FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS
TIC 60153 – Análisis Cuantitativo
Examen 2 (50 %): Modelos de Regularización

Nombre: _____ Código: _____ Nota: _____

Profesor: **Santiago Ortiz - Henry Velasco** Grupo: **01** Fecha: _____ de 20__

Notas:

- Todas las respuestas, gráficas, tablas y operaciones deben ser debidamente justificadas.
- La información que sea obtenida de alguna fuente debe ser citada y referenciada en el documento a entregar.

1) Considere el conjunto de datos **"`Boston Housing Data`"** presentados en Harrison and Rubinfeld (1978). Defina como variable respuesta a la columna *MEDV*. Realice una partición 80-20, donde el primer 80 % de los datos son datos de entrenamiento y el restante 20 % son datos para prueba. Realizar:

   - Genere los modelos de regresión por regularización **Ridge**, **LASSO** y **Elastic-Net** para los datos de entrenamiento. Encuentre los valores óptimos de α^* y λ^* junto a su respectiva gráfica de evolución de los coeficientes de regresión. Compare los modelos en términos de la selección de variables, interprete los coeficientes y escriba la ecuación ajustada de regresión para cada caso. Finalmente, realice una predicción con las observaciones de prueba y determine cual de los tres modelos es el mejor en capacidad predictiva **(RMSE)**.
   - Sea $Z \in \mathbb{R}^{n\times p}$ una muestra de datos multivariantes. La distancia de Mahalanobis (MD), presentado por Mahalanobis (1936) y conocida como la distancia estadística, se define de la siguiente forma:

$$\text{MD}(z_i) = \sqrt{(z_i - \hat{\mu})'\hat{\Sigma}^{-1}(z_i - \hat{\mu})}, \quad \text{para } i = 1, \ldots, n.$$

     Donde $x_i \in \mathbb{R}^{p\times 1}$ denota una observación (fila), pero transpuesta, de la muestra, $\hat{\mu} \in \mathbb{R}^{p\times 1}$ el vector de medias (vector con los promedios de cada columna o variable) y $\hat{\Sigma}^{-1} \in \mathbb{R}^{p\times p}$ la inversa de la matriz de covarianza de los datos. MD puede interpretarse como la distancia que tiene cada individuo de la muestra al individuo ideal o equilibrio (representado por $\hat{\mu}$), considerando la estructura de dependencia de las variables.

     Esta distancia tiene múltiples usos, como por ejemplo la detección de observaciones atípicas. Se dice que una observación z_i es atípica sí $\text{MD}(z_i) > \chi^2_{(q,p)}$, donde q denota un percentil, usualmente $q \in [0.95, 0.99]$, de la distribución Chi-squared de parámetro p.

     De acuerdo a la información anterior: Detecte las observaciones atípicas multivariantes de la muestra, usando $q = 0.90, 0.95, 0.975, 0.99$ y en un gráfico muestre las distancias calculadas para cada individuo y los puntos de corte $(\chi^2_{(q,p)})$ ¿Qué puede concluir acerca de los individuos atípicos para cada punto de corte? ¿Quienes son? ¿El gráfico de distancias muestra alguna observación NO atípica que para usted si lo fue? ¿Qué puede concluir acerca del comportamiento de q en el valor del punto de corte? Justifique detalladamente.
   - Realice dos modelos de regresión **Robusto–Regularizado**, combinando la técnica de detección de atípicos multivariantes más la regularización **Elastic-Net**. Encuentre los valores

óptimos de α^* y λ^*, interprete los coeficientes de regresión, escriba la respectiva ecuación de regresión resultante y valide los supuestos del modelo $\left(\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)\right)$. Finalmente, realice una predicción con las observaciones de prueba y determine el **RMSE** de estos dos modelos. Compare los resultados con los modelos estimados en el ítem 1.

NOTA: La aplicación de la técnica de detección de atípicos debe realizarla de la siguiente manera:

- **Modelo 1:** Estandarizar las variables y calcular MD usando el vector de medianas en vez del vector de medias y la matriz de correlación de Spearman en vez de la matriz de covarianzas usual. Luego, eliminar los atípicos y con la muestra limpia estimar el modelo regularizado.

- **Modelo 2:** Estandarizar las variables y calcular MD usando el vector de medianas en vez del vector de medias y la matriz de correlación de Spearman en vez de la matriz de covarianzas usual. Luego, determinar como observaciones NO atípicas los primeros $\lceil \frac{n+p+1}{2} \rceil$ individuos con las distancias más pequeñas. Finalmente, con la muestra limpia estimar el modelo regularizado.

  Ejemplo: Si $n = 1000$ y $p = 100$ entonces $\lceil \frac{n+p+1}{2} \rceil = \lceil \frac{1000+100+1}{2} \rceil = \lceil 550.5 \rceil = 551$. Es decir, debe identificar como NO atípicos los 551 individuos de la muestra con las distancias de Mahalanobis más pequeñas.

2) El conjunto de datos ‘‘`YearPredictionMSD`’’ contiene información sobre canciones de música popular y el año en que se grabaron. Incluye 515345 observaciones y 90 características, como la intensidad media del sonido, la varianza del espectro de frecuencia y la correlación entre las características espectrales. El objetivo es predecir el año en que se grabó la canción.

   - Carque el conjunto de datos usando la función `read_csv` del paquete **pandas** y el como primer argumento el link `https://archive.ics.uci.edu/ml/machine-learning-databases/00203/YearPredictionMSD.txt.zip`, use como segundo argumendo *header = None*.
   - Divida el conjunto de datos en características o variables explicativas X y variable objetivo Y, tenga en cuenta que se quiere modelar el año en que se grabó la canción.
   - Reduzca la dimensión de las variables. Para ello, use un modelo de regresión **LASSO** con un coeficiente de penalización de 10, para extraer características importantes del conjunto de variables explicativas.
   - Con el conjunto de variables reducido, ajuste un modelo de regresión OLS e interprete su significancia y su R cuadrado ajustado.
   - Revise los supuestos de los errores, y con los hallazgos del ítem anterior, concluya sobre la conveniencia de usar este modelo para predecir el año de grabación de la canción.

3) El conjunto de datos conocido como ‘‘`California Housing Dataset`’’ puede ser cargado del paquete **sklearn**. La variable objetivo es el valor medio de la vivienda para los distritos de California, expresado en cientos de miles de dólares ($100000). Este conjunto de datos se derivó del censo de EE.UU. de 1990, usando como unidad de censo el grupo de bloques. Un grupo de bloques es la unidad geográfica más pequeña para la que La Oficina del Censo de EE.UU. publica datos de muestra (un grupo de bloque generalmente tiene una población de 600 a 3000 personas).

Un hogar es un grupo de personas que residen dentro de una casa. Dado que el promedio. El número de habitaciones y dormitorios en este conjunto de datos se proporciona por hogar, estas columnas pueden tomar valores sorprendentemente grandes para grupos de bloques con pocos hogares y muchas casas vacías, como centros vacacionales.

- Lea el conjunto de datos usando la función **fetch_california_housing** del paquete **sklearn.datasets**, guardelos en una variable llamada **california_housing** y con el comando *print(california_housing.DESCR)* observe la descripción general del dataset y en especial qué es cada una de las variables de entrada.
- Separe las variables explicativas X de la variable respuesta Y, para acceder a ellas use los comandos *california_housing.data* y *california_housing.target*. Considere la conveniencia de incluir las variables Longitud y Latitud al modelo. Haga un análisis exploratorio de las correlaciones entre las variables y comente al respecto.
- Ajuste un modelo de regresión **LASSO** con un coeficiente de penalización pequeño, iterativamente ajuste este valor para eliminar variables explicativas y corregir el problema de multicolinealidad, en cada iteración calcule las correlaciones de las variables explicativas y pare cuando no se encuentren correlaciones altas.
- Ajuste un modelo de regresión OLS con las variables reducidas y revise los residuales.
- Realice una detección de atípicos usando la distancia de Mahalanobis, eliminelos y vuelva a ajustar el modelo OLS, valide los supuestos del modelo.

Pautas

- Entregar un documento de `RMarkdown/Jupyter` (en PDF) con la solución y rutinas de código empleadas (fecha máxima de entrega: Mayo 31 hasta las 23:59). Enviar por correo electrónico a ambos profesores (Intu).
- El documento a entregar debe contener todos los procedimientos, códigos y gráficos necesarios que den debida justificación a lo realizado. Además, debe estar presentado en formato artículo.
- Realizar en equipos conformados por 3-4 participantes (mandatorio).

Referencias

Harrison, D., and Rubinfeld, D. L. (1978), "Hedonic Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81 – 102.

Mahalanobis, P. C. (1936), "On the Generalized Distance in Statistics," *Proceedings of the National Institute of Science of India*, 2, 49 – 55.