Tarea 1

Profesor: Felipe Tobar
Auxiliares: José Díaz, Diego Garrido, Jou-Hui Ho, Luis Muñoz, Diego Troncoso
Consultas: Diego Garrido, Diego R. Troncoso (U-cursos)
Período: 13/4/2020 — 20/4/2020

Formato entrega: Informe en formato PDF, con una extensión máxima de 3 páginas (puede usar un formato de doble columna), presentando y analizando sus resultados, y detallando la metodología utilizada. Adicionalmente debe entregar el jupyter notebook (o el código que haya generado) con la resolución de la tarea.

P1. Descomposición Sesgo-Varianza (2.0 puntos)

(a) (0.5 puntos) Considere el conjunto de observaciones $D = \{(x_i, y_i)\}_{i=1}^N$, relacionadas mediante el modelo lineal

$$y_i = \theta^\top x_i + \epsilon_i,$$

donde $\{\epsilon_i\}_{i=1}^N$ son observaciones i.i.d con $\mathbb{E}[\epsilon] = 0$ y $\text{var}(\epsilon) = \sigma^2$, θ es un parámetro fijo y desconocido. Para un nuevo par, denotado (x, y), no contenido en el conjunto de observaciones D, considere la predicción de y mediante $\hat{y} = \hat{\theta}^\top x$, donde $\hat{\theta}$ es un estimador del parámetro θ basado en D. Muestre que el costo cuadrático esperado de predecir y con $\hat{y}$ (recuerde que las esperanzas se toman con respecto a la ley de ϵ) admite la siguiente descomposición sesgo-varianza:

$$\mathbb{E}\left[(\hat{y} - y)^2\right] = \text{var}(\hat{y}) + \text{sesgo}^2(\hat{y}) + \sigma^2,$$

donde: i) el primer término es la varianza de la variable aleatoria $\hat{y}$, con $\text{var}(\hat{y}) = \mathbb{E}\left[(\hat{y} - \mathbb{E}[\hat{y}])^2\right]$ (recuerde que $\hat{\theta}$ es variable aleatoria, pues depende de D), ii) el segundo término es el sesgo al cuadrado de la variable aleatoria $\hat{y}$, con $\text{sesgo}(\hat{y}) = E\left[\hat{y} - y\right] = E\left[\hat{y}\right] - \theta^\top x$, y iii) σ^2 es la varianza de ϵ.

Solución:

$$\begin{aligned}
E[(\hat{y} - y)^2] &= E\left[((\hat{y} - E[\hat{y}]) - (E[y] + \epsilon - E[\hat{y}]))^2\right] \\
&= E[(\hat{y} - E[\hat{y}])^2] + E[(E[y] + \epsilon - E[\hat{y}])^2] - 2E[(\hat{y} - E[\hat{y}])(E[y] + \epsilon - E[\hat{y}])] \\
&= var(\hat{y}) + E[E[y]^2 + \epsilon^2 + E[\hat{y}]^2 + 2\epsilon E[y] - 2E[y]E[\hat{y}] - 2\epsilon E[\hat{y}]] \\
&- 2E[\hat{y}E[y] + \hat{y}\epsilon - \hat{y}E[\hat{y}] - E[\hat{y}]E[y] - E[\hat{y}]\epsilon + E[\hat{y}]^2] \\
&= var(\hat{y}) + E[y]^2 + \sigma^2 + E[\hat{y}]^2 - 2E[y]E[\hat{y}] - 2E[\hat{y}]E[y] + 2E[\hat{y}]^2 + 2E[\hat{y}]E[y] - 2E[\hat{y}]^2 \\
&= var(\hat{y}) + \sigma^2 + E[y]^2 + E[\hat{y}^2] - 2E[y]E[\hat{y}] \\
&= var(\hat{y}) + sesgo^2(\hat{y}) + \sigma^2
\end{aligned}$$

(b) (1.0 puntos) Para los parámetros de mínimos cuadrados $\theta_{MC} = (X^\top X)^{-1} X^\top Y$ y de mínimos cuadrados regularizados $\theta_{MCR} = (X^\top X + \rho I)^{-1} X^\top Y$ calcule el sesgo y la varianza de la predicción.

Solución:
MC

(i) **Sesgo**

$$\begin{aligned} E[\hat{y} - y] &= E[x^T \hat{\theta}_{MC}] - E[x^T\theta + \epsilon] = E[x^T (X^T X)^{-1} X^T Y] - x^T\theta \\ &= E[x^T (X^T X)^{-1} X^T (X\theta + \epsilon)] - x^T\theta = E[x^T\theta + x^T (X^t X)^{-1} X^T \epsilon] - x^T\theta \\ &= x^T\theta - x^T\theta = 0 \end{aligned}$$

(ii) **Varianza**

$$\begin{aligned} E[(\hat{y} - E[\hat{y}])^2] &= E[\hat{y}^2] - E[\hat{y}]^2 = E[(x^T\theta + x^T (X^T X)^{-1} X^T \epsilon)^2] - (x^T\theta)^2 \\ &= E[(x^T\theta)^2 + 2x^T\theta w^T \epsilon + w^T \epsilon\epsilon^T w] - (x^T\theta)^2, w^T = x^T (X^T X)^{-1} X^T, w^T\epsilon = \epsilon^T w \\ &= (x^T\theta)^2 + w^T\sigma^2 I w - (x^T\theta)^2 = \sigma^2 x^T (X^T X)^{-1} X^T X (X^T X)^{-1} x = \sigma^2 x^T (X^T X)^{-1} x \end{aligned}$$

MCR

(i) **Sesgo**

$$\begin{aligned} E[\hat{y} - y] &= E[x^T \hat{\theta}_{MCR}] - x^T\theta = E[x^T (X^T X + \rho I)^{-1} X^T (X\theta + \epsilon)] - x^T\theta \\ &= x^T (X^T X + \rho I)^{-1} X^T X\theta - x^T\theta = x^T (I + \rho (X^T X)^{-1})^{-1}\theta - x^T\theta \\ &= x^T \left[(I + \rho (X^T X)^{-1})^{-1} - I \right]\theta \end{aligned}$$

(ii) **Varianza**

$$\begin{aligned} E\left[(\hat{y} - E[\hat{y}])^2\right] &= E[\hat{y}^2] - E[\hat{y}]^2 = E\left[\left(w^T (X\theta + \epsilon)\right)^2\right] - E\left[w^T (X\theta + \epsilon)\right]^2, w^T = x^T (X^T X + \rho I)^{-1} X^T \\ &= E\left[(w^T X\theta)^2 + 2w^T X\theta\epsilon + w^T \epsilon\epsilon^T w\right] - (w^T X\theta)^2 = w^T\sigma^2 I w \\ &= \sigma^2 x^T (X^T X + \rho I)^{-1} X^T X (X^T X + \rho I)^{-1} x \end{aligned}$$

(c) (0.5 puntos) Analice las expresiones anteriores. ¿Cuáles son las ventajas y desventajas de ambos estimadores? ¿Qué aseveraciones puede hacer de la comparación de ambos criterios (MC y MCR) con respecto de sus sesgos y varianzas?

Solución:

La desventaja de MCR es que la predicción es sesgada, puesto que para x y θ cualquiera la única forma de que el sesgo sea 0 es tomando $\rho = 0$, lo que resultaría en el estimador de MC. Como ventaja de MCR se tiene que la varianza de la predicción es menor que la de MC, lo que lo vuelve más robusto, siendo menos sensible a correlaciones espúreas u outliers presentes en el conjunto de entrenamiento teniendo la capacidad de generalizar mejor en conjunto de datos fuera de muestra.

Mostrar que la varianza de la predicción de MCR es menor que MC no es trivial, por tanto, debe demostrarse. Por demostrar que $var(\hat{y}_{MC}) - var(\hat{y}_{MCR}) > 0$:

$$
\begin{aligned}
var(\hat{y}_{MC}) - var(\hat{y}_{MCR}) &= \sigma^2 x^T (X^T X)^{-1} x - \sigma^2 x^T W (X^T X)^{-1} W^T x,\ W = (X^T X + \rho I)^{-1} X^T X \\
&= \sigma^2 x^T W [W^{-1} (X^T X)^{-1} W^{T^{-1}} - (X^T X)^{-1}] W^T x \\
&= \sigma^2 x^T W [(X^T X)^{-1} (X^T X + \rho I)(X^T X)^{-1} (X^T X + \rho I)(X^T X)^{-1} - (X^T X)^{-1}] W^T x \\
&= \sigma^2 x^T W [2\rho (X^T X)^{-2} + \rho^2 (X^T X)^{-3}] W^T x \\
&= \sigma^2 x^T [(X^T X + \rho I)^{-1}][2\rho I + \rho^2 (X^T X)^{-1}][(X^T X + \rho I)^{-1}]^T x \\
&= \theta^2 x^T \Sigma(\rho) x > 0,\ \Sigma(\rho) = [(X^T X + \rho I)^{-1}][2\rho I + \rho^2 (X^T X)^{-1}][(X^T X + \rho I)^{-1}]^T,\ \forall \rho > 0
\end{aligned}
$$

Para llegar la desigualdad de la útlima linea un pequeño repaso de algebra lineal:

Una matriz semidefinida positiva $W \succeq 0$ cumple que $x^T W x \geq 0\ \forall x$, si $x^T W x > 0\ \forall x \neq 0$ se dice que W es definida positiva $W \succ 0$. Algunas propiedades:

1. $A \succeq 0, B \succeq 0 \implies A + B \succeq 0$
2. $A \succ 0, B \succeq 0 \implies A + B \succ 0$
3. $A \succeq 0, \lambda \geq 0 \implies \lambda A \succeq 0$
4. $A \succ 0, \lambda > 0 \implies \lambda A \succ 0$
5. $A \succeq 0 \implies A^{-1} \succeq 0$
6. $A \succ 0 \implies A^{-1} \succ 0$
7. Si $M \succeq 0$ y Q tiene rango completo en las columnas $\implies Q^T M Q \succeq 0$
8. Si $M \succ 0$ y Q tiene rango completo en las columnas $\implies Q^T M Q \succ 0$
9. $A \succeq 0, B \succeq 0$ y $AB = BA$, entonces $AB \succeq 0$
10. $A \succ 0, B \succ 0$ y $AB = BA$, entonces $AB \succ 0$

La matriz $X^T X$ es semidefinida positiva, la matriz I es definida positiva, $(X^T X + \rho I)^{-1}$ es definida positiva $\forall \rho > 0$ (propiedad 2, 4 y 6), además tiene rango completo en las columnas puesto que es invertible, similarmente $2\rho I + \rho^2 (X^T X)^{-1}$ es definida positiva $\forall \rho > 0$, luego por propiedad (8) $\Sigma(\rho) = [(X^T X + \rho I)^{-1}][2\rho I + \rho^2 (X^T X)^{-1}][(X^T X + \rho I)^{-1}]^T$ es definida positiva $\forall \rho > 0$.

Otra forma de demostrar lo mismo es darse cuenta que $var(\hat{y})$ es decreciente en ρ, para esto debemos demostrar que su derivada es no positiva:

$$
\begin{aligned}
var(\hat{y})' &= -\sigma^2 x^T (X^T X + \rho I)^{-2} X^T X (X^T X + \rho I)^{-1} x - \sigma^2 x^T (X^T X + \rho I)^{-1} X^T X (X^T X + \rho I)^{-2} x \\
&= -\sigma^2 x^T [(X^T X + \rho I)^{-1} ((X^T X + \rho I)^{-1} X^T X + X^T X (X^T X + \rho I)^{-1})(X^T X + \rho I)^{-1}] x^T, \\
&= -\sigma^2 x^T (X^T X + \rho I)^{-1} [2(I + \rho (X^T X)^{-1})^{-1}](X^T X + \rho I)^{-1} x^T \\
&= -\sigma^2 x^T \Sigma(\rho) x^T < 0,\ \Sigma(\rho) = (X^T X + \rho I)^{-1} [2(I + \rho (X^T X)^{-1})^{-1}](X^T X + \rho I)^{-1} \succ 0\ \forall \rho > 0
\end{aligned}
$$

Tenemos que $2(I + \rho (X^T X)^{-1})$ es definida positiva por propiedad 2 y 6, ya que $I \succ 0$ y $\rho (X^T X)^{-1} \succeq 0$.

P2. Regresión Lineal (4.0 puntos) Solución

Como primer paso debe instalar Anaconda una distribución de Python que proporciona el stack básico de Python para la ciencia de datos, debe descargar la versión Python 3.7. Anaconda incluye Jupyter Notebook, un entorno de desarrollo interactivo web.

Para esta pregunta se pide implementar el estimador de mínimos cuadrados y de mínimos cuadrados regularizados con regularizador *ridge* para un conjunto de datos. Para la implementación de los estimadores **solo está permitido el uso de operaciones de álgebra lineal**, para esto pueden utilizar el stack de numpy.

Como base de datos del experimento se utilizará el archivo **Housing.csv**, este corresponde a un dataset que posee precios de casas de algunas localidades de USA. Este dataset consta de una variable X, que corresponde al ingreso promedio de la población en esa área (*Avg Area Income*), e Y, que corresponde al precio de las casas. El objetivo de esta pregunta es aprender una función lineal que relacione X e Y, así un agente inmobiliario que no tiene información sobre el precio de las casas en una nueva localidad pueda fijarle un precio a estas en base al ingreso promedio del área. Para una correcta comparación de los estimadores la base de datos viene dividida en dos conjuntos, entrenamiento (*in-sample*) y validación (*out-of-sample*).

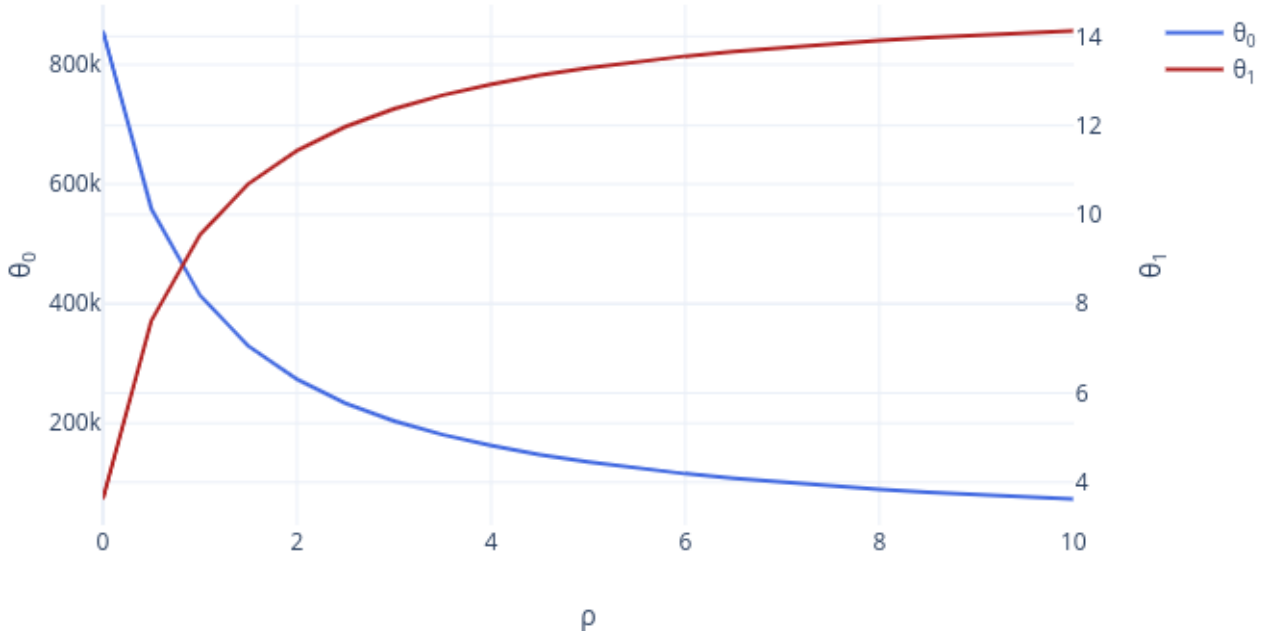
Para esto deberá:

(a) (0.5 puntos) Cargar los datos desde el archivo **Housing.csv** y graficarlos.

(b) (1.0 puntos) Implemente el estimador de mínimos cuadrados regularizados usando regularización *ridge* y obtenga el vector de parámetros θ para diferentes valores de $\rho \in [0, 10]$ incluyendo los límites. Para esto implemente la función *reg_lineal(X,Y,ρ)*. Ejemplo de parámetros estimados para $\rho \in [0, 1]$, donde θ_0 es el intercepto y θ_1 la pendiente.

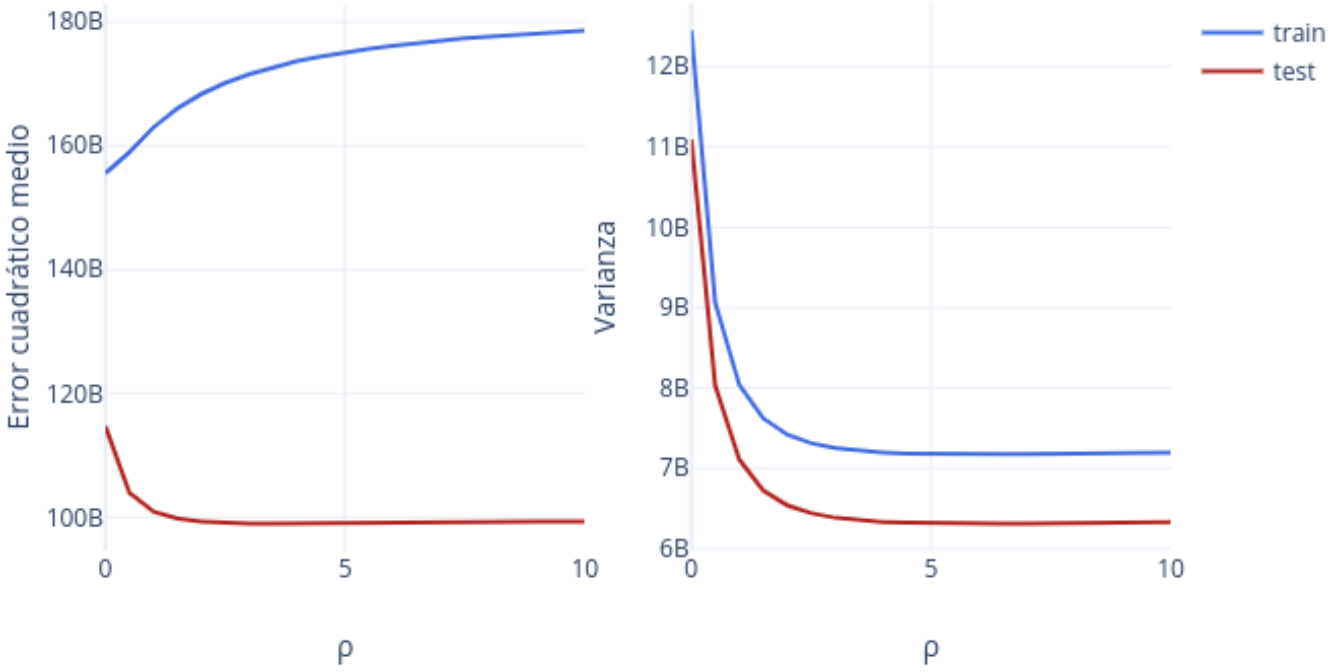
ρ	θ_0	θ_1
0.0	856479.11	3.62
0.5	558528.88	7.62
1.0	414376.44	9.55
1.5	329368.67	10.69
2.0	273301.82	11.44
2.5	233546.35	11.98
3.0	203888.06	12.37
3.5	180913.64	12.68
4.0	162592.48	12.93
4.5	147640.86	13.13
5.0	135207.49	13.29
5.5	124705.59	13.44
6.0	115717.52	13.56
6.5	107937.97	13.66
7.0	101138.54	13.75
7.5	95145.00	13.83
8.0	89822.07	13.90
8.5	85063.18	13.97
9.0	80783.18	14.02
9.5	76913.25	14.08
10.0	73397.14	14.12

(c) (0.5 puntos) Grafique el valor de los parámetros estimados para los valores de ρ.



(d) (0.5 puntos) Grafique el error cuadrático medio y la varianza de la predicción para los valores de ρ tanto en el conjunto de entrenamiento como en el de validación.

La varianza de la predicción no es $var(\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2$, si se calcula así se tendrá un curva creciente en ρ, se debe calcular a partir del promedio de las varianza de la predicción sobre observaciones individuales usando la formula derivada en la pregunta 1. Notar que el σ^2 de la formula se estima sobre los datos de entrenamiento, por ende σ^2 se estima una vez para cada ρ.



(e) (0.5 puntos) Grafique la ecuación de la recta con los parámetros estimados para diferentes valores de ρ junto a los datos de entrenamiento y validación.

(f) (1.0) Discuta cómo elegir el valor apropiado de ρ en base a los resultados obtenidos en los puntos anteriores.

Notar que de la figura del punto (b) se observa que la norma del vector de parámetros es decreciente en ρ, el efecto de esto se observa en (e), donde las rectas obtenidas fijando ρ mayor son menos afectadas o cargadas en la dirección de los dos outliers de la esquina inferior derecha presentes en el conjunto de entrenamiento.

Es importante destacar que el desempeño en el conjunto de entrenamiento empeora a medida que ρ aumenta y de hecho es mínimo para $\rho = 0$, pero le objetivo de una tarea predictiva es obtener un modelo con alta capacidad de generalización, es decir, que tenga un buen desempeño en un conjunto de datos no visto en entrenamiento, en este caso ρ puede calibrarse de tal manera que la disminución de la varianza de la predicción compense el incremento del sesgo al cuadrado, pudiendo tener un error cuadrático medio (nuestra médida de desempeño escogida) igual o menor al obtenido por MC en un conjunto de validación, para los ρ probados el error cuadrático medio alcanza su valor mínimo en $\rho = 3.5$.

Notar que buscar ρ que minimice la varianza de la predicción no está bien, ya que si tomamos $\rho = \infty \implies \theta = 0 \implies var(\hat{y}) = 0$, luego se tiene que la varianza es mínima y el error cuadrático medio es máximo.

No se permite el uso de paquetes predefinidos para regresión lineal. Estos pueden ser considerados para contrastar los propios resultados pero no para resolver la pregunta. E.g., `numpy.polyfit`, `scipy.stats.linregress`, `sklearn.linear_model.LinearRegression`.