UNIVERSIDAD EAFIT

Estadística No Paramétrica en Ciencia de los Datos
Taller 2, Semestre 2024 1

Conjuntos de datos para el taller:

- En el archivo *"Temperaturas.xlsx"* se tiene el comportamiento de la temperatura diaria de diferentes provincias en la
- En el archivo de *"accidentes_barrio.csv"* se encuentra para diferentes barrios de la ciudad de Medellín, para diferentes horas del día, la cantidad de accidentes acumulados ocurridos en dicha hora para cada barrio.

Utilizando los conjuntos de datos disponibles, de solución a las siguientes preguntas *(**Nota:** todos los puntos tienen un mismo peso porcentual en la calificación)*

1. Del conjunto de datos de Temperatura, realice la selección de una provincia y realice la estimación de la densidad. Para esta estimación considere:
    a. Tabla de frecuencias
    b. Estimación utilizando al menos 2 Kernels. Al variar el ancho de banda, ¿qué puede concluir de su efecto al estimar la densidad?
    c. Mediante la tabla de frecuencias y los Kernels ajustados, estime la probabilidad de que aleatoriamente se obtenga un valor menor o igual a la media, mediana y moda de las temperaturas de la provincia. ¿Con dichas probabilidades considera que hay algún sesgo en los datos?
    Hint: Para estimar la probabilidad utilizando la tabla de frecuencias, asuma una distribución uniforme para cada clase.

2. Utilizando la densidad estimada en el punto anterior, realice la simulación de 1000 datos provenientes de la distribución de la provincia seleccionada, verifique visualmente si los datos simulados tienen un comportamiento distribucional similar a los datos reales. Ahora, repita el proceso de simulación 10000 veces, en donde en cada iteración va a realizar la estimación del promedio de la temperatura, realice un histograma de dichos promedios, ¿qué observa? ¿Qué propiedad estadística soporta los hallazgos encontrados?

3. Verifique que el Kernel gaussiano cumple las propiedades de ser un Kernel

4. De las siguientes afirmaciones, diga si son verdaderas o falsas justificando:
    a. La suma de Kernels es un Kernel
    b. Cualquier Kernel puede ser utilizado para generar una regresión no paramétrica

5. Cree un modelo de regresión donde la variable explicativa (Y) es la provincia que usted eligió, y las variables regresoras (X) son las demás provincias del conjunto de datos. Utilice 3 aproximas robustas y/o no paramétricas para estimar el modelo de regresión, e identifique cuales provincias son relevantes para explicar la provincia elegida (presente evidencia estadística para justificar su respuesta).

6. Considere la siguiente función de densidad conjunta

$$f(x,y) = \begin{cases} \frac{3}{16}(x^2+y^2) & si\ 0 < x < y < 2 \\ 0 & En\ otro\ caso \end{cases}$$

Realice la estimación de la regresión no paramétrica E(Y|X). Adicionalmente, con el modelo de regresión teórico encontrado, realice la simulación de 1000 valores de la Y agregando ruido blanco con desviación de 0.1. Posteriormente, realice la estimación de 1 modelo de regresión robusto y/o no paramétrico, además de la regresión lineal tradicional y compare el desempeño de haber estimado un modelo teórico con una versión no paramétrica VS una paramétrica.

7. Considere el planteamiento del modelo de regresión tradicional paramétrico:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Que busca minimizar

$$Error = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

Cuya solución es

$$\hat{\beta}_1 = \frac{COV(X,Y)}{VAR(X)}\ (2)\ \text{ y }\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \qquad (3)$$

Plantee maneras de ***robustecer*** (1), (2) y (3). Realice la implementación de la versión robusta planteada utilizando como Y la provincia seleccionada, en función de cualquier provincia adicional que desee. ***Nota:*** no es necesario implementar la solución numérica de la optimización del problema, es suficiente ajustar el modelo y presentar el error obtenido utilizando la función de coste robusta.

8. Utilice el método de rangos para identificar que provincias del conjunto de datos son distintas a las elegidas

9. Para el conjunto de datos de accidentes, utilice la medida de profundidad de su elección, identifique cuales son los 10 barrios más profundos. ¿Dentro de lo que usted conoce de la planificación territorial de la ciudad de Medellín, tiene sentido que esos barrios sean los más profundos a nivel de accidentalidad?

10. Utilice la medida de profundidad que desee para identificar el 5% de curvas outliers que hay en el conjunto de datos. ¿Dentro de lo que usted conoce de la planificación territorial de la ciudad de Medellín, tiene sentido que esos barrios sean outliers a nivel de accidentalidad?

11. Consulte cómo es el funcionamiento de un boxplot en el caso funcional, además realice la implementación de este sobre el conjunto de datos de accidentalidad, y compare los resultados obtenidos VS los outliers obtenidos en el punto anterior.

Entregable: Se recibirá un trabajo por equipo (equipos creados en EAFIT Interactiva), en el cual debe de contener (1) informe/documento escrito que contenga el enunciado de cada punto, además de su respectiva solución y análisis, adicionalmente (2) el código organizado y documentado utilizado para dar solución a la implementación computacional realizada en el taller. Código que no ejecute o compile, se tomará como razón para disminuir la puntuación de los ejercicios computacionales.

Fecha de entrega: jueves 16 de mayo 2024.