Ejercicio 1

En este ejercicio trabajaremos con el conjunto de datos `airquality` de la librería base de `R`. Si bien los árboles de regresión son particularmente útiles cuando el número de covariables es grande, en este ejercicio entrenaremos un árbol con una única covarible (en base a muy pocas observaciones) para entender el método con el cual se construye. Concretamente, entrenaremos un árbol para predecir el nivel de ozono en función del viento.

1. Explorar el conjunto de datos `airquality`. Recordar que se puede utilizar el comando `data(‘‘airquality’’)` para cargarlo al environment.

2. Ajustar un árbol de regresión para predecir Ozono (`Ozone`) en función de Viento (`Wind`), utilizando las primeras 20 observaciones de `airquality` (sin datos faltantes en `Ozone` ni `Wind`) y graficarlo.

3. Con el objetivo de entender la construcción del árbol obtenido en el punto anterior, determinar empíricamente el primer split propuesto. Para ello:

   a) Considerar todos las posibles particiones binarias de los datos de `Wind` (que dan origen a dos regiones con un mínimo de 5 datos cada una) y para cada una de ellas calcular el RSS del árbol resultante. ¿Cuánto vale el mínimo RSS?
   (Hint.: reordenar el data frame utilizado como muestra de entrenamiento en base a `Wind`.)

   b) ¿Para qué partición se alcanza el mínimo RSS? ¿A qué valor de la variable `Wind` corresponde? Entonces, ¿por qué el split propuesto por el árbol de `R` es `Wind` <9.45?

   c) ¿Cuánto vale el RSS del árbol?

Ejercicio 2

Volviendo a considerar el conjunto de datos `airquality`:

1. Dividir aleatoriamente los datos en dos muestras. Definir una primera muestra de tamaño 76, como de entrenamiento y la otra, de tamaño 77, como muestra de testeo. (Utilizar semilla 2).

2. Entrenar un árbol de regresión para predecir Ozone en función del resto de las covariables en base a la muestra de entrenamiento.

   a) ¿Cuántos nodos terminales tiene el árbol?

   b) ¿Qué variables fueron utilizadas?

   c) ¿Cuál es el RSS del árbol? ¿Cuál es el MSE de entrenamiento?

d) ¿Cuál es la variable más importante para predecir el nivel de Ozono?

3. Obtener el valor predicho para la cuarta observación completa de la muestra de entrenamiento y su residuo correspondiente. Calcular el residuo "a mano" y comparar su valor con el objeto `residuals` del `summary`.

4. Predecir el nivel de Ozono en el Día número 7 del próximo mes de Junio, con una Radiación Solar de 120, Viento de 10.5 y una Temperatura de 62. Calcular dicha predicción en base al gráfico del árbol y corroborar con la función `predict`.

5. Calcular el MSE_{test}, es decir, error cuadrático medio en la muestra de testeo.

Ejercicio 3

Para ejercitar la poda de un árbol, descargar del campus el conjunto de datos `zariguella.csv`.

1. Entrenar un árbol de regresión para predecir la edad de una zarigüella (en años) en función del resto de las covariables (las cuales están compuestas por 7 medidas morfométricas de cada una de ellas), en base a una submuestra de entrenamiento (utilizando semilla 2), considerando aproximadamente la mitad de los datos. (Tener en cuenta que hay que sacar la primer columna, ya que la misma solo contiene el número de observación)

2. Graficar el error de $k-$fold CV (con $k = 10$), que se obtiene al realizar la secuencia de subárboles que resulta de minimizar la función de costo complejidad para distintos valores de α (con semilla igual a 3), versus la cantidad de nodos terminales.

3. Podar el árbol entrenado en el punto 1, utilizando el tamaño óptimo según lo analizado en el ítem anterior, graficarlo e identificar qué ramas fueron podadas.

4. Calcular el MSE_{test} del árbol podado e interpretar el valor de su raíz cuadrada.