Métodos avanzados en Estadística **Curso 2021/2022**
Máster en Ciencia de Datos

Ejercicios: estimación no paramétrica de funciones de densidad

1. Sea X^* una v.a. cuya función de densidad es un estimador del núcleo con parámetro de suavizado h y núcleo gaussiano. Calcula la esperanza y la varianza de X^*.

2. Realiza un análisis descriptivo de los datos británicos de ingresos familiares en 1975 (reescalados dividiendo por la media) contenidos en el fichero Datos-ingresos.txt. En concreto, calcula su media, mediana y desviación típica, y un estimador del núcleo de la función de densidad. Comenta los resultados.

3. Los datos del fichero lipidos.txt corresponden a la concentración de colesterol y triglicéridos (mg/dl) en pacientes evaluados por tener un dolor en el pecho. De estos pacientes, 51 no presentaron evidencia de enfermedad cardiaca mientras que 320 sí la presentaron.

(a) Representa un estimador del núcleo de la densidad para la variable correspondiente a la concentración de triglicéridos (utiliza primero todos los datos y después trata por separado a los individuos sanos y enfermos). Experimenta utilizando distintos núcleos y distintos valores del parámetro de suavizado. Comenta el resultado.

(b) Representa un estimador del núcleo bidimensional para el vector de variables correspondiente a las concentraciones de triglicéridos y colesterol, tratando por separado los datos de los individuos enfermos y los sanos. Comenta el resultado.

4. Sea $\hat{f}(x)$ un estimador del núcleo con parámetro de suavizado h y núcleo simétrico K. Demuestra que:

$$\int \hat{f}^2(x)dx = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} K * K\left(\frac{X_j - X_i}{h}\right),$$

donde $K * K(u) = \int K(u-y)K(y)dy$ es la convolución de K.

5. Sea F_n la función de distribución empírica correspondiente a una muestra $X_1, \ldots, X_n$ y sea $X_1^*, \ldots, X_n^*$ una remuestra extraída de F_n. Se calcula un estimador del núcleo $\hat{f}_h^*(x)$ a partir de $X_1^*, \ldots, X_n^*$ y otro $\hat{f}_h(x)$ a partir de $X_1, \ldots, X_n$. Demuestra que

$$\mathrm{E}_{F_n}\left[(\hat{f}_h^*(x) - \hat{f}_h(x))|X_1, \ldots, X_n\right] = 0.$$

¿Qué implica este resultado sobre el uso del bootstrap para estimar el sesgo de los estimadores del núcleo?

6. Sean $X_1, \ldots, X_n$ v.a.i.i.d. de una distribución con densidad f. Se considera el estimador del núcleo $\hat{f}$ con núcleo rectangular $K(x) = \mathbb{I}_{[-1/2,1/2]}(x)$ y parámetro de suavizado h.

(a) Calcula el sesgo y la varianza de $\hat{f}(x)$, para un valor de x fijo.
(b) Demuestra que tanto el sesgo como la varianza tienden a cero si $h \to 0$ y $nh \to \infty$.

7. Considera una variable aleatoria con distribución beta de parámetros $\alpha = 3$, $\beta = 6$.

(a) Representa gráficamente la función de densidad y la función de distribución.
(b) Simula una muestra de tamaño 20 de esta distribución. A continuación, representa en los mismos gráficos del apartado (a) las estimaciones de F y f obtenidas respectivamente mediante la función de distribución empírica F_n y un estimador del núcleo $\hat{f}$ obtenidos a partir de la muestra simulada.
(c) Verifica empíricamente el grado de aproximación alcanzado en las estimaciones de F y f. Para ello, genera 200 muestras de tamaño 20 y para cada una de ellas evalúa el error (medido en la norma del supremo, es decir, el máximo de las diferencias entre las funciones) cometido al aproximar F por F_n y f por $\hat{f}$. Por último, calcula el promedio de los 200 errores obtenidos.