Clase 3 - Regresión lineal (parte 2)

Aprendizaje de Máquinas - MA5204

Felipe Tobar

Department of Mathematical Engineering &
Center for Mathematical Modelling
Universidad de Chile

16 de marzo de 2021

UNIVERSIDAD
DE CHILE

Mínimos cuadrados regularizados

En la clase anterior se vio que el criterio de mínimos cuadrados fallaba cuando los datos presentaban muestras atípicas ya que la penalización cuadrática desplazaba y rotaba el hiperplano regresor.

Una forma estándar de corregir parcialmente este problema, es agregar una penalización en la norma del parámetro, logrando que el funcional de costo promueva ajuste a los datos pero también busque parámetros de baja magnitud.

De esta forma, el nuevo funcional de costos corresponde a:

$$J_\rho = \left\| Y - \tilde{X}\theta \right\|_2^2 + \rho \left\| \theta \right\|_p^p, \ p \geq 0,$$

donde $||\cdot||_p$ denota la norma ℓ_p, es decir, $||\theta||_p = \left(\sum_{j=1}^{M+1} |\theta_j|^p \right)^{\frac{1}{p}}$ y el parámetro $\rho \geq 0$ tiene el rol de balancear la importancia entre ajuste (primer término) y regularidad de la solución (segundo término).

¿Por qué penalizar la norma del parámetro?

Se podría pensar que el criterio de mínimos cuadrados regularizados (MCR) no puede ser mejor que MC ya que MC se preocupa específicamente de ajustar los datos de forma óptima, mientras que MCR agrega un término que *sesga* la predicción.

Este análisis no es del todo correcto ya que:

- MC solo ajusta los datos de entrenamiento y no necesariamente eso produce una buena generalización a datos no observados.
- MC presenta una alta varianza cuando el parámetro es de norma alta ya que pequeñas variaciones en la entrada pueden provocar grandes variaciones en la salida.
- Por el mismo argumento, si bien MCR agrega un sesgo en la predicción, este es compensado por la disminución de la norma del parámetro θ.
- Lo anterior será visto formalmente en la descomposición sesgo-varianza para el caso $p = 2$.

Ridge regression o regularización de Tikhonov ($p = 2$)

Para el caso $p = 2$, el funcional tiene la forma

$$J_\rho = \left\| Y - \tilde{X}\theta \right\|_2^2 + \rho \left\| \theta \right\|_2^2 .$$

Esta regularización es denominada regularización de Tikhonov.

El funcional es estrictamente monótono por lo que tiene un único mínimo, el cual puede ser encontrado utilizando la CPO:

$$\begin{aligned}
&\nabla_\theta J_\rho = 0 \\
&\iff -2(Y - \tilde{X}\theta)^\top \tilde{X} + 2\rho\theta^\top = 0 \\
&\iff -Y^\top \tilde{X} + \theta^\top \tilde{X}^\top \tilde{X} + \rho\theta^\top = 0 \\
&\iff \theta^\top = Y^\top \tilde{X}(\tilde{X}^\top \tilde{X} + \rho\mathbb{I})^{-1} \\
&\iff \theta = (\tilde{X}^\top \tilde{X} + \rho\mathbb{I})^{-1} \tilde{X}^\top Y
\end{aligned}$$

Se observa que la condición $N > M$ ya no es necesaria ya que la matriz $\tilde{X}^\top \tilde{X} + \rho\mathbb{I}$ puede tener determinante tan grande como se quiera haciendo variar ρ, solucionando el problema de invertibilidad.

Descomposición sesgo-varianza

Una forma de analizar la efectividad de una estimación de un parámetro es mediante la descomposición sesgo-varianza. Se asumirá que las muestras están compuestas de una parte determinista $\tilde{x}$ y un ruido aditivo ϵ (de media nula), es decir:

$$y_i = \theta^\top \tilde{x}_i + \epsilon_i$$

Luego, si $\hat{\theta}$ es una estimación del parámetro θ, se puede probar que para una nueva observación $(\tilde{x}_\star, y_\star)$, el *error cuadrático esperado* asociado a la predicción $\hat{y}_\star = \hat{\theta}^\top \tilde{x}_\star$ se puede descomponer de la siguiente forma:

$$\mathbb{E}\left(y_\star - \hat{y}_\star\right)^2 = \text{Sesgo}(\hat{y}_\star)^2 + \text{Varianza}(\hat{y}_\star) + \sigma^2,$$

Donde los 3 sumandos de la descomposición corresponden a:

- $\text{Sesgo}(\hat{y}_\star) := \mathbb{E}(\hat{y}_\star) - y_\star$.
- $\text{Varianza}(\hat{y}_\star) = \mathbb{E}(\mathbb{E}(\hat{y}_\star) - \hat{y}_\star)^2$.
- σ^2 es la varianza de *ruido* ϵ y no puede ser controlada por la elección de $\hat{\theta}$.

Es decir, al momento de ajustar un parámetro hay que preocuparse en su sesgo y en su varianza.

Sesgo-varianza para MC y ridge regression

Para el caso de las estimaciones por mínimos cuadrados y ridge regression se tienen los siguientes resultados:

Teorema (descomposición sesgo-varianza para MC)

$$Sesgo(\hat{f}_\star) = 0$$
$$Varianza(\hat{f}_\star) = \sigma^2 \tilde{x}_\star^\top (\tilde{X}^\top \tilde{X})^{-1} \tilde{x}_\star$$

Teorema (descomposición sesgo-varianza para ridge regression)

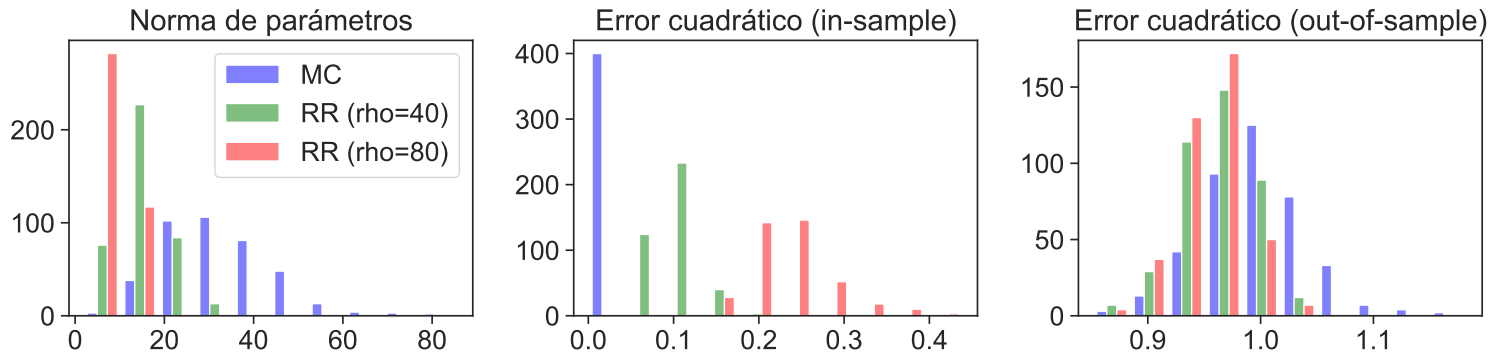
$$Sesgo(\hat{f}_\star) = \tilde{x}_\star^\top \left((\mathbb{I} + \rho(\tilde{X}^\top \tilde{X})^{-1})^{-1} - \mathbb{I} \right) \theta$$
$$Varianza(\hat{f}_\star) = \sigma^2 \tilde{x}_\star^\top (\tilde{X}^\top \tilde{X} + \rho\mathbb{I})^{-1} \tilde{X}^\top \tilde{X} (\tilde{X}^\top \tilde{X} + \rho\mathbb{I})^{-1} \tilde{x}_\star$$

Además, se puede probar que la varianza del estimador de ridge regression es menor a la varianza del estimador de MC.

De esta forma, el parámetro ρ juega el papel de balancear el sesgo y la varianza del estimador.

Ejemplo de ridge regression vs. mínimos cuadrados

- Sean $N = 1000$ muestras relacionados linealmente.
- Se utilizarán solo $N' = 15$ muestras para entrenar el modelo usando MC, y RR con $\rho \in \{40, 80\}$.
- Se repetirá el proceso 400 veces para analizar como se comportan los estimadores dentro de muestra (con respecto a los datos de entrenamiento) y fuera de muestra (con respecto a los $N - N'$ datos no usados para entrenar).



- A mayor ρ, los parámetros encontrados tienen menor magnitud.
- MC se comporta mejor en evaluación dentro de muestra.
- RR se comporta mejor en evaluación fuera de muestra.

Por lo tanto, se observa que incluir un sesgo en el ajuste del modelo (RR) puede ayudar a generalizar y no sobreajustar cuando se tienen pocos datos.

Equivalencia con un problema de optimización con restricción

Si bien el problema de MCR incluye una penalización sobre la norma de θ, el problema también puede ser visto como el problema de MC con la restricción adicional que fija la norma del parámetro. En efecto, dicho problema viene dado por:

$$\min_{\theta} \left\| Y - \tilde{X}\theta \right\|_2^2 \quad \text{s.a.} \quad ||\theta||_p^p = \tau$$

donde $\tau \geq 0$ es una constante fija. El lagrangiano del problema es el siguiente:

$$L(\theta, \lambda) = \left\| Y - \tilde{X}\theta \right\|_2^2 + \lambda \left(||\theta||_p^p - \tau \right)$$

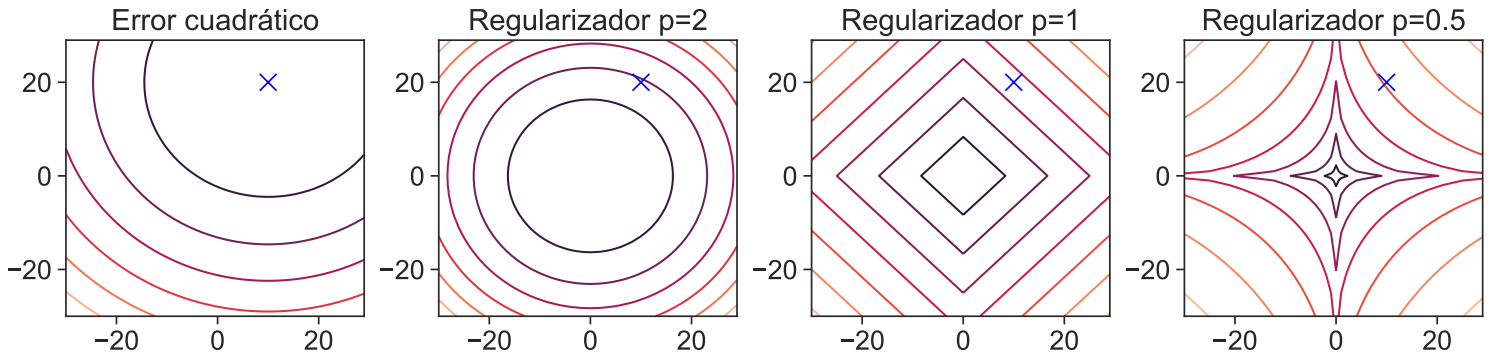
donde λ es el multiplicador de Lagrange asociado al problema. Usando la condición de primer orden se tiene que:

$$\begin{aligned} \frac{\partial L}{\partial \theta} = 0 \quad &\Rightarrow \quad \frac{\partial}{\partial \theta} \left(\left\| Y - \tilde{X}\theta \right\|_2^2 + \lambda ||\theta||_p^p \right) = 0 \\ \frac{\partial L}{\partial \lambda} = 0 \quad &\Rightarrow \quad ||\theta||_p^p = \tau, \end{aligned}$$

lo cual recupera la forma del problema de MCR. Es decir, se puede interpretar MCR como MC con una restricción sobre la norma.

Curvas de nivel de los regularizadores

Con la interpretación anterior, podemos entender distintos regularizadores (distintos $p \geq 0$) mediante sus curvas de nivel. La siguiente imagen ilustra las curvas de nivel correspondientes al costo cuadrático (izquierda) y al término de regularización para $p \in \{0.5, 1, 2\}$.



Se observa cómo las curvas de nivel atraen el mínimo hacia el origen de distinta forma:

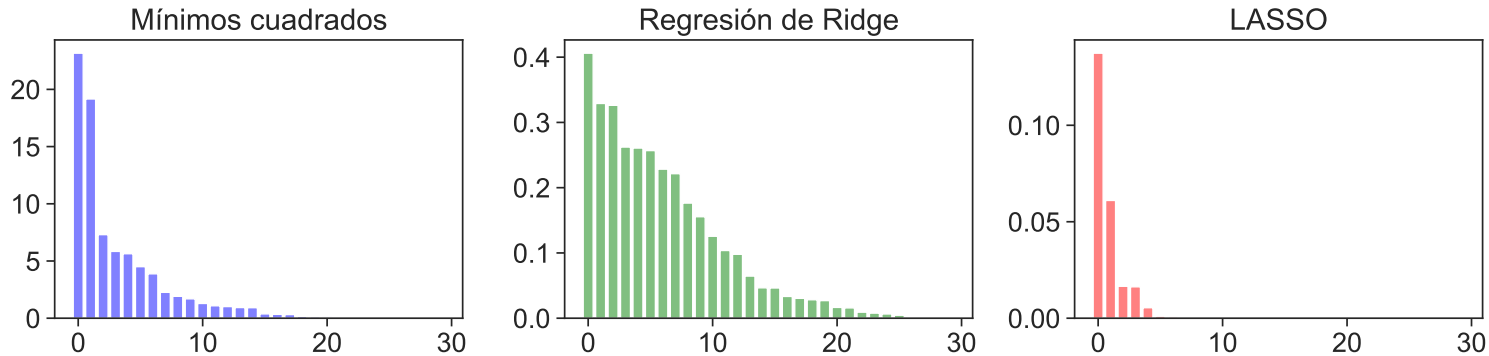
- $p = 2$ lleva la solución al origen de forma homogénea.
- $p \in \{0.5, 1\}$ lleva la solución a los bordes, es decir, privilegiando soluciones ralas.

Para $p \leq 1$, la forma de la curva de nivel permite que, usualmente, la solución del problema se concentre el las puntas del *diamante*, llevando algunas coordenadas del parámetro θ directamente a cero. Por esto decimos que $p \leq 1$ tiene la propiedad de *selección de variables*.

LASSO y selección de características

El caso $p = 1$ es conocido como LASSO y como se vio, posee la propiedad de selección de características.

Para ilustrar esta propiedad, se utilizará un dataset fijo, el cual tiene $N = 569$ muestras y un dimensión de entrada de $M = 30$ y se usarán $2/3$ de los datos para entrenar y el $1/3$ restante para calcular puntajes fuera de muestra.



Se observa la disminución en la norma de θ al usar MCR, así como también la selección de características realizada por LASSO.

	in-sample	out-of-sample
MC	**0.7896**	0.6911
RR	0.6905	0.6903
LASSO	0.7452	**0.7242**

Notar que LASSO tiene el mejor puntaje fuera de muestra.

Elastic net y consideraciones generales

Para concluir esta sección, es importante tener en cuenta que:

- Una posible variación de MCR consiste en utilizar la regularización LASSO y ridge al mismo tiempo, resultando en lo que se denomina *elastic net regularization*, cuyo funcional de costo es

$$J_\rho = \left\| Y - \tilde{X}\theta \right\|_2^2 + \lambda_1 \left\| \theta \right\|_2^2 + \lambda_2 \left\| \theta \right\|_1$$

- Como se vio en la descomposición sesgo-varianza, el *hiperparámetro* ρ determina el balance entre regularización (cuán sesgado) y ajuste (cuán bien replica los datos de entrenamiento. en la práctica, para elegir este hiperparámetro se puede evaluar el desempeño fuera de muestra para distintos valores de ρ con la finalidad de elegir un valor apropiado. Una manera de realizar la evaluación de desempeño para diferentes ρ será vista al estudiar la *validación cruzada*.