Tarea 1

Fecha de entrega: 1 de abril 2024 (en clase)

Profesor: Felipe Tobar
Auxiliares: Augusto Tagle, Camilo Carvajal Reyes, Fernando Fetis, Juan Carlos Cuevas

Formato de entrega: PDF con extensión máxima de 3 páginas, presentando y analizando sus resultados, y detallando la metodología utilizada. Recomendación: Incorpore gráficos para apoyar su análisis. Adicionalmente debe entregar el jupyter notebook (o el código que haya generado) con la resolución de la tarea y debe fijar semillas antes de los procesos aleatorios en su código para poder replicar resultados.

En el siguiente enlace encontrará un conjunto de datos de que contiene reseñas de productos de *Amazon*. Para el desarrollo de las siguientes preguntas ocupe sólo un 20 % de estos datos, eligiéndolos aleatoriamente usando su rut (sin verificador) como semilla aleatoria.

(a) (2 puntos) Implemente la clase `Linear Regression` utilizando como base los métodos que se describen en el código entregado al final de la tarea. Esta clase deberá ser utilizada para el resto de la tarea. En esta sección no puede usar librerías para implementar su clase (excepto *numpy*).

Consideremos un modelo de regresión lineal simple (mínimos cuadrados) para predecir la calificación por estrellas de cada reseña utilizando solo tres características del conjunto de datos. Consideramos entonces que la etiqueta y_{sr}, que corresponde a la columna `star_rating` (cantidad de estrellas) estará dada por:

$$y_{sr} = \theta_0 + \theta_1 x_{vp} + \theta_2 x_{tv} + \theta_3 x_{lr} + \epsilon\,,$$

donde x_{vp} corresponderá a la columna `verified_purchase`, x_{tv} a la columna `total_votes` y x_{lr} a la columna `length_of_review` (creada por usted). Por otro lado, θ_0, θ_1, θ_2 y θ_3 son los parámetros del modelo y $e \sim \mathcal{N}(0, \sigma^2)$ corresponde al ruido.

(b) (0.5 puntos) ¿Cuál es la distribución de las calificaciones en el conjunto de datos? Grafique.

(c) (1.5 punto) Divida **su** conjunto de datos en dos, considerando un 90 % para entrenamiento y el 10 % restante para prueba. Entrene el modelo por mínimos cuadrados en el dataset de entrenamiento y reporte los valores de θ_0, θ_1, θ_2 y θ_3. Dé una breve explicación de estos parámetros (i.e., que es lo que representan). Adicionalmente, reporte el error cuadrático medio (ECM) del modelo en el conjunto de entrenamiento y en el de prueba, y explique los resultados obtenidos.

(d) (1 punto) Implemente LASSO y Ridge Regression. Utilice distintos valores para ρ y grafique el ECM en función de éste para ambos datasets (entrenamiento y prueba). Grafique la norma de los parámetros encontrados en para los distintos modelos e indique qué puede observar de las magnitudes obtenidas para cada método.

(e) (0.5 punto) Plantee un modelo utilizando características polinomiales, dejando explícita su formulación matemática. Implemente dicho modelo y reporte su ECM en los conjuntos de entrenamiento y prueba.

(f) (0.5 puntos) De todos los modelos/métodos de regresión utilizados indique cuál recomendaría usted. Fundamente su respuesta.

Estrutura para la clase Linear Regression

```
class LinearRegression:

    def __init__(self):
        pass

    def fit(self):
        pass

    def predict(self):
        pass
```