Introducción al Aprendizaje Estadístico - Actividades de Clase - Unidad 6
Lucía Babino - Paula Spano

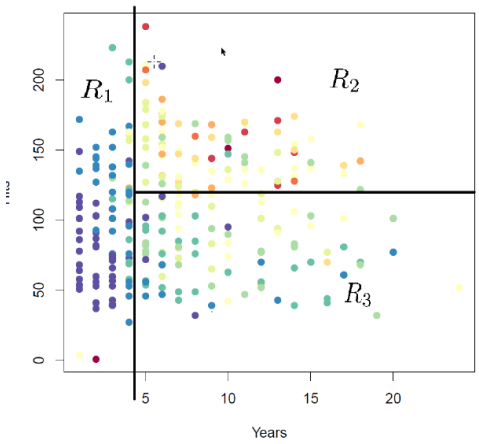
Ejercicio 1: Calentando motores

1. Elaborar un ejemplo para la aplicación del método de árboles de regresión con una variable de respuesta y dos covariables.

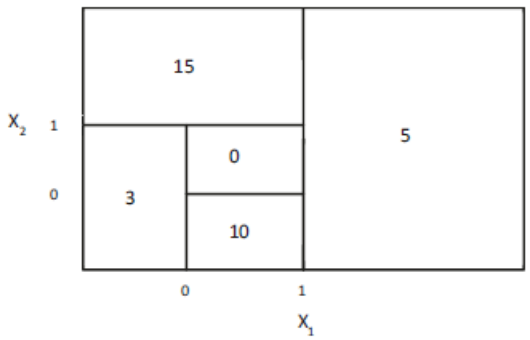
   a) Proponer una partición obtenida mediante divisiones recursivas binarias, que contenga, al menos, seis regiones.

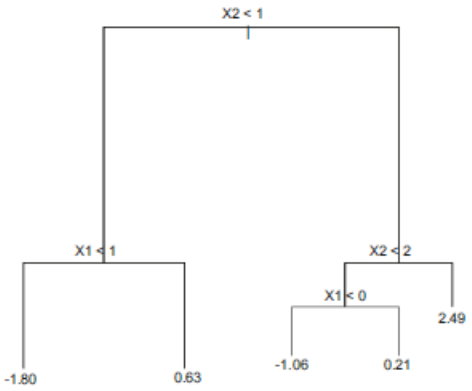
   b) Dibujar el árbol correspondiente a la partición propuesta, etiquetando todos los aspectos que se deben incluir en dicho gráfico.

   (Hint: los gráficos tienen que tener un aspecto similar a los siguientes, presentados en la teórica)



2. Mirando el siguiente gráfico:





a) Representar el árbol correspondiente a la partición del espacio de covariables propuesta en el gráfico de la izquierda.

b) Crear un diagrama similar al propuesto en el ítem anterior, que represente las particiones del árbol ilustrado en el gráfico de la derecha.

Ejercicio 2: Conjunto de datos Boston

El dataset Boston de la librería ISLR2, contiene información respecto a los precios medianos de las viviendas (`medv`) en 506 suburbios de Boston.

El objetivo es predecir dicho valor.

3. Importar en `R` la librería **ISLR2** y explorar el conjunto de datos `Boston`. Utilizar el help para ver el significado de las variables. Analizar si es necesario modificar la clase de alguna/s de ellas.

4. Crear una muestra de entrenamiento que tenga la mitad de las observaciones (extraídas aleatoriamente del dataset original utilizando semilla 1) y utilizarla para entrenar un árbol de regresión mediante el comando `tree` para predecir `medv` en función del resto de las variables. Guardar el resto de los datos en otro dataset que utilizaremos luego como muestra de testeo para evaluar la performance del árbol. Guardar el árbol en un objeto llamado `tree_boston`.

5. Realizar un `summary` de `tree_boston` y determinar

   a) ¿Qué variables fueron consideradas para la construcción del árbol?

   b) ¿Cuántas hojas lo componen?

   c) Utilizando la función `predict()`, calcular la RSS (suma de cuadrados de los residuos) del árbol en la muestra de entrenamiento. ¿Dicho valor se encuentra en el summary?

6. Graficar el árbol obtenido en el punto anterior. En base a este gráfico, responder:

   a) ¿cuál es el factor más importante para predecir el precio mediano de las viviendas en un suburbio de Boston?

   b) para cada una de las variables `rm` y `lstat` indicar si mayores valores de esa variable están asociados con mayores o menores precios.

   c) predecir el valor mediano de una vivienda que cumple con las siguientes características:

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat
1.19294	0	21.89	0	0.624	6.326	97.7	2.271	4	437	21.2	12.26

7. Utilizar el comando `predict` para obtener la misma predicción del ítem anterior.

8. Calcular el MSE_{test}, es decir el error cuadrático medio en la muestra de testeo, el MSE_{train} y comparar. ¿A qué cree que se debe lo que observa?

Ejercicio 3: Podando el árbol

9. Calcular el error de k-fold CV (con $k = 10$ folds) de la secuencia de subárboles que resulta de minimizar la función de costo complejidad para distintos valores de α. (Utilizar semilla 3)

10. Graficar el error de CV de cada subárbol en función de su tamaño, ¿cuál elige?

11. Podar el árbol maximal para obtener un subárbol de tamaño 5 según el método de cost-complexity, graficarlo e identificar qué rama/s fueron podadas del árbol maximal.

12. Calcular el MSE_{test} tanto para el árbol maximal como para el podado y comparar. ¿Es coherente lo que ocurre en este caso con lo observado al calcular los errores de CV?

13. Cacular la raíz cuadrada de MSE_{test} del árbol maximal e interpretar.