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.

(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.

(c) (1.0 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?

P2. Regresión Lineal (4.0 puntos)

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,rho)`.

(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.

(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.

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`.