Curso: Métodos basados en Kernels para aprendizaje automático
Profesora: Tamara Fernández

Tarea 2

Instrucciones de entrega:

- Entregue un informe en formato **pdf** con sus resultados, pseudo codigo (si corresponde) y conclusiones. Además incluya su codigo en R o Python.
- Comente todas sus soluciones, y utilice herramientas gráficas, por ejemplo use el paquete `rgl` (u otro), si es que estima que esto le ayudará a presentar de mejor manera sus resultados.
- **Fecha de entrega:** 23 de Septiembre

IMPORTANTE: La siguiente tarea debe ser realizada sin la ayuda de ningún paquete o librería. Las únicas librerias permitidas para su realización son la librería `caret` para utilizar la función `preProcess` para estandarizar.

Ejercicio 1. Considere la base de datos `datos.txt` que contiene la siguiente información:

Nombre	Fórmula	Descripción
`pointY`	y	variable respuesta
`pointX`	x	covariables
`pointZ`	z	
`point_X1_Z1`	xz	
`point_X1_Z1`	xz	
`point_X2_Z0`	x^2	
`point_X0_Z2`	z^2	
`point_X3_Z0`	x^3	
`point_X0_Z3`	z^3	
`point_X2_Z1`	x^2z	
`point_X1_Z`	xz^2	
`kfold`	grupo en cross-validation	Número entre 1 y 5

1. En esta parte usted debe implementar un modelo de regresión lineal múltiple para predecir la variable respuesta y solo utilizando como variables independientes las siguientes covariables: x y z (correspondientes a `pointX` y `pointZ`). Su implementación debe incluir la estandarización de covariables y el uso de $k = 5$ fold cross-validation. Para utilizar k-fold cross validation ocupe los grupos en la variable `kfold`.

   Responda:

   a) Escriba en una ecuación el modelo ajustado y reporte en una tabla el estimador de mínimos cuadrados obtenido en cada iteración de cross-validation.

   b) Obtenga un gráfico de dispersión para mostrar:
   - la relación entre x y la variable respuesta y, y
   - la relación entre z y la variable respuesta y.

   En cada gráfico incluya los valores ajustados encontrados en una iteración (cualquiera) de cross-validation.

   c) Reporte en un gráfico boxplot el coeficiente R^2 y el error cuadrático medio predictivo (obtenidos en cada iteración de cross-validation). Interprete y discuta los resultados obtenidos.

2. En esta parte se le va a pedir que implemente un modelo de regresión Ridge. Para esta implementación usted debe usar todas las covariables presentes en la base para ajustar su modelo. Su implementación debe incluir la estandarización de covariables y el uso de $k = 5$ fold cross-validation. Para utilizar k-fold cross validation ocupe los grupos en la variable `kfold`.

a) Reporte una tabla con los estimadores de ridge y el parámetro de regularización λ para cada iteración de cross-validation. El parámetro de regularización λ debe ser escogido en cada iteración de cross-validation de tal forma que se minimíce el error cuadrático medio predictivo. Para esto considere valores de λ en `seq(0.1,3,0.05)`.

b) Obtenga un gráfico de dispersión para mostrar:

- la relación entre x y la variable respuesta y, y
- la relación entre z y la variable respuesta y.

En cada gráfico incluya los valores ajustados encontrados en una iteración cualquiera de cross-validation (utilizando el parámetro óptimo de regularización encontrado).

c) Reporte en un gráfico boxplot el coeficiente R^2 y el error cuadrático medio predictivo obtenidos en cada iteración de cross-validation (considerando el modelo con parámetros óptimos). Interprete y discuta los resultados obtenidos.

d) Para una iteración de cross-validation muestre el encogimiento de los parámetros a medida de que el parámetro de regularización aumenta (incremente el rango de λ de ser necesario). Basado en este gráfico, decida que variables son más relevantes en su modelo.

3. En esta parte se le va a pedir que implemente kenel ridge regression usando la función de penalización:

$$\mathcal{L}_\lambda(f) = \sum_{i=1}^{n}(y_i - f(x_i, z_i))^2 + \lambda\|f\|^2_{\mathcal{H}}, \quad f \in \mathcal{H}$$

Note que para su implementación usted solo deberá usar las covariables x y z. Su implementación debe incluir la estandarización de covariables y el uso de $k = 5$ fold cross-validation. Para utilizar k-fold cross validation ocupe los grupos en la variable `kfold`.

a) Para cada iteración de cross-validation implemente el modelo anterior utilizando la función de kernel dada por:

$$K(u, v) = \exp\left\{-\frac{\|u - v\|^2}{\ell^2}\right\} \qquad \text{donde} \quad v, u \in \mathbb{R}^2$$

Para su impementación use parámetros (ℓ, λ) óptimos en el sentido de que minimicen el error cuadrático medio predictivo. Para encontrar estos parámetros busque en una grilla con al menos 500 combinaciones de parámetros distintos. Reporte en una tabla los parámetros (ℓ, λ) escogidos en cada iteración de cross-validation.

b) Obtenga un gráfico de dispersión para mostrar:

- la relación entre x y la variable respuesta y, y
- la relación entre z y la variable respuesta y.

En cada gráfico incluya los valores ajustados encontrados en una iteración de cross-validation (considerando los parámetros óptimos encontrados en cada iteración).

c) Reporte en un gráfico boxplot el coeficiente R^2 y el error cuadrático medio predictivo. Interprete y discuta los resultados obtenidos. Incluya en su discusión la comparación con los modelos anteriores.