Tarea 2 de Aprendizaje Máquina, 2022

Universidad Iberoamericana

09 de marzo de 2022

1. Construya un núcleo de orden 2 con soporte en $[-1, 1]$ y que sea un polinomio de grado 2.

2. Sea $K = \mathbb{1}_{[-1/2,1/2]}(x)$ y suponga que se cuenta con una muestra $X_1, \ldots, X_n$ provenientes de una distribución con densidad f.

   (a) Obtenga el estimador de densidad kernel y demuestre que este estimador puede ser escrito de la forma:
   $$\hat{f}_h(x) = \frac{1}{nh} B$$
   donde B es una variable aleatoria binomial con parámetros n y p, tal que
   $$p = \int_{x-h/2}^{x+h/2} f(u)du$$

   (b) A partir del inciso anterior y usando el teorema del valor medio deduzca que existe $x^\star \in [x - h/2, x + h/2]$ tal que
   $$\mathbb{E}\left[\hat{f}_h(x)\right] = f(x^\star)$$
   y
   $$\mathbb{V}\left[\hat{f}_h(x)\right] = \frac{1}{n} f(x^\star)\left[\frac{1}{h} - f(x^\star)\right]$$

   (c) A partir del inciso anterior podemos deducir que si f cumple ser suave y sin cambios bruscos, el estimador kernel será insesgado y su varianza tenderá a cero cuando $n \to \infty$, siempre que n tienda a infinito más rápido de lo que h tiende a cero.

3. Para el siguiente ejercicio necesitará la base de datos `ethanol` que se encuentra en la paquetería `lattice` de `R`. Los datos registrados corresponden con la emisión de NO de ciertos motores así como razón de aire/etanol usado (E) y la compresión del motor (C). Se desea estimar el NO emitido en función de E y C. Construya un modelo aditivo generalizado para este problema. ¿Cuánto NO emitiría un motor si $E = 0.9$ y $C = 12$? ¿Qué cantidad de esas emisiones serían debido al valor de E y qué cantidad debido al valor de C?

4. Los datos `bones_mineral_density.csv` son mediciones relacionadas con la absorción de minerales en niños, más información sobre los datos se encuentra en la liga `https://web.stanford.edu/~hastie/ElemStatLearn/datasets/bone.info.txt`. Es de particular interés estimar la variable `spnbmd` en función de la edad y el género del niño,la variable `idnum` es un identificador del niño, que omitiremos en este ejercicio.

(a) Considere las observaciones correspondintes a niños, estime la función de densidad de la variable `spnbmd` por el método de histograma y por el método de estimador kernel y preséntelos en una misma gráfica. Repita el ejercicio para las observaciones correspondientes a niñas.

(b) Realice un gráfico de dispersión, donde en el eje horizontal se encuentre la edad de los niños y en el eje vertical la variable `spnbmd` y discriminando por el género del niño. Es decir ponga en un color distinto los puntos correspondientes a niños y los puntos correspondientes a niñas.

(c) Para cada uno de los grupos ajuste splines cúbicos con 7 nodos igualmente espaciados entre 11.25 y 23.5, incluyendo estos valores. Realice nuevamente el gráfico de dispersión y muestre sobre el mismo gráfico los estimadores obtenidos.

(d) Repita el inciso anterior pero usando los métodos de splines suavizados, polinomios locales y el estimador de Nadaraya-Watson. *Hint:* Recuerde que el estimador de Nadaraya-Watson es un caso particular de estimador de polinomios locales. Los parámetros de estos estimadores los puede ajustar por validación cruzada o a "ojo" variando el valor del parámetro y viendo cuál ajusta mejor a los datos.

(e) Comente brevemente sus observaciones, diga similitudes o diferencias que encuentre entre los distintos métodos. Si tuviera que presentar estos resultados ante un comité médico, ¿qué método preferiría? ¿Cambiaría su respuesta si el resultado fuera a ser utilizado de manera interna por una máquina?

Fecha de entrega: 23 de marzo de 2022.