**Curso:** Métodos basados en Kernels para aprendizaje automático. 2023
**Profesora:** Tamara Fernández

# Tarea 3.

**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:** 19 de Octubre
- Note que los puntos base de Tarea suman más que 6 puntos. En el caso de obtener más de 6 puntos, los puntos extras pueden ser utilizados en la Tarea 1 o en la Tarea 2.

**Ejercicio 1** (Explorando kernels, 3 puntos)**.** En este ejercicio el objetivo es explorar el comportamiento de diferentes kernels en la solución de kernel regression. Para esto considere la base de datos `starbucks` del paquete `openintro`, y considere la variable `carb` como variable respuesta y la variable `protein` como variable independiente. Para este ejercicio considere estandarización de la variable independiente y utilice toda la base de datos como set de entrenamiento.

1. Considere la función de **kernel constante** dada por:

$$K_c(x, y) = c, \qquad c > 0$$

   *a*) Justifique que el kernel anterior es efectivamente un kernel (**Hint**: utilice las propiedades dadas en el capítulo de algebra de kernels).

   *b*) Utilice el kernel anterior para ajustar un modelo de regresión con kernels utilizando las variables descritas en el enunciado. Elija el parámetro c, de tal forma que se minimice el error cuadrático medio calculado en el set de entrenamiento (considere al menos 100 candidatos). Elija apropiadamente su parámetro de regularización $\lambda$.

   *c*) Grafique los valores ajustados obtenidos para su modelo.

   *d*) Repita los items a-c para los kernels dados por:

      1) kernel lineal

$$K_c(x, y) = 1 + c^2 xy, \qquad c > 0$$

      2) kernel polinomial (grado 2)

$$K_c(x, y) = (c + xy)^2, \qquad c > 0$$

      3) kernel gaussiano

$$K_c(x, y) = \exp\{-(x - y)^2/c^2\}, \qquad c > 0$$

**Ejercicio 2** (Kernel Browniano)**.** En este ejercicio nuestra inspiración es el **movimiento browniano**. El movimiento browniano es el movimiento aleatorio y descontrolado de una particula, que se origina debido a constantes interacciones a escala microscópica con el ambiente. El movimiento browniano es muy importante en economía, física, y en la ciencia en general, ya que este moviemiento se ha observado en una gran variedad de fenómenos, por ejemplo: en el movimiento de particulas de polen en el agua, en el modelamiento de portafolios en economía, en astronomía un estrella masiva puede experimentar el movimiento browniano debido a la interacción gravitacional con estrellas de menor tamaño, etc.

Sea $X_t$ la posición de una particula, que se mueve en los números reales a tiempo $t$. Si la particula sigue una trayectoria browniana, entonces la covarianza entre las posiciones a tiempo $t$ y $s$, es decir, $X_t$ y $X_s$, está dada por $Cov(X_t, X_s) = K(t, s)$, donde la función $K : [0, \infty) \times [0, \infty) \to \mathbb{R}$ es:

$$K(t, s) = \sigma^2 \min\{t, s\},$$

que se conoce como el kernel browniano.

1. (2 puntos) La primera parte de este ejercicio, tiene como objetivo verificar que $K$ es un kernel (usando la definición de kernel que aprendimos en clases, es decir, simétrico y semi-definido positivo).

   a) Sea $g : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}$ una función que se evalua en una covariable $\boldsymbol{x}$, y en un número real $t \in [0, \infty)$.
   Defina $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ por

   $$K(\boldsymbol{x}, \boldsymbol{x}') = \int_0^1 g(\boldsymbol{x}, t) g(\boldsymbol{x}', t) dt$$

   Verifique que $K$ es kernel: es decir que es simetrico y semi-definido positivo. **Hint:** para verificar que es semi-definido positivo use directamente la definición y manipule adecuadamente los términos.

   b) Muestre que $K(x, x') = \min(x, x')$ es un kernel, donde $x$ y $x'$ son números reales no-negativos.
   **Hint:** Dados $b \geq 0$, defina la función $g_b : [0, \infty) \to \mathbb{R}$, tal que $g_b(t)$ que toma valor 1 si $t$ está en el intervalo $[0, b]$, si no toma el valor 0. Use que $\int_0^\infty g_b(t) dt = b$. Combine con el ejercicio anterior.

   c) (Caso particular) Muestre que la función dada por $J : [0, 1) \times [0, 1) \to \mathbb{R}$ dado por

   $$J(x, x') = \frac{1}{1 - \min(x, x')}$$

   es un kernel.
   **Hint:** recuerde la suma geométrica aprendida en su curso de cálculo.

2. (2 puntos) Considere los datos del archivo ”datos.txt”. Donde se encuentran muestras de $t \to f(t)$ para distintos valores de $t$. La primera columna contiene tiempos $t$ y en la segunda contiene el valor $f(t)$.

   a) Utilize kernel regression (con parámetro de regularización $\lambda = 0{,}01$), usando el kernel Squared-exponential

   $$K_1(\boldsymbol{x}, \boldsymbol{x}') = \exp(-(\boldsymbol{x} - \boldsymbol{x}')^2 / \ell^2)$$

   y usando el kernel browniano

   $$K_2(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \min(\boldsymbol{x}, \boldsymbol{x}').$$

   Encuentre parametros $\sigma^2$ y $\ell^2$ de tal forma que, en ambos casos, el **error cuadrático medio** calculado en el set de entreamiento sea del orden de $10^{-6}$, es decir, su error debe ser del tipo 0,00000x, es decir, 5 ceros antes de su primera cifra significativa.

   **Ojo 1:** Se pide calcular el error cuadratico medio, **no** el error cuadratico medio predictivo, asi que use todas las observaciones que consideró en el set de entrenamiento para calcularlo.

   **Ojo 2:** Note que en este ejercicio **no se pide hacer cros-validation**, asi que puede usar toda la base de datos como set de entrenamiento.

   b) Grafique los mapas de calor de los kerneles $K_1$ y $K_2$ con los parámetros encontrados y algunas funciones asociadas al RKHS ¿Qué opina sobre los resultados? ¿Qué kernel le parece más confiable para este problema?