Introducción a la Estadística y Ciencia de Datos – Primer cuatrimestre 2024

Trabajo Práctico Final – Regresión No Paramétrica

Este trabajo práctico es de carácter obligatorio, y la nota formará parte de la calificación final de la materia. Se debe entregar un informe en formato `pdf` con la resolución y resultados del ejercicio, incluyendo todos los gráficos que crean pertinentes y el archivo `.Rmd` donde se realizaron los cálculos y se programó la implementación del análisis pedido.

Los archivos deben llevar el nombre de los autores en orden alfabético, por ejemplo: Alen-Bianco-Parada.Rmd. Se debe respetar la formación de los grupos inscriptos.

1. Teórico

Consideremos una muestra de n vectores aleatorios independientes que cumplen $(X_i, Y_i) \sim (X, Y)$, donde $X, Y \in \mathbb{R}$ y

$$E(Y|X = x) = m(x)\,,$$

siendo m una función a valores reales suave. Dado un núcleo K que satisface las propiedades a)-e) vistas en clase teórica y una ventana h, consideremos el estimador de Nadaraya–Watson de m dado por

$$\widehat{m}_h(x) = \sum_{i=1}^{n} Y_i \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{\ell=1}^{n} K\left(\frac{X_\ell - x}{h}\right)} = \sum_{j=1}^{n} Y_i w_{i,h}(x)\,. \tag{1}$$

Llamemos $\mathbf{Y}$ al vector de respuestas con $i-$ésima componente dada por Y_i.

(a) Llamemos $\widehat{Y}_i$ a los valores predichos, siendo $\widehat{Y}_i = \widehat{m}_h(X_i)$ y sea $\widehat{\mathbf{Y}}$ el vector de predichos donde la $i-$ésima componente es $\widehat{Y}_i$. Probar que $\widehat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$, es decir que se trata de una transformación lineal en $\mathbf{Y}$, y hallar una expresión para $\mathbf{S}$ definiendo claramente cada una de sus componentes.

(b) El método de validación cruzada tiene como función objetivo a

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_h^{-i}(X_i))^2$$

para la minimización en h, donde el estimador $\widehat{m}_h^{-i}(\cdot)$ se computa como (1) pero sin la observación (X_i, Y_i). A primera vista, el cálculo de CV es computacionalmente alto, sin embargo una simplificación muy relevante está dada por la siguiente igualdad

$$\widehat{m}_h^{-i}(X_i) = \frac{\widehat{m}_h(X_i) - Y_i\, w_{i,h}(X_i)}{1 - w_{i,h}(X_i)}\,. \tag{2}$$

De esta manera, se obtiene que

$$\mathrm{CV}(h) \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{m}_h(X_i))^2}{(1 - w_{i,h}(X_i))^2}\,, \tag{3}$$

con lo cual no es necesario recalcular el estimador para cada observación.

Probar las igualdades (2) y (3).

2. Práctico

En el archivo `individuals.csv` se encuentran datos de un estudio antropométrico en 3900 niños realizado en 1977. Se tomaron un total de 87 medidas corporales tradicionales y funcionales en una muestra de bebés, niños y jóvenes que representan la población estadounidense de a lo sumo 20 años a mediados de la década del setenta.

Fuente: math.nist.gov/ SRessler/anthrokids/).

En particular, las variables `BUTTOCK.KNEE.LENGTH` y `HIP.CIRCUMFERENCE` corresponden a las medidas de la longitud del fémur y de la circunferencia de la cadera, ambas expresadas en milímetros. La variable `SEX` reporta el sexo de los individuos, siendo `1` el masculino y `2` el femenino.

De aquí en más consideraremos la población femenina.

(a) Cargue los datos y realice un diagrama de dispersión de `HIP.CIRCUMFERENCE` (eje x) vs. `BUTTOCK.KNEE.LENGTH` (eje y) para las observaciones que corresponden al sexo femenino. Considere remover las observaciones con valor igual a 0. ¿Qué sugiere este gráfico?

(b) Sea $X_1, \ldots, X_n$ una muestra aleatoria donde $, X_i \sim F$, siendo F estrictamente creciente con densidad f. Sea x_p tal que $F(x_p) = p$ el $p-$ésimo cuantil de la distribución. Supongamos que estimamos el cuantil x_p mediante estadísticos de orden: $X_{[np]}$.

Bajo las condiciones de regularidad dadas, se puede demostrar que estos estimadores son asintóticamente normales, es decir,

$$\sqrt{n}(X_{[np]} - x_p) \xrightarrow{D} N\left(0, \frac{p(1-p)}{[f(x_p)]^2}\right) .$$

Así, si $\tilde{X}_n$ es la mediana muestral y $\tilde{\mu}$ la mediana poblacional, tenemos que

$$\sqrt{n}(\tilde{X}_n - \tilde{\mu}) \xrightarrow{D} N\left(0, \frac{1}{[4f(\tilde{\mu})]^2}\right) .$$

Interesa la distribución de la longitud del contorno de cadera de la población femenina en distintos grupos etários (o sea de edad), para lo cual consideraremos la edad en meses registrada en `AGE.IN.MONTHS`.

Teniendo en cuenta el primer, segundo y tercer cuartil de la variable `AGE.IN.MONTHS` forme 4 grupos etários y estime la mediana de `HIP.CIRCUMFERENCE` en cada uno de ellos. ¿Cuáles son sus estimaciones?

Calcule un intervalo de confianza *bootstrap* normal de nivel 0.95 para cada una de las 4 medianas y realice un gráfico que resulte ilustrativo.

Comente brevemente cómo implementa los intervalos normales bootstrap solicitados.

(c) Se quiere evaluar un ajuste de regresión para las variables `HIP.CIRCUMFERENCE` (x) y `BUTTOCK.KNEE.LENGTH` (y).

i) Investigue el comando `ksmooth` de `R` y ajuste una regresión no paramétrica usando el núcleo normal. Utilice los argumentos `bandwidth`=100 y `bandwidth`=50. Agregue al gráfico del ítem (a) las dos regresiones obtenidas mediante el comando `lines` usando distintos colores para cada ventana. ¿Qué sugiere el gráfico obtenido? ¿Cuál de las dos ventanas usaría?

ii) Implemente un código que realice la búsqueda de la ventana óptima para `ksmooth` con núcleo normal para el parámetro `bandwidth`. Utilice el criterio de convalidación cruzada basado en *leave-one-out* y realice la búsqueda en una grilla de `bandwidth` entre 20 y 50 con paso 1. Grafique la función objetivo y represente allí el parámetro óptimo hallado de acuerdo al criterio que está utilizando.

iii) Implemente una función `nwsmooth` para el cómputo del estimador de Nadaraya–Watson dado en (1) con núcleo normal. Utilice el criterio de convalidación cruzada basado en *leave-one-out* y realice la búsqueda en una grilla de ventanas. Decida si mantener la misma grilla o no y justifique su decisión.

Agregue al gráfico del ítem (a) la regresión obtenida mediante el comando `lines` con un nuevo color.

iv) Vuelva a realizar el diagrama de dispersión de (a) y superponga la estimación de la regresión no paramétrica que obtuvo con la ventana óptima, tanto en su implementación de `nwsmooth` del ítem iii) como en la de `ksmooth` del ítem iii). Asimismo, superponga la recta que obtiene utilizando el método de mínimos cuadrados. Compare y concluya acerca de qué ajuste propondría.

(d) Como se describió en clase, el estimador de Nadaraya-Watson $\widehat{m}_h(X_i)$ definido en (1) puede pensarse como la mejor constante que ajusta localmente a la regresión en el punto x, es decir como solución de

$$\underset{a}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - a)^2 K\left(\frac{X_i - x}{h}\right) . \tag{4}$$

Esta idea podría generalizarse y podríamos buscar el mejor polinomio local, así por ejemplo, podríamos encontrar la mejor recta local. De esta manera, buscaríamos un nuevo estimador mediante

$$(\widehat{a}_0, \widehat{a}_1) = \underset{a_0,a_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - a_0 - a_1(X_i - x))^2 K\left(\frac{X_i - x}{h}\right) \tag{5}$$

y definiríamos el estimador lineal local como

$$\widetilde{m}_h(x) = \widehat{a}_0 .$$

En la Sección 2.2 del trabajo de Loader (2004) que adjuntamos se muestra cómo se puede calcular matricialmente este estimador.

i) Implementar una función `linearsmooth` que calcule el estimador lineal local de la regresión usando un núcleo normal para un conjunto de datos dado y una ventana determinada.

ii) Aplicar el estimador implementado a los datos de `HIP.CIRCUMFERENCE` (x) y `BUTTOCK.KNEE.LENGTH` (y) con una ventana $h = 40$ (Aclaración: este valor de la ventana corresponde a la selección por validación cruzada de nuestra implementación y se las brindamos ya que es MUY lenta la determinación).

iii) Graficar en un mismo *plot* el estimador de Nadaraya–Watson obtenido con la ventana óptima hallada en el ítem (c) iii) y el estimador del ítem anterior. Comparar.