Práctica 6: Método basado en Árboles de Decisión

Evelyn G. Coronel
Redes Neuronales y Aprendizaje Profundo para Visión Artificial
Instituto Balseiro

(5 de noviembre de 2020)

EJERCICIO 1

Item A

Este ejercicio, se utilizan arboles de clasificación y regresión para obtener la cantidad de unidades de asientos para bebés se compran, en función de las características del producto y del lugar donde se venden. Las variables a tener en cuenta del conjunto de datos son:

1. Sales: Cantidad de asientos vendidos en una tienda.
2. CompPrice: Precio del asiento según la competencia.
3. Income: Ingreso de dinero medio en miles de USD del lugar donde se localiza la tienda.
4. Advertising: Inversión en propaganda de la tienda en miles de USD.
5. Population: Población local en miles.
6. Price: Precio de venta por cada asiento.at each site
7. ShelveLoc: Posición del asiento en las góndolas, se clasifica las posición como Bad, Good and Medium (Malo, Bueno y Medio).
8. Age: Edad promedio de la población local.
9. Education: Nivel de educación
10. Urban: *Yes* si la tienda está ubicada en un lugar urbano, *No* caso contrario.
11. US: *Yes* si la tienda está ubicada en Estados Unidos, *No* caso contrario.

Además de estas variables, se agregó una última variable llamada *High* es *Yes* si las ventas superan las 8 unidades, *No* caso contrario.

Para realizar los ajustes, se transformaron los datos descriptos *Yes* y *No* como 1 y 0 respectivamente, y *Good*, *Medium*, *Bad* con 3, 2 y 1. Además se separó el conjunto de datos en un 80 % de entrenamiento y un 20 % de validación. Las variables *Sales* y *High* se separan del resto de los atributos para hacer los ajustes.

Item B

Usando un árbol de decisión de clasificación, se utilizó la columna *High* como salida esperada y el resto de los atributos como entrada. El árbol está instanciado de tal forma que el ajuste se realice hasta que las hojas sean puras, es decir, que el coeficiente de gini del nodo sea 0. Con esto se obtuvo un árbol con las características que se presentan en la Tabla I.

Profundidad	11
Hojas	50
Error de entrenamiento	0 de 320
Precisión de entrenamiento	1
Error de validación	3 de 80
Precisión de validación	0.6625

Tabla I: Características del árbol de decisión de clasificación posterior al ajuste

Las primeras dos ramas del árbol se muestran en la Fig.1. En este caso se llega hasta un profundidad de 11 porque no estamos limitando el valor de gini de las hojas, la clasificación se realiza hasta que los datos de entrenamiento tengan una clasificación fiel, por eso se observa el overfitting del árbol. El primer nodo usa la posición en la góndola para clasificar los datos, ya que este parámetro divide mejor los datos, dicho de otra forma, tiene el coeficiente de gini mayor de todos los nodos.

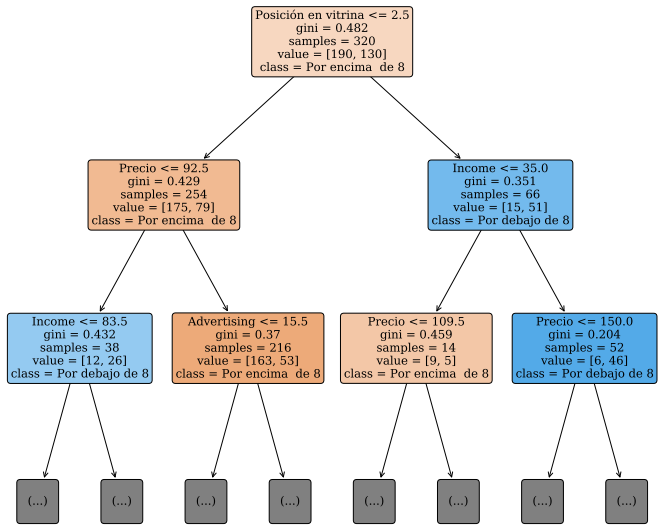

Fig. 1: Las primeras dos ramas de la estructura del árbol de clasificación para el item B.

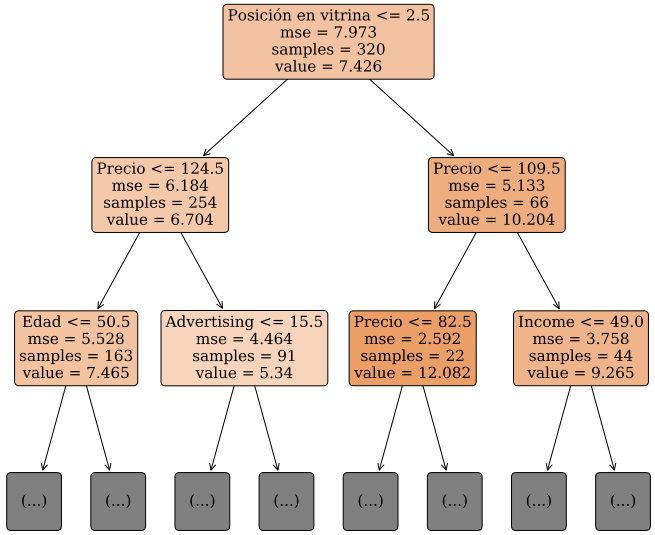

Fig. 2: Las primeras dos ramas de la estructura del árbol de clasificación para el item C.

Item C

El árbol instanciado en este item es el árbol de regresión, los parámetros por defecto también hacen que el árbol regresione hasta tener hojas con gini nulo. Como se ven en la Tabla II se muestran las características del árbol luego de realizar el ajuste. Nuevamente se tiene un overfitting con el conjunto de entrenamiento.

Tal como el caso del árbol de clasificación, el atributo que mejor separa los datos es la posición de los asientos en las góndolas. Esto tiene sentido porque eso es independiente del tipo de árbol que se utilize.

Profundidad	18
Hojas	320
Error de entrenamiento	0.0
Precisión de entrenamiento	1
Error de validación	0.07
Precisión de validación	0.2759

Tabla II: Características del árbol de regresión posterior al ajuste

A pesar del overfitting del árbol, el árbol aprende la correlación de los atributos con la cantidad de asientos. Esto se observa en la Fig.3, dado que los ejes x e y son la cantidad de asientos reales y predichos, mientras más cerca estén los puntos a la línea, mejor es la predicción del árbol.

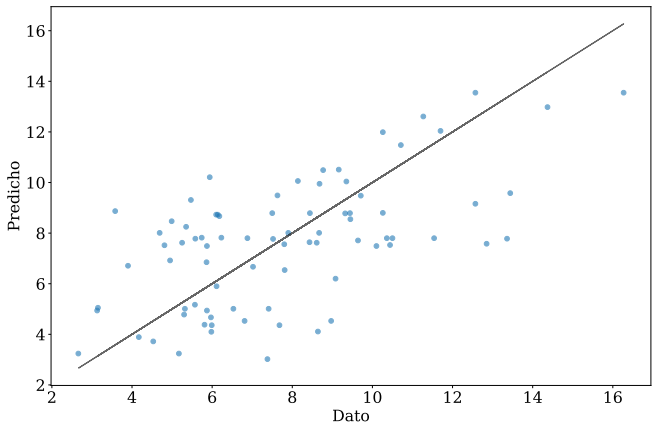

Fig. 3: Comparación entre el dato y la predicción del árbol de regresión para else item C.

Item D

Item E

Item F

Item G

Item H