Datos LIDAR

La técnica conocida como LIDAR (light detection and ranging) usa la reflexión de luz de láser emitida para detectar compuestos químicos en la atmósfera. Esta técnica ha probado ser una herramienta muy eficiente para el monitoreo de la distribución de diversos elementos polulantes en la atmósfera (Sigrist, 1994).

En el archivo lidar.txt (disponible para su descarga en el campus) se encuentran datos medidos con la técnica LIDAR. La variable "range" (`rango`) es la distancia recorrida antes de que la luz sea reflejada de regreso hacia su fuente. La variable "int.conc" (`logratio`) es el logaritmo del cociente de la luz recibida de dos fuentes de luz láser de distinta frecuencia.

1. A partir de los datos de lidar.txt realizar un diagrama de dispersión (scatter plot) de `rango` (eje x) vs. `logratio` (eje y). Describir la relación entre ambas variables.

2. Utilizar la función `pred_prom_loc`, definida en el ítem 11 de la guía de "Actividades de Clase - Unidad 2", para predecir el `logratio` de un `rango` de 602. Realizar la predicción con ventana `h = 5`. Superponer al gráfico de dispersión del ítem anterior el punto correspondiente a esta predicción en color rojo y relleno.

3. Graficar la función predictora del `logratio`, `pred_prom_loc`, para `h = 5` en base a los datos de Lidar y superponerla al gráfico de dispersión del primer ítem. Repetir para `h = 10` y `h = 30` y comparar.

4. Utilizar la función `ksmooth` de `R` para graficar la función predictora del `logratio` (en base a `rango`) mediante el método de Nadaraya Watson con núcleo normal y ancho de ventana `h = 5`. Superponer la función al diagrama de dispersión del primer ítem.

5. Repetir para `h=10,30,50` y graficar las 4 funciones (para cada `h=5,10,30,50`) en un mismo gráfico junto con los puntos observados. Comparar los resultados obtenidos con las 4 ventanas.

6. Con la función `ksmooth`, predecir el `logratio` para los valores observados del `rango` usando cada una de las 4 ventanas del ítem anterior y luego computar el Error Cuadrático Medio de entrenamiento ($\text{MSE}_{\text{train}}(h)$) para cada h. ¿Cuál de las 4 ventanas consideradas da el menor $\text{MSE}_{\text{train}}(h)$? ¿Cómo se puede justificar lo que está ocurriendo?

7. Hallar mediante el criterio de Validación Cruzada - Leave One Out (LOOCV) la ventana óptima, h_{opt}. Realizar la búsqueda en una grilla para valores de `h` entre 3 y 165 con paso 1. Graficar h vs. $CV(h)$.

8. Computar la pérdida (ó error) de Validación Cruzada asociada a la estimación provista por el método de Nadaraya-Watson con la ventana óptima hallada, $CV(h_{\text{opt}})$, y también el Error Cuadrático Medio de entrenamiento, $\text{MSE}_{\text{train}}(h_{\text{opt}})$. ¿Cuál es mayor? ¿A qué cree que se debe?

9. Graficar los puntos observados y la función de regresión estimada por el método de Nadaraya-Watson con la ventana óptima hallada.

10. Por otra parte, calcule la predicción de `logratio` por el método de vecinos más cercanos para un valor de `rango` igual a 570 utilizando $k = 5$ vecinos. Superponga al gráfico de dispersión de las observaciones, el punto correspondiente a esta predicción.

11. Implemente una función `pred_knn(x, y, x_nuevo, k)` que en base a un conjunto de valores `x`, sus correpondientes valores `y`, un nuevo valor de x, `x_nuevo`, donde queremos predecir y la cantidad `k` de vecinos que vamos a utilizar, calcule la predicción de y mediante el método de vecinos más cercanos. ¿Qué puede hacer para corroborar que está bien implementada?

12. Utilizar la función definida en el ítem anterior para obtener predicciones para `logratio` por el método de vecinos más cercanos para un valor de `rango` igual a 570 utilizando `k = 20 y 40` y comparar los resultados obtenidos aquí junto con el que utiliza `k = 5`.

13. Explorar la función `knn.reg` de la librería `FNN`. Repetir los ítems 7, 8 y 9.