Profesor: Dr. Oldemar Rodríguez Rojas
CA-0404 Modelos Lineales
Regresión Clásica, Lasso y Ridge
Fecha de Entrega: Lunes 6 de septiembre a las 2pm
Reglas generales:

1. La solución de la tarea se debe subir en el Aula Virtual, no pueden ser enviadas por correo.
2. Cada día de atraso en la entrega de la tarea implica una pérdida de 20 puntos.
3. Las tareas son estrictamente de carácter individual, tareas idénticas se les asignará cero puntos.
4. Todas las tareas tienen el mismo valor en la nota final del curso, es decir, el promedio de las notas obtenidas en la tareas será la nota final del curso.
5. Todos los ejercicios tienen el mismo valor.

Tarea Número 2

- **Pregunta 1:** Complete todas las demostraciones que quedaron pendientes en las presentaciones de la clase.
- **Pregunta 2:** Para la demostración del Teorema 12 de la presentación de la clase realice lo siguiente:

  1. Construya un ejemplo a mano 3×4, es decir, tres variables predictoras y una variable a predecir, en el que se verifique que da el mismo resultados si los β's se calculan usando la forma clásica o si calculan proyectando la variable a predecir sobre el espacio generado por las tres variables predictoras. **Nota**: Las tres variables (vectores) predictoras no deben ser desde el inicio ortogonales.
  2. Pruebe el Teorema 12.

- **Pregunta 3:** En este ejercicio usaremos la tabla de datos que viene en el archivo `Uso_Bicicletas.csv`. Este es un conjunto de datos de usuarios de la empresa de alquiler de bicicletas por horas *Capital Bike* en Washington D.C. Las columnas de la tabla de datos son:

  - `Fecha`: No es una variable.
  - `Estacion`: Estación del año, es una variable categórica ordinal por lo que se puede dejar como numérica.
  - `Hora`: Hora del día, es una variable categórica ordinal por lo que se puede dejar como numérica.
  - `Feriado`: Si el día es feriado o no, es una variable categórica.
  - `DiaSemana`: Día de la semana, es una variable categórica ordinal por lo que se puede dejar como numérica.
  - `DiaTrabajo`: Si el día es de trabajo o no, es una variable categórica.
  - `TipoClima`: Tipo de clima, es una variable categórica.

- `SensacionTermica`: Sensación térmica, es una variable numérica.
- `TemperaturaReal`: Temperatura real, es una variable numérica.
- `Humedad`; Humedad relativa, , es una variable numérica.
- `VelocidadViento`: Velocidad de viento, es una variable numérica.
- `UsuariosCasuales`: Cantidad de usuarios en día pero que no están registrados como clientes. Es una variable numérica.
- `UsuariosRegistrados`: Cantidad de usuarios en día que sí están registrados como clientes. Es una variable numérica.
- `TotalUsuarios`: Cantidad total de usuarios en día.

La variable a predecir es `TotalUsuarios`.

1. Cargue la tabla de datos en `R`. Asegúrese de codificar adecuadamente las variables y de ignorar la columna `Fecha`. Además asegúrese de seleccionar la variable `TotalUsuarios` como la variable a predecir. Use para entrenar el modelo el 80 % de los datos.
2. Calcule el resumen numérico para la variable a predecir y explique el promedio.
3. Grafique la matriz de correlación e interprete la correlación entre las variables `TotalUsuarios` y `UsuariosRegistardos`.
4. Ejecute la Regresión Lineal, observe que los coeficientes de `UsuariosRegistrados` = 0.999999999999999 y `UsuariosCasuales` = 0.999999999999998 son distintos de cero (son prácticamente 1) y los coeficientes de las demás variables son casi 0, ¿Qué significa esto?
5. ¿Que relación observa entre las variables `UsuariosRegistrados` y `UsuariosCasuales` con respecto a la variable a predecir `TotalUsuarios`? Esto implica que las variables `UsuariosRegistrados` y `UsuariosCasuales` deben ser ignoradas en la construcción de la regresión ¿Por qué?
6. Ejecute nuevamente una Regresión Lineal, pero esta vez elimine (ignore) desde `R` las variables `UsuariosRegistrados` y `UsuariosCasuales` y usando el valor absoluto de los coeficientes β determine cuáles son las 3 variables que mayor importancia tienen en la regresión.
7. Para la Regresión Lineal del item 6 interprete la Raíz Error Cuadrático Medio y el Error Relativo.
8. Para la Regresión Lineal del item 6, según la correlación entre la predicción y la variable a predecir, ¿son buenas o no las predicciones de esta regresión?
9. Corra un modelo de regresión Penalizada tipo Lasso. ¿Por qué se puede decir que prácticamente el resultado es el mismo que el de la Regresión Lineal? Justifique usando el gráfico Coeficientes y lambda y con base en los Coeficientes β.
10. Para la Penalizada Lasso del item anterior seleccione Mejor Lambda Log(x) = igual a 3. ¿Cuántos Coeficientes β se anulan? Observe que la correlación entre la predicción y la variable a predecir son casi iguales que en el caso de la Regresión Lineal, con base en esta observación ¿cuál modelo prefiere la Regresión Lineal o la Regresión Lasso (con $\log(\lambda) = 3$), justifique su respuesta. **Nota:** Debido a los procesos de optimización que se ejecutan dentro del método Lasso dos ejecuciones de este item podrían dar diferente, además podría causar que en alguna de las ejecuciones no se anule ningún β (coeficiente).

11. Corra un modelo de regresión Penalizada tipo Ridge con Mejor Lambda Log(x) = que sugiere el métoodo. ¿Qué se puede concluir con base en la correlación entre la predicción y la variable a predecir?

- **Pregunta 4:** Un cliente nos contrata para estudiar una posible oportunidad de negocio, y para ver si le es rentable quiere una predicción de las ventas potenciales de asientos de niños para autos en su tienda. Para ello hacemos uso de los datos `AsientosNinno.csv` los cual contienen detalles de ventas de asientos de niños para auto en una serie de tiendas similares a las del cliente; y además los datos incluyen variables que definen características de la tienda y su localidad. La tabla de datos está formada por 400 filas y 13 columnas. Seguidamente se explican las variables que conforman la tabla.

  - `Ventas`: Ventas de asientos de niños para autos en cada localidad, en miles de unidades (**variable a predecir**)
  - `PrecioCompt`: Precio promedio por asiento de niño cobrado por la competencia en cada localidad.
  - `Ingreso`: Nivel de ingreso promedio de los habitantes de la región, en miles de dólares.
  - `CercaniaEsc`: Variable numérica.
  - `Publicidad`: Presupuesto que asigna cada tienda a publicidad, en miles de dólares.
  - `Poblacion`: Tamaño de la población en cada región, en miles.
  - `Precio`: Precio cobrado por la tienda por los asientos de niño para auto.
  - `CalidadEstant`: Indica la calidad de ubicación de los asientos de niño en los estantes de la tienda (Bueno, Medio o Malo).
  - `Edad`: Edad promedio de los habitantes de la localidad.
  - `Educacion`: Años de educación promedio de los habitantes de cada región.
  - `Urbano`: Indica si la tienda está localizada en una zona urbana o no (1 = Sí, 0 = No).
  - `USA`: Indica si la tienda está ubicada en Estados Unidos o no (1 = Sí, 0 = No).
  - `Desarrollo`: Variable numérica.

  Realice lo siguiente:

  1. Cargue la tabla de datos en `R`. Asegure de que todas las variables categóricas que tiene esta tabla queden codificadas de forma adecuada (es decir, que `R` las esté interpretando como variables categóricas). Para medir el error tome un 25 % de la tabla de datos como tabla de testing.
  2. Según el gráfico de Correlación ¿cuál es la variable que mejor correlaciona con la variable a predecir?
  3. Ejecute una Regresión Lineal y usando el valor absoluto de los coeficientes β determine cuáles son las 5 variables que mayor importancia tienen en la regresión. ¿Está la variables que tenía mayor poder predictivo según los dos ejercicios anteriores?

4. Corra un modelo de Regresión Penalizada tipo Lasso ¿Cuáles variables se anulan y por qué? ¿Cuál modelo prefiere Regresión Lineal o Lasso? Según lo anterior y basado en los índices de calidad de ambos métodos, justifique su respuesta. **Nota:** Debido a los procesos de optimización que se ejecutan dentro del método Lasso dos ejecuciones de este item podrían dar diferente, además podría causar que en alguna de las ejecuciones no se anule ningún β (coeficiente).
5. Observe que las variables `CalidadEstant`, `Urbano` y `USA`, a pesar de ser variables categóricas, aparecen en los resultados de la regresión e incluso tienen coeficientes β asociados. Esto se debe a que internamente `R` automáticamente las convierte en *Códigos Disyuntivos Completos (Variables Dummies)*. `R` las codifica como `Urbano1` y `USA1` dado que son variables dicotómicas por lo que basta ver la categoría 1 (la categoría 0 es completamente complementaria).
6. En el caso de la variable `CalidadEstant` `R` también automáticamente la convierte en *Códigos Disyuntivos Completos (Variables Dummies)*, solo aparecen las modalidades `CantidadEstantMalo` y `CantidadEstantMedio` porque la tercera categoría `Bueno` es complementaria a las dos anteriores.

   ¿Qué sucede si en `R` forzamos a la variable `CantidadEstant` a ser variable *Código Disyuntivo Completo (Dummy)*? Nótese que en `R` cuando se recodifica una variable hay que generar de nuevo las tablas de `training-testing`.

- **Pregunta 5:**

1. Programe en **R** una función `lm2(...)` que recibe como parámetro una tabla de aprendizaje y retorna un modelo de Regresión Lineal, es decir, calcula y retorna $\beta = (X^tX)^{-1}X^ty$.
2. Programe en **R** una función `predict2(...)` que recibe como parámetro el modelo construido en la pregunta anterior, una tabla de testing de modo tal que retorna la predicción para esta tabla de testing.
3. Usando la tabla de datos `uscrime.csv` compare los resultados de las funciones `lm(...)`, `lm2(...)`, `predict(...)` y `predict2(...)`.
4. Usando la tabla de datos `uscrime.csv` y la función de **R** denominada `system.time(...)` compare los tiempos de ejecución de las funciones `lm(...)`, `lm2(...)`, `predict(...)` y `predict2(...)`.

- **Pregunta 6:** Demuestre que la Regresión Ridge puede ser obtenida mediante Regresión Lineal clásica usando una versión aumentada de la tabla de datos de la siguiente manera: Se aumenta la tabla de datos X con p filas adicionales $\sqrt{\lambda}I$; y se aumenta y con p ceros, es decir:

$$\widetilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda}I \end{bmatrix}$$

$$\widetilde{y} = \begin{bmatrix} y \\ \vec{0} \end{bmatrix}$$

donde $\vec{0} = (0, 0, \ldots, 0)^T$, o sea que $\widetilde{X}$ es de tamaño $(n+p) \times m$ que $\widetilde{y}$ es de tamaño $(n+p) \times 1$.

Mediante la introducción de datos artificiales que tienen valor de respuesta cero, el procedimiento de ajuste reduce los coeficientes de la regresión a valores cercanos cero. Está forma de

introducir funciones de penalización fue propuesta a Abu-Mostafa (1995), donde las restricciones del modelo se implementan mediante la adición datos artificiales.

- **Pregunta 7:** Considere una regresión con modelo de p parámetros, ajustado por mínimos cuadrados en un conjunto de datos de entrenamiento $(x_1, y_1), \ldots, (x_n, y_n)$ de una población tomada al azar. Sea $\hat{\beta}$ el estimado de mínimos cuadrados. Supongamos que se tienen algunos datos de prueba $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_m, \tilde{y}_m)$ tomados al azar de la misma población que los datos de entrenamiento. Si $R_{tr}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2$ y $R_{te}(\beta) = \frac{1}{m}\sum_{i=1}^{m}(\tilde{y}_i - \tilde{x}_i\beta)^2$, pruebe que:

$$E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})],$$

donde la esperanza (media) es calculada en todo el espacio aleatorio en cada expresión.

- **Pregunta 8:**

(a) Supongamos que ejecutamos una regresión Ridge con parámetro λ en una sola variable X, y se obtiene el coeficiente a. Ahora incluimos una copia exacta $X^\star = X$ y volvemos a calcular la regresión Ridge. Demuestre que ambos coeficientes son idénticos y calcule su valor. Demuestre en general que si m copias de la variable X_j son incluidas en la regresión Ridge, entonces sus coeficientes son todos iguales.

Sugerencia: Considere matrices como las siguientes:

$$X = \begin{pmatrix} x_1 & x_1 \\ \vdots & \vdots \\ x_n & x_n \end{pmatrix}$$

$$X = \begin{pmatrix} x_1 & x_1 & x_1 \\ \vdots & \vdots & \vdots \\ x_n & x_n & x_n \end{pmatrix}$$

(b) ¿Qué pasa en Regresión Lasso? ¿Ocurre lo mismo?

Entregables: Debe entregar un documento autreproducible HTML con todos los códigos y salidas, incluya pruebas de ejecución de las funciones programadas. No olvide poner un título para cada pregunta. Las demostraciones las puede entregar en papel a mano.