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

Tarea 1

Instrucciones de entrega:

- **Entregue un informe en formato pdf** con sus resultados, pseudo codigo (si corresponde) y conclusiones. Además incluya su codigo en R.
- **Fecha de entrega: 26 de agosto 2022.**
- **Puntaje:** La primeras 3 preguntas son obligatorias y suman un total de 60 puntos (se consideran 10 puntos bases). La cuarta pregunta es trabajo adicional y **opcional**. En el caso de que la nota de la tarea exceda los 70 puntos (con punto base), la nota final es un 70.

Ejercicio 1 (15 puntos). Funciones y graficos en R: En este ejercicio introducimos la función de kernel llamada **Locally Periodic**, definida por:

$$K_{LP}(x, x') = \sigma^2 \exp\left\{-\frac{2\sin(\pi\|x - x'\|/p)^2}{\ell^2}\right\} \exp\left\{-\frac{\|x - x'\|^2}{2\ell^2}\right\}, \quad \text{donde} \quad x, x' \in \mathbb{R}^2$$

1. Escriba una función en R que le permita evaluar el kernel Locally Periodic.
2. Grafique el comportamiento de la función anterior en terminos de $\|x - x'\|$. Para esto use $\sigma = 1$, $p = 0.5$ y $\ell = 1$.
3. Repita el item anterior para $\ell = 10$. Comente lo que observa.

Ejercicio 2 (15 puntos). Ciclos en R: Considere dos matrices de **A** y **B** de dimensiónes $n \times n$ y $n \times 50$, respectivamente.

1. Escriba un método/función basado en ciclos `for` que le permita calcular **AB** (multiplicación usual de matrices). El método debe recibir como parámetros de entrada las matrices **A** y **B**.
2. Investigue la función `system.time` para calcular el tiempo de ejecución para valores de n entre 10 y 10000. Grafique n vs el tiempo de ejecución, y comparelo con el tiempo de ejecución obtenido al usar `A%*%B` (multiplicación nativa de R). Comente.

Ejercicio 3 (30 puntos). Regresión lineal: En las siguientes preguntas trabajaremos con la base de datos `starbucks` del paquete `openintro`, que reune datos nutricionales para 77 productos vendidos por Starbucks. Esta base de datos contiene la siguiente información:

- `item`: Nombre del producto vendido.
- `calories`: Número de calorias.
- `fat`: Grasas (gramos).
- `carb`: Carbohidratos (gramos).
- `fiber`: Fibra (gramos).
- `protein`: Proteinas (gramos).
- `type`: Tipo de producto: `bakery, bistro box, hot breakfast, parfait, petite, salad,` y `sandwich`

Parte I: Regresión lineal simple: En esta parte del ejercicio, el objetivo es modelar la cantidad de proteinas (`protein`) de un producto en terminos de sus calorias (`calories`).

1. Use la función `sample` para dividir la base de datos en $k = 5$ conjuntos disjuntos, elegidos al azar, de tamaño lo más parecido posible. Agregue en una columna nueva una variable que indique a que grupo pertenece cada observación (fila).

2. Utilizando el item anterior, cree un conjunto de datos de entrenamiento y testeo utilizando 4 y 1 grupos respectivamente.

3. Utilizando el set de entrenamiento creado en el item anterior, estime los coeficientes de regresión utilizando:

   (a) La ecuación que entrega la solución del problema de mínimos cuadrados

   (b) La función `lm()`

   (c) Repita la parte a) usando la columna `calories` **estandarizada**. Explique como se relacionan los parámetros estimados con los del item a). Justifique esta relación encontrando ecuaciones que la expliquen.

   En todos los casos considere un intercepto. ¿Qué concluye al comparar las soluciones en a), b) y c)? Basado en sus modelos de regresión, ¿En cuántas proteinas aumenta un producto por cada caloria adicional? ¿Es esta cantidad la misma en todos los modelos?

4. Grafique las rectas de regresión para los casos a) y c). Compare.

5. Utilice los estimadores de minimos cuadrados encontrados en los items 3.a) y 3.c) para predecir la variable `protein` en el set de testeo. Recuerde que en el caso de haber estandarizado en el set de entrenamiento, debe aplicar exactamente la misma estandarización a sus datos en el set de testeo. Repita los gráficos del item anterior, pero esta vez agregue los valores predichos por su modelo y los valores reales para las observaciones en el conjunto de datos de testeo.

6. Para los modelos estimados en 3.a) y 3.c) calcule el error cuadrático medio predictivo dado por la siguiente formula

$$ECMP = \frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i - \bar{\boldsymbol{x}}_i^{\intercal}\widehat{\boldsymbol{w}})^2 \tag{1}$$

   donde $(y_1, \bar{\boldsymbol{x}}_{\mathbf{1}}), \ldots, (y_N, \bar{\boldsymbol{x}}_{\boldsymbol{N}})$ son los datos correspondientes al set de testeo, y donde $\widehat{\boldsymbol{w}}$. Recuerde que en el caso de haber estandarizado, sus datos en el set de testeo deben considerar la misma estandarización. Qué le parece este error?

7. Repita el item anterior $k = 5$ veces de tal forma que cada uno de los 5 grupos que definió en el item 1 sea ocupado como set de testeo. Utilice boxplots para graficar el error cuadrático medio predictivo. Comente lo que observa.

Parte II: Regresión lineal multiple: Un investigador plantea modelar el número de proteinas encontradas en un producto en términos de su grasa (`fat`), carbohidratos (`carb`), fibra (`fiber`) y calorias (`calories`). Para ello propone el siguiente modelo

$$\texttt{protein} = w_0 + w_1 \cdot \texttt{fat} + w_2 \cdot \texttt{carb} + w_3 \cdot \texttt{fiber} + w_4 \cdot \texttt{calories} + \epsilon,$$

donde ϵ denota un error normal de media 0 y varianza σ^2.

1. El investigador posee conocimiento experto que le permite determinar que el parámetro poblacional w_2 cumple con $w_2 = -0.8$. Use esta información, y la función de pérdida dada por el Error Cuadrático Medio, para establecer el problema de optimización que debe plantear para encontrar estimadores para (w_0, w_1, w_3, w_4).

2. Utilice el estimador encontrado en el item anterior para calcular el error cuadrático medio predictivo definido en la ecuación (1). Para resolver esta pregunta use $k = 5$ fold cross-validation.

Ejercicio 4 (10 puntos). Pregunta Opcional: El puntaje de esta pregunta es extra y se suma al puntaje total de la tarea.

1. Muestre que si la matriz $\mathbf{X}$ es de rango completo, entonces $\mathbf{X}^\intercal\mathbf{X}$ también tiene rango completo.

2. Sea $\mathbf{A}$ una matriz de $d \times d$ cuyas entradas dependen en un escalar θ. La derivada de $\mathbf{A}$ respecto a θ se define mediante

$$\frac{\partial}{\partial\theta}\mathbf{A} = \begin{pmatrix} \frac{\partial}{\partial\theta}\mathbf{A}_{11} & \frac{\partial}{\partial\theta}\mathbf{A}_{12} & \cdots & \frac{\partial}{\partial\theta}\mathbf{A}_{1d} \\ \frac{\partial}{\partial\theta}\mathbf{A}_{21} & \frac{\partial}{\partial\theta}\mathbf{A}_{22} & \cdots & \frac{\partial}{\partial\theta}\mathbf{A}_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial\theta}\mathbf{A}_{d1} & \frac{\partial}{\partial\theta}\mathbf{A}_{d2} & \cdots & \frac{\partial}{\partial\theta}\mathbf{A}_{dd} \end{pmatrix}$$

   Suponga que $\mathbf{A}$ tiene inversa dada por $\mathbf{A}^{-1}$. Muestre que

$$\frac{\partial}{\partial\theta}\mathbf{A}^{-1} = -\mathbf{A}^{-1}\left(\frac{\partial}{\partial\theta}\mathbf{A}\right)\mathbf{A}^{-1}.$$