

EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2023 — Extensiones Árboles.
Objetivo: Que veremos.
Lectura recomendada: Referencia.

1. INTRODUCCIÓN

Los árboles de decisión por construcción son insensibles a:

- Variables continuas con *outliers*.
- Variables continuas con diferentes escalas.
- Variables categóricas (no se necesitan transformar a *dummies*).

El algoritmo de ajuste de un árbol de decisión, en regresión por ejemplo, busca minimizar la función de pérdida

$$R_{\alpha}(T) = \sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|, \quad (1)$$

donde $\hat{y}_{R_m} = \sum_{i: x_i \in R_m} y_i / n_m$ es el promedio de las respuestas en la región m -ésima, a través de ir buscando secuencialmente las regiones R_m que logren la máxima disminución de RSS.

La búsqueda se puede lograr por medio de un algoritmo voraz (*greedy*) que escoge variables y puntos de corte.

1.0.1. Para pensar: En nuestra última discusión sobre árboles de decisión es que tienen un sesgo al utilizar variables categóricas con muchas etiquetas. ¿Por qué?

1.1. Motivación

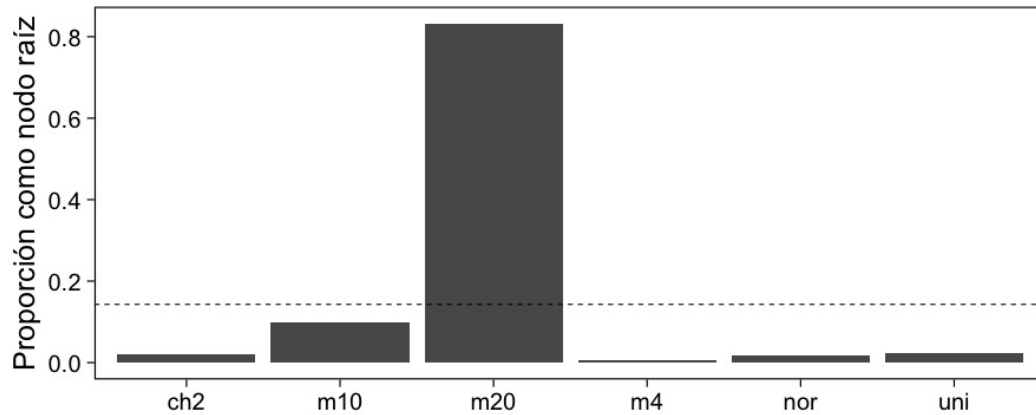
Consideremos el experimento siguiente. Tenemos una colección de 7 atributos y una respuesta. Los datos son generados de manera aleatoria.

```
1 generate_data(nsamples = 100) ▷  
2 print(n = 5)
```

De manera artificial consideraremos que la respuesta y está relacionada con los demás atributos (sabemos que no es cierto).

```
1 fit_tree ← function(engine){  
2   data_train ← generate_data()  
3  
4   tree_spec ← decision_tree(tree_depth = 2) ▷  
5   set_engine(engine) ▷  
6   set_mode("regression")  
7  
8   tree_spec ▷  
9   fit(y ~ ., data = data_train)  
10 }
```

Repetimos este proceso (generar datos aleatorios y ajustar un árbol) un número determinado de veces y registramos cuántas veces cada atributo fue utilizado en el nodo raíz.



En contraste, con un modelo denominado `conditional trees` somos capaces de evitar ese sesgo. Incluso podemos evitar escoger variables de corte que no tienen asociación con la respuesta.

