Cuadrados mínimos con regularización e interpolación de Legendre

Tomás Bossi
tomasbossi97@gmail.com

Francisco Valentini
ft.valentini@gmail.com

Jazmín Vidal
jazmin.vidald@gmail.com

Resumen

Este informe aborda el problema de mínimos cuadrados lineales y su extensión a través de dos métodos: la regresión polinomial y la regularización. En primer lugar, presentamos la formulación del problema de mínimos cuadrados lineales, la técnica de polinomios de Legendre para implementar la regresión polinomial, y la regularización L2 o Ridge como estrategia para reducir la varianza del método. Luego, desarrollamos cómo implementar la solución de mínimos cuadrados con y sin regularización mediante la descomposición SVD. Finalmente, proponemos una aplicación práctica en la que buscamos el grado de polinomio y el valor de regularización que optimizan la capacidad de generalizar de un modelo lineal para un conjunto de datos dado.

1 Introducción

Dado un conjunto de n observaciones para dos variables (x_i, y_i), $i = 1, \ldots, n$, el problema de mínimos cuadrados lineales consiste en hallar una función o modelo lineal $f(x) = \beta_0 + \beta_i x$ que minimice la suma de los cuadrados de los errores $r_i = y_i - f(x_i)$, es decir,

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} r_i^2 = \min_{\beta_0,\beta_1} \|r\|_2^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

donde $r = [r_1, \ldots, r_n]$ es el vector de residuos y $\|\cdot\|_2$ es la norma euclidiana [Watkins, 2004]. El resultado de esta minimización es una función lineal que "ajusta bien" a los datos, en el sentido de que los errores $(y_i - f(x_i))^2$ tienden a ser pequeños.

Matricialmente, este problema se puede representar como:

$$\min_{\beta} \|r\|_2^2 = \min_{\beta} \|X\beta - y\|_2^2, \tag{1}$$

donde $y = [y_1, \ldots, y_n]$ es un vector de datos observados, $\beta = [\beta_0, \beta_1]$ es el vector de coeficientes o incógnitas a obtener y $X \in \mathbb{R}^{n \times 2}$ está dada por

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

El modelo lineal se puede extender añadiendo variables adicionales elevando la variable original x a una potencia. Por ejemplo, una regresión cúbica usa tres variables, x, x^2, y x^3, además de y. Este enfoque se conoce como regresión polinomial y es una manera sencilla de ajustar datos de manera no lineal pero resolviendo el problema de cuadrados mínimos lineales [James et al., 2013].

Si decidimos aproximar nuestros datos con un polinomio de grado d, entonces el problema consiste en buscar $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d$ que minimice la suma de los cuadrados de los errores. En el problema

matricial 1, entonces, buscamos $\beta \in \mathbb{R}^{d+1}$ y X se define como

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^d \end{bmatrix}. \tag{2}$$

En este trabajo asumimos que hay más observaciones que coeficientes a obtener ($n > d + 1$), de modo que el vector de soluciones β del problema de mínimos cuadrados es único [Watkins, 2004]. En particular, el vector es tal que satisface

$$X^T X\beta = X^T y, \tag{3}$$

y está dado por

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

La matriz $X^T X$ del sistema 3 puede estar mal condicionada, lo que puede llevar a errores numéricos. Esto es particularmente cierto en el caso de la regresión polinomial y cuando d es grande porque las columnas $1, x, x^2, \dots, x^d$ no son ortogonales [Strang, 2022]. Para polinomios de grado diez, por ejemplo, es prácticamente imposible resolver el sistema 3 mediante eliminación gaussiana porque cada error de redondeo se amplifica en más de 10^{13} [Strang, 2022].

Una manera de evitar este problema es usar polinomios de Legendre en lugar de los polinomios estándar. El procedimiento consiste en encontrar combinaciones de $1, x, x^2, \dots, x^d$ que sean ortogonales entre sí mediante la técnica de Gram-Schmidt [Strang, 2022]. De esta manera obtenemos una matriz X ortogonal y, por ende, la matriz $X^T X$ está bien condicionada. Los polinomios construidos de esta manera se llaman polinomios de Legendre y son ortogonales entre sí en el intervalo $-1 \leq x \leq 1$ [Strang, 2022].

Cuando la relación entre y y las variables $x, x^2, \dots, x^d$ es efectivamente cercana a lineal, los coeficientes β que resuelven 3 pueden tener varianza alta [James et al., 2013]. Esto quiere decir que un cambio pequeño en los datos con los que se ajusta el modelo, pero que surjan del mismo proceso generador de datos, puede generar un cambio grande en los coeficientes que se obtienen. Esto es particularmente cierto cuando d es casi tan grande como n. En este caso, el modelo se ajusta demasiado bien a los datos con los que se obtuvieron los coeficientes y no generaliza bien a datos nuevos [James et al., 2013].

Una manera de reducir la varianza es implementar un ajuste que tienda a restringir los valores de β a valores pequeños. Esto se puede lograr añadiendo un término de penalización al problema de minimización 1, de modo que el problema se convierta en

$$\min_{\beta} \|X\beta - y\|_2^2 + \lambda\|\beta\|_2^2 = \min_{\beta_0,\beta_1,\dots,\beta_d} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \dots - \beta_d x_i^d)^2 + \lambda \sum_{j=1}^{d} \beta_j^2, \tag{5}$$

donde $\lambda \geq 0$ es el parámetro que el tamaño de la penalización, y β_0 se excluye de $\|\beta\|_2^2$ porque no se penaliza. Este procedimiento se conoce como regularización L2 o Ridge [James et al., 2013]. La solución en este caso es

$$\beta = (X^T X + \lambda I)^{-1} X^T y, \tag{6}$$

donde I es la matriz identidad de dimensión $d + 1 \times d + 1$.

Cuando $\lambda = 0$, la regularización no tiene efecto y obtenemos los mismos coeficientes que en la expresión 4. A medida que λ aumenta, los coeficientes β se hacen más pequeños y la varianza se reduce. En el límite, cuando $\lambda \to \infty$, los coeficientes β tienden a cero y la varianza es mínima [James et al., 2013].

Para hallar un modelo lineal con un nivel de varianza adecuado, entonces, podemos ajustar el valor del parámetro λ. Una manera de optimizar este valor consiste en usar un conjunto de datos de validación. Esto implica particionar aleatoriamente el conjunto de datos disponibles en dos partes: un conjunto de entrenamiento y otro conjunto de validación. Para cada valor de λ que se quiera probar, se halla β con la expresión 6 usando el conjunto de entrenamiento y se evalúa el error en el conjunto de validación con una medida apropiada. Una cantidad ampliamente usada para medir el error es el error cuadrático medio (ECM), que se obtiene como

$$ECM = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2, \tag{7}$$

donde $f(x_i)$ es el valor ajustado para el dato de validación x_i con los coeficientes β obtenidos con el conjunto de entrenamiento.

Así como optimizamos el valor de λ, también podemos optimizar el grado d del polinomio. Para esto, podemos usar el mismo procedimiento de validación, pero variando también el grado del polinomio. Si fijamos posibles valores de d y λ y evaluamos el ECM de validación para cada combinación posible, podemos hallar el par de valores que optimiza la capacidad predictiva del modelo, es decir, la capacidad de generalizar. Los parámetros λ y d se conocen como hiperparámetros porque no se estiman directamente a partir de los datos, sino que se ajustan mediante una estrategia de validación.

1.1 Objetivos del trabajo

Los objetivos del presente trabajo son:

1. Implementar la solución al problema de cuadrados mínimos lineales a partir de la descomposición SVD, con y sin regularización
2. Mostrar una aplicación de regresión polinomial con regularización. La misma consiste en hallar el grado de polinomio y el valor de regularización que minimiza el error de validación para un conjunto de datos predeterminado.

La organización del trabajo se estructura de la siguiente manera: en la sección 2 describimos las implementaciones de la solución a mínimos cuadrados mediante SVD con y sin regularización (sección 2.1), y presentamos el experimento de búsqueda de hiperparámetros de regresión polinomial con regularización (sección 2.2). En la sección 3 presentamos los resultados del experimento. Finalmente, concluimos el trabajo en la sección 4.

2 Desarrollo

En la primera subsección desarrollamos cómo se puede implementar la solución a mínimos cuadrados mediante SVD con y sin regularización. En la segunda subsección describimos el experimento de búsqueda de hiperparámetros de regresión polinomial con regularización.

2.1 Algoritmos de cuadrados mínimos

En el sistema de ecuaciones normales 3, la matriz $X^T X$ puede estar mal condicionada, lo que puede llevar a errores numéricos. Una forma de mitigar este problema es usar la descomposición SVD de X para resolver el sistema [Strang, 2022].

Si conocemos la descomposición SVD de X tal que $X = U\Sigma V^T$, entonces podemos expresar la solución de mínimos cuadrados 4 como

$$\beta = (X^T X)^{-1} X^T y = V\Sigma^{-1} U^T y \tag{8}$$

para el caso en que X tiene rango completo. El algoritmo 1 muestra cómo implementar esta solución. Asumimos que tenemos a disposición una función SVD que calcula la descomposición SVD de una matriz X, devolviendo las matrices U y V^T y un vector $valores_singulares$ con los valores singulares de X. También asumimos que tenemos una función $\mathrm{Diag}(1/x)$ que devuelve una matriz diagonal D con los elementos x^{-1} en la diagonal. Entonces, dadas las entradas X e y, el algoritmo calcula la descomposición SVD de X y luego calcula la solución de mínimos cuadrados β mediante la expresión 8.

Algoritmo 1 Solución de mínimos cuadrados con descomposición SVD

function MINIMOSCUADRADOS(X, y)
  $U, valores\_singulares, V^T \leftarrow \mathrm{SVD}(X)$
  $D \leftarrow \mathrm{DIAG}(1/valores\_singulares)$
  $\beta \leftarrow {V^T}^T \cdot D \cdot U^T \cdot y$
  return β
end function

Para el caso en el que queremos resolver el problema de mínimos cuadrados con regularización Ridge, también podemos usar la descomposición SVD de X para hallar la solución de la ecuación 6. Partiendo de esta

expresión, podemos reescribir la solución como

$$\begin{aligned}
\beta &= (X^T X + \lambda I)^{-1} X^T y \\
&= (V\Sigma^T U^T U \Sigma V^T + \lambda I)^{-1} V\Sigma^T U^T y \\
&= (V\Sigma^T \Sigma V^T + \lambda I)^{-1} V\Sigma^T U^T y \\
&= (V\Sigma^2 V^T + \lambda V V^T)^{-1} V\Sigma^T U^T y \\
&= (V(\Sigma^2 + \lambda I)V^T)^{-1} V\Sigma^T U^T y \\
&= V(\Sigma^2 + \lambda I)^{-1} V^T V\Sigma^T U^T y \\
&= V(\Sigma^2 + \lambda I)^{-1} \Sigma^T U^T y \\
&= V\Sigma(\Sigma^2 + \lambda I)^{-1} U^T y.
\end{aligned} \tag{9}$$

Este desarrollo considera una versión cuadrada de Σ en la que se eliminan las filas nulas (lo cual no altera los resultados), una adaptación de U y V para que tengan las dimensiones correctas, que todos los valores singulares son positivos porque X tiene rango completo, que $\Sigma = \Sigma^T$ porque Σ es cuadrada y diagonal, y que $UU^T = VV^T = I$ porque V y U son ortogonales.

El algoritmo 2 muestra cómo implementar la solución 9. Como antes, asumimos que tenemos a disposición las funciones SVD y Diag. En este caso, la operacion $Diag(x/(x^2 + \lambda))$ devuelve una matriz diagonal $D \in \mathbb{R}^{d+1 \times d+1}$ tal que $D_{ii} = x_i/(x_i^2 + \lambda)$. Entonces, dadas las entradas X, y y λ, el algoritmo calcula la descomposición SVD de X y luego calcula la solución β mediante la expresión 9.

Algoritmo 2 Solución de mínimos cuadrados con regularización Ridge con descomposición SVD

function MINIMOSCUADRADOSRIDGE(X, y, λ)
  $U, valores\_singulares, V^T \leftarrow$ SVD(X)
  $D \leftarrow$ DIAG($valores\_singulares/(valores\_singulares^2 + \lambda)$)
  $\beta \leftarrow {V^T}^T \cdot D \cdot U^T \cdot y$
  return β
end function

Ambas funciones fueron implementadas en Python 3.9 [1]. Para manipular matrices usamos la biblioteca NumPy [2]. En particular, usamos el método `linalg.svd` de la biblioteca para obtener la descomposición SVD, y la función `diag` para construir matrices diagonales.

2.2 Experimento

El experimento consiste en encontrar una regresión polinomial con regularización Ridge que generalice bien a partir del conjunto de datos predeterminado.

[1] `https://www.python.org/downloads/release/python-390/`
[2] `https://numpy.org/`

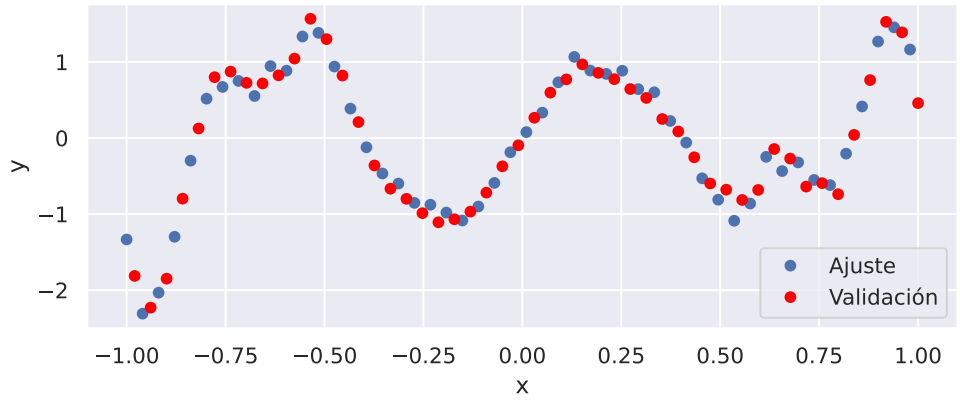


Figura 1: Gráfico de dispersión del conjunto de datos. Se diferencian los datos pertenecientes a los conjuntos de ajuste y validación.

El conjunto de datos que usamos está conformado por 100 observaciones de una variable x y una variable y, particionados aleatoriamente en dos partes iguales, un conjunto de entrenamiento y un conjunto de validación (ver Figura 1). A partir de estos datos, construimos la matriz X según la expresión 2 con polinomios de Legendre de grado d.

Para hallar un modelo que generalice, buscamos el grado de polinomio d y el valor de regularización λ que minimizan el error de validación. En particular, tomamos como posibles valores $d = 1, 2, \ldots, 49$ y 100 valores λ uniformemente distribuidos en una escala logarítimica en base 10 en el intervalo $[-7, 2]$, y evaluamos todas las combinaciones posibles (un total de 4900 combinaciones). Para cada configuración, ajustamos el modelo con el algoritmo 9 en los datos de entrenamiento y evaluamos el ECM en los datos de validación según la expresión 7.

Realizamos todos los análisis y visualizaciones con Python 3.9. Para construir los polinomios de Legendre usamos la función `polynomial.legendre.legvander` de la biblioteca NumPy.

3 Resultados

En esta sección presentamos los resultados del análisis propuesto en la sección 2.2.

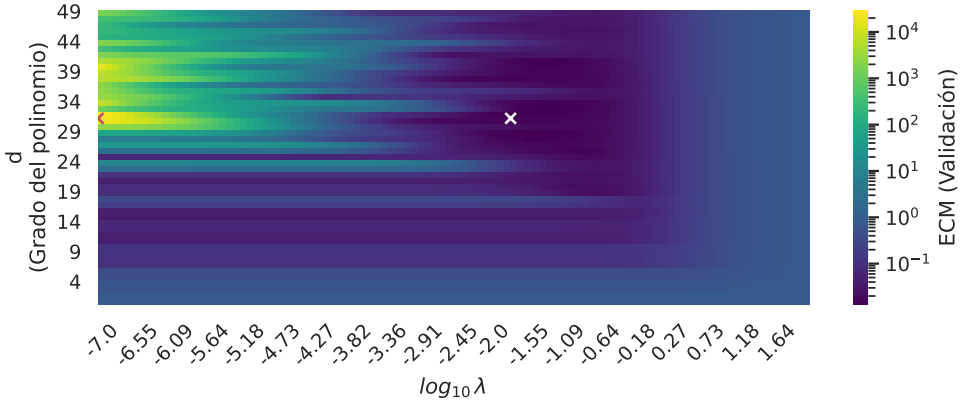


Figura 2: Mapa de calor que indica el error cuadrático medio de validación para la grilla de hiperparámetros explorada. En valores de λ que exceden el rango visualizado, el error de validación no mejora. Se señala con una cruz blanca al punto asociado al par de ECM mínimo ($\lambda \approx 10^{-1.73}, d = 32$), y con media cruz roja al asociado al de ECM máximo en todo el rango graficado ($\lambda = 10^{-7}, d = 32$).

El ECM del ajuste a los datos de validación es mínimo para el polinomio de grado $d = 32$ y coeficiente de regularización $\lambda \approx 10^{-1.73}$, con ECM $\approx 10^{-1.88}$. Casualmente, para el rango de hiperparámetros visualizado en el heatmap, $d = 32$ es también el grado polinomial asociado al error máximo, con ECM $\approx 10^{4.47}$ (figura 2). El ajuste sobre los datos de validación mediante este polinomio puede ser visualizado en la figura 3.

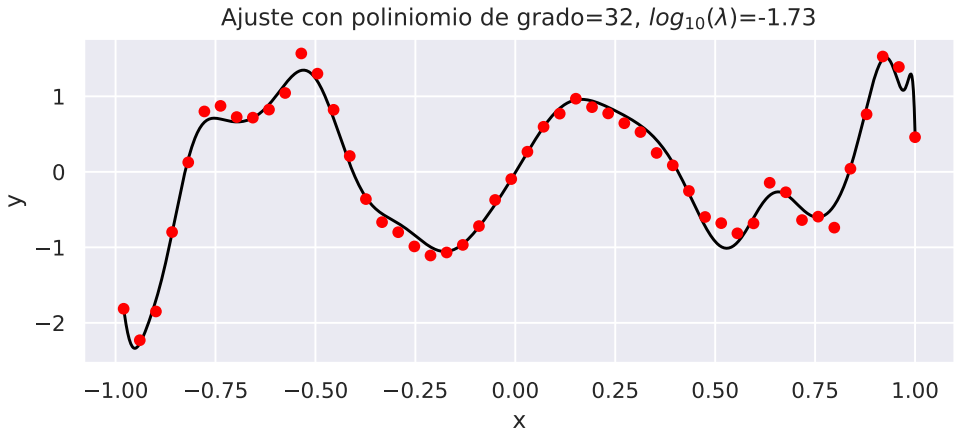


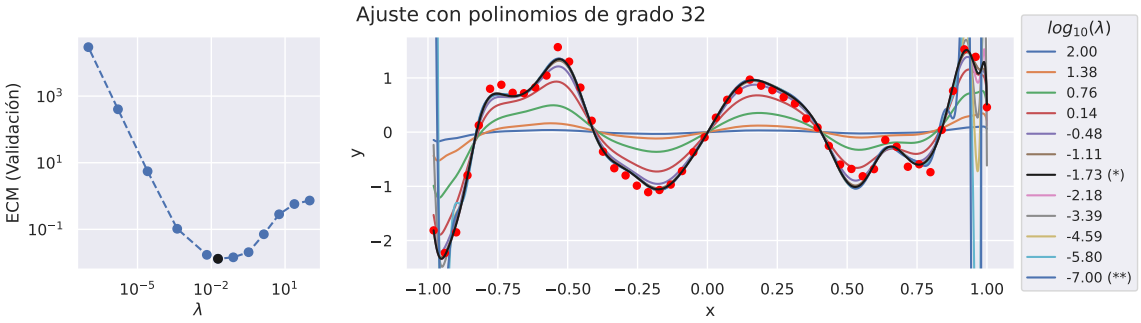
Figura 3: Valores ajustados del modelo con menor error de validación. En rojo, los datos de validación.

La figura 2 expone varias propiedades interesantes del espacio de hiperparámetros (λ, d), que fueron exploradas en más profundidad en la figura 4 y que se comentan a continuación.

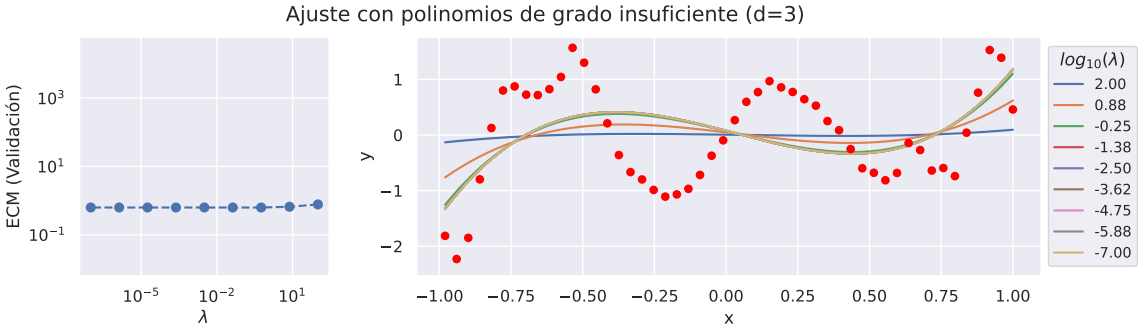
Los máximos de ECM en todo el rango visualizado en el heatmap se encuentran en la región de alto grado polinomial y bajo coeficiente de regularización, donde es razonable que lo que esté ocurriendo sea un sobreajuste a los datos de entrenamiento y un consecuente mal ajuste en términos de ECM a los datos de validación. Esto puede visualizarse en la figura 4a para el caso particular con $d = 32$, cuando se utilizan valores muy pequeños de λ. En relación a esto, una de las caracteristicas más llamativas del heatmap es la presencia de un bandeado horizontal para valores de d suficientemente grandes ($d \gtrsim 20$), con la característica de que para d fijo y al incrementar λ el ECM tiende a primero disminuir hasta llegar a una región óptima para eventualmente volver a subir. Como puede visualizarse en la figura 4a, esto parece deberse a que con un primer incremento de λ se logra disminuir el sobreajuste a los datos de entrenamiento; eventualmente se llega a un mínimo de ECM y el incremento de λ más allá de este punto produce una disminución en la calidad del ajuste por exceso de regularización.

Para grados polinomiales muy bajos (alrededor de $d = 3$) el ECM es, en terminos relativos a los errores en el resto del espacio, aproximadamente constante e independiente de λ (moviendose horizontalmente a lo largo del heatmap con d fijo). Esto evidentemente se debe a que cuando se tiene un grado polinomial demasiado bajo para los datos a ser ajustados, todo ajuste va a ser mediocre; no hay coeficiente de regularización que pueda salvar esa condición (figura 4b).

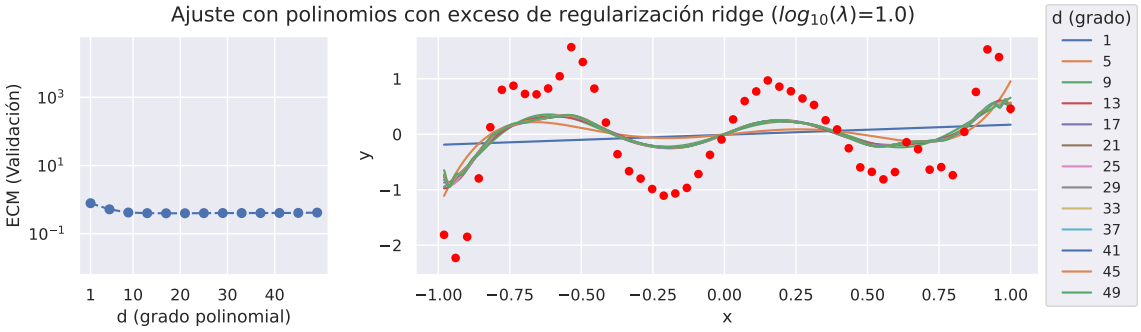
Finalmente, algo análogo ocurre para valores muy grandes de λ, más allá del valor óptimo hallado para $d = 32$, para los cuales el ECM es aproximadamente constante e independiente del grado polinomial (moviendose verticalmente a lo largo del heatmap con λ fijo). En este caso lo que ocurre es que para valores muy grandes de λ el error de cuadrados mínimos pasa a ser dominado por el error de regularización, lo que produce que todos los coeficientes de la combinación lineal de polinomios de Legendre tiendan a 0, generandose como resultado un polinomio aproximadamente de la forma $f(x) = 0$ independientemente del grado (figura 4c).



(a) Polinomios de grado $d = 32$ con coeficiente de regularización λ variable. (*) Polinomio asociado al mejor ajuste (figuras 2 y 3). En el gráfico de ECM en función de λ se encuentra resaltado el punto asociado al polinomio. (**) Polinomio asociado al peor ajuste dentro del rango graficado en la figura 2.



(b) Polinomios de grado insuficiente $d = 3$ con coeficiente de regularización λ variable.



(c) Polinomios de grado d variable con coeficiente de regularización excesivo $\lambda = 10$.

Figura 4: ECM y gráfico de ajuste a los datos de validación para polinomios asociados a series de pares (λ, d). En todos los casos se grafica a la izquierda el ECM de validación, en función de λ en (a) y (b) y del grado polinomial en (c), y a la derecha se muestran los ajustes correspondientes a cada valor del error, con los datos de validación en rojo. Para los gráficos de ECM se usó siempre la misma escala en el eje y con el fin facilitar la comparación de errores entre subfiguras.

4 Conclusiones

En este trabajo implementamos la solución al problema de mínimos cuadrados lineales mediante la descomposición SVD, con y sin regularización. También usamos la técnica de polinomios de Legendre para implementar regresión polinomial, la cual nos permite ajustar datos de manera no lineal resolviendo el problema de cuadrados mínimos lineales de manera computacionalmente eficiente. Mediante una aplicación práctica en la que buscamos el grado de polinomio y el valor de regularización que minimizan el error de validación, mostramos que la regularización es una estrategia efectiva para optimizar la capacidad predictiva de un modelo de ajuste.

Referencias

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

Gilbert Strang. *Introduction to linear algebra*. SIAM, 2022.

David S Watkins. *Fundamentals of matrix computations*. John Wiley & Sons, 2004.