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:** 20 de Septiembre

Ejercicio 1 (Regresión ridge)**.** Considere datos $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, donde para cada individuo indexado en $i = 1, \ldots, n$, tenemos una variable respuesta y_i, y un vector de covariables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$ de dimensión d. Además, considere el modelo lineal que relaciona la respuesta y_i con el vector de covariables $\boldsymbol{x}_i$:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot x_{i1} + \ldots + \beta_d \cdot x_{id} + \epsilon_i \\ &= \beta_0 + \boldsymbol{x}_i^{\intercal} \boldsymbol{\beta} + \epsilon_i, \end{aligned}$$

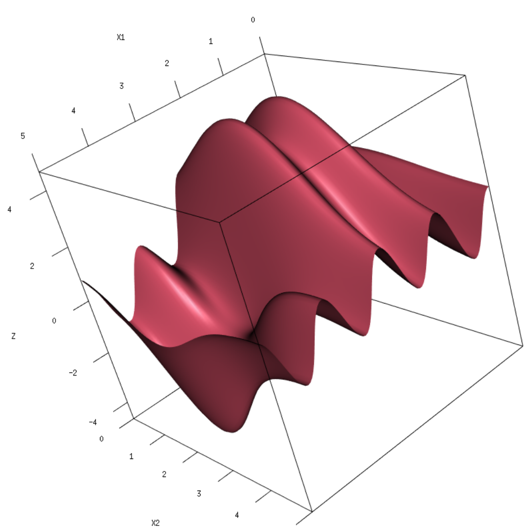
donde $\epsilon_1, \ldots, \epsilon_n$ son ruidos independientes con media 0, y varianza constante.

1. Sea $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$, y considere la función de perdida dada por:

$$L_\lambda(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \boldsymbol{x}_i^{\intercal} \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|^2$$

   Encuentre explicitamente (i.e., resuelva el problema de optimización derivando e igualando a cero) estimadores $\widehat{\beta}_0$ y $\widehat{\boldsymbol{\beta}}$ que minimizen la función de perdida anterior. (Note que este es el problema estándar resuelto en Ridge Regression, pero donde no se penaliza al intercepto). Discuta que ocurre cuando i)$\lambda = 0$, ii)$\lambda \to \infty$.

Ejercicio 2 (Kernel Ridge Regression)**.** En este ejercicio trabajaremos con datos simulados que se encuentran en el archivo `sim.txt`. Este archivo contiene información sobre 3 variables: `y` (variable respuesta), y `X1,X2` (covariables 2-dimensionales asociadas). La función que genero los datos luce de esta forma:



1. Implemente un modelo de Kernel Ridge Regression basado en la función de perdida:

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

   donde $\mathcal{H}$ es el RKHS asociado a la función de kernel dada por el kernel squared-exponential $K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\ell^2}\right\}, \quad \text{donde } \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^2.$$

   En su implementación elija $\ell = 0{,}5$ y $\lambda = 1$. Evalue la estimación del modelo en terminos del R cuadrado, y del error cuadratico medio predictivo. Para esto último considere $k = 10$ cross-validation. Muestre sus resultados ¿Qué opina de los resultados obtenidos?

2. Implemente el mismo modelo anterior, pero esta vez encuentre parámetros (ℓ, λ) óptimos en el sentido de que minimicen el error cuadrático medio predictivo. Utilizando estos parámetros óptimos vuelva a evaluar su estimación en terminos del R cuadrado, y del error cuadratico medio predictivo. Repita este procedimiento usando $k = 10$ cross-validation y compare con el item anterior. Comente sobre sus resultados.

   Hint 1: Para encontrar los valores óptimos de los parámetros cree una grilla de valores para (ℓ, λ).

   Hint 2: Para una comparación más justa, use los mismos conjuntos para $k = 10$ cross-validation en este, y en el item anterior.

3. Note que en el item anterior, al usar k=10 cross-validation, se obtiene una colección de 10 parámetros optimos $(\ell_1, \lambda_1), \ldots, (\ell_{10}, \lambda_{10})$ (uno por cada set de testeo). ¿Qué puede decir sobre la distribución de estos parámetros? (utilice algún gráfico para mostrar la distribución) ¿Son muy distintos de los ocupados en el item 1?

4. Una generalización del squared exponential kernel $K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ esta dada por la siguiente ecuación:

$$\begin{aligned} K(x, x') &= \exp\left\{-(\boldsymbol{x} - \boldsymbol{x}')^\intercal \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{x}')\right\} \\ &= \exp\left\{-\left(\frac{x_1 - x_1'}{\ell_1}\right)^2 - \left(\frac{x_2 - x_2'}{\ell_2}\right)^2\right\} \end{aligned}$$

   donde $\boldsymbol{x} = (x_1, x_2)$, $\boldsymbol{x}' = (x_1', x_2')$ y $\Sigma = \begin{pmatrix} \ell_1^2 & 0 \\ 0 & \ell_2^2 \end{pmatrix}$, (note que esta generalización permite 2 parametros de length-scale distintos). Repita el item 2, pero esta vez use la generalización del squared-exponential kernel y busque parámetros óptimos $(\ell_1, \ell_2, \lambda)$ (en esta ocasión necesitará buscar en una grilla 3 dimensional). Comente los resultados obtenidos. Según los resultados que encuentre, comente porque utilizar esta generalización podría ser ventajoso, o desventajoso.