Tarea 1

Fecha de entrega: 28 de abril 2023

Profesor: Felipe Tobar
Auxiliares: Catherine Benavides, Camila Bergasa, Víctor Caro,
Camilo Carvajal Reyes, Diego Cortez M., Stefano Schiappacasse

Formato de entrega: PDF con extensión máxima de 5 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.

P1 - Regresión Lineal Bayesiana (40 %)

Sea X una variable aleatoria en $\mathbb{R}^p$ e Y variable aleatoria a valores $\mathbb{R}$. Consideremos el modelo de regresión lineal:

$$Y = \beta_0 + \beta^T X + \epsilon\,, \text{ con } \epsilon \sim \mathcal{N}(0, \sigma^2),$$

con $\beta = (\beta_1, \ldots, \beta_p)^T$. Suponemos σ^2 conocido. Nos interesamos en la estimación Bayesiana de los parámetros $\theta = (\beta_0, \beta_1, \ldots, \beta_p)$. Suponemos que la distribución X no depende de θ. Sean $(x_i, y_i)_{1 \leq i \leq N}$ datos.

(a) (1 punto) Demuestre que el estimador *máximo a posteriori* con un prior constante (que no depende de θ) equivale al estimador de máxima verosimilitud.

Supongamos que usamos el prior siguiente:

$$p(\beta_i) = \frac{1}{\sqrt{2\pi}\tau} = e^{-\frac{\beta_i^2}{2\tau^2}}$$

(b) (2 puntos) Compute el estimador *máximo a posteriori*. Identifique a qué estimador conocido corresponde.

(c) (2 puntos) Considere los casos

- $\tau \to \infty$
- $\tau \to 0$

Argumente qué sucede en cada caso, tanto en términos del estimador encontrado, como en términos "Bayesianos" (esto es, piense en la forma que tendrían los priors en ambos casos).

Ahora denote $\rho = \frac{\sigma^2}{\tau^2}$, que lo interpretaremos como un hiper-parámetro de nuestro modelo.

(d) (1 punto) Demuestre que el estimador resultante de la minimización es insesgado si y sólo si $\rho = 0$.

P2 - Regresión lineal y clasificación (%60)

En el siguiente enlace encontrará un conjunto de datos de valores, que incluye 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. Puede utilizar la biblioteca `scikit-learn` para sus modelos, excepto cuando se indique lo contrario.

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 dado 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.

(a) (0 puntos)¿Cuál es la distribución de las calificaciones en el conjunto de datos? Grafique.

(b) (1 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.

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

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

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

(f) (2.5 puntos) Implemente una clase *Logistic Regression* a mano utilizando como base los métodos que se describen en el código que figura al final de la tarea. Entrene su clase con su dataset de entrenamiento. Reporte el ECM en ambos conjuntos de datos. Comente sus resultados **Observación:** En esta sección no puede usar librerías para implementar su clase (excepto *numpy*).

(g) (0.5 puntos) ¿Cuál, según usted, es mejor manera de plantear este problema, como un problema de regresión o de clasificación? Concluya sobre los resultado obtenidos mediante regresión lineal y regresión logística. Defina explícitamente cómo comparará ambos resultados siendo que resuelven problemas distintos.

```python
def softmax_():
    pass

class MultipleLogRegression:

    def ___init___(self):
        pass

    def fit(self):
        pass

    def predict(self):
        pass
```