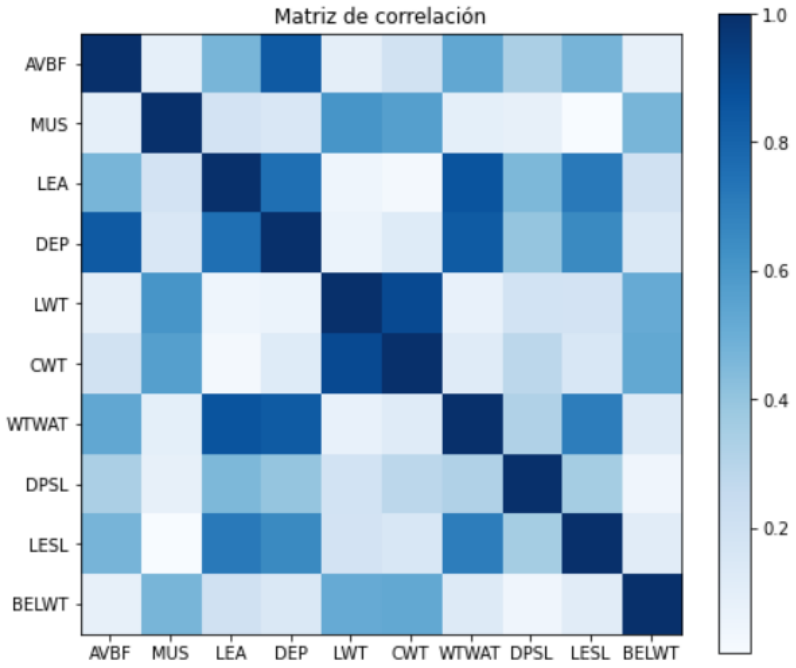

Figura 3: Gráfica de la matriz de correlación. Los coeficientes de correlación están en valor absoluto.

Así pues, para los datos del porcentaje de grasa en puercos se ajustaron los modelos de regresión Ridge y LASSO por validación cruzada de $K = 9$ iteraciones. Como valores candidatos de λ, se utilizaron 50 puntos logarítmicamente espaciados entre 1 y 100. En la siguiente tabla se muestra el valor de λ seleccionado por validación cruzada así como el *score* obtenido para ese valor.

	Regresión Ridge	Regresión LASSO
λ	3.089	4.498
score	9.788	9.448

Adicionalmente, en la siguiente tabla se muestra el coeficiente de determinación[1] y el error cuadrático medio sobre todo el conjunto de datos.

	Regresión Ridge	Regresión LASSO
Coeficiente de determinación	0.741	0.763
Error cuadrático medio	6.339	5.798

[1]El coeficiente de determinación mide la proporción de puntos que se pueden explicar por el modelo y se calcula como $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Finalmente, los coeficientes de regresión estimados fueron los siguientes

	Regresión Ridge	Regresión LASSO
Intercepto	0.458	0
AVBF	0.559	0
MUS	-0.98	-0.88
LEA	-0.527	0
DEP	2.478	5.394
LWT	0.237	0.21
CWT	0.024	0.023
WTWAT	-2.047	-2.084
DPSL	0.484	0
LESL	-0.587	-0.477
BELWT	1.185	1.177

En la Figura 4 se muestra la gráfica de los coeficientes de regresión Ridge y LASSO en función del parámetro de regularización λ. En la misma figura se muestra el valor de λ seleccionado por validación cruzada como una línea vertical de color negro.

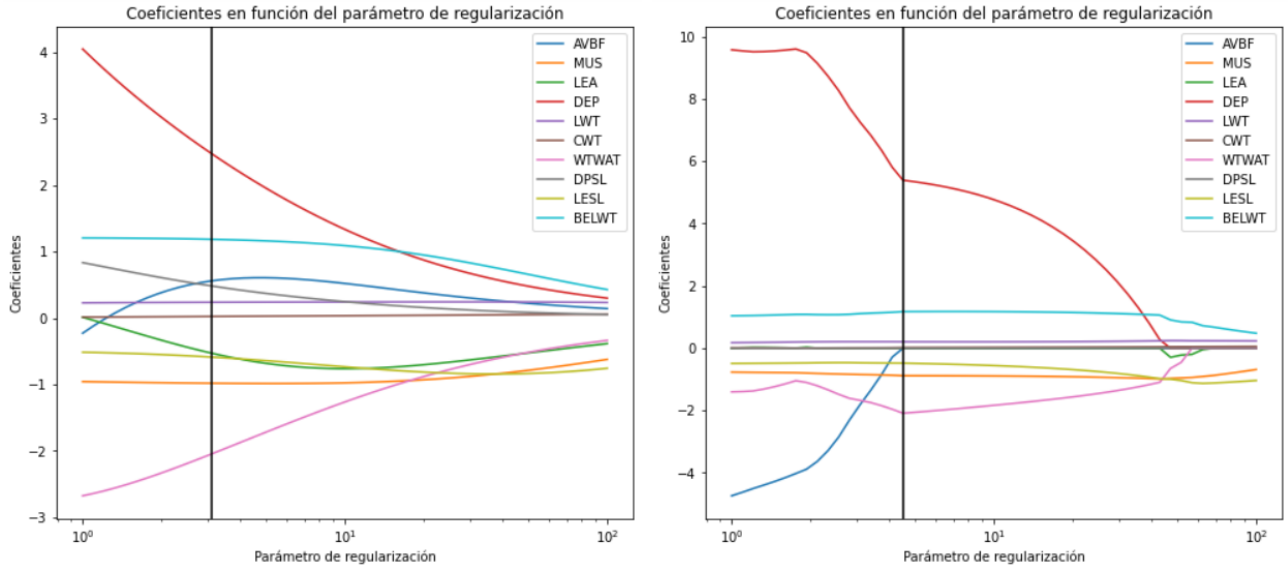

Figura 4: Gráfica de los coeficientes de regresión Ridge (izquierda) y LASSO (derecha) en función del parámetro de regularización λ.

Referencias

[Rod] Joaquín Amat Rodrigo. *Regularización Ridge, Lasso y Elastic Net con Python*. URL: https://www.cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python.html. (Consultado: 19.05.2021).

[Wie] Wessel van Wieringen. *Lasso regression*. URL: http://www.few.vu.nl/~wvanwie/Courses/HighdimensionalDataAnalysis/WNvanWieringen_HDDA_Lecture56_LassoRegression_20182019.pdf. (Consultado: 19.05.2021).

[Wika] Wikipedia. *Multicollinearity*. URL: https://en.wikipedia.org/wiki/Multicollinearity. (Consultado: 19.05.2021).

[Wikb] Wikipedia. *Ridge regression*. URL: https://en.wikipedia.org/wiki/Ridge_regression. (Consultado: 19.05.2021).

[Wikc] Wikipedia. *Validación cruzada*. URL: https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada. (Consultado: 19.05.2021).