ARBOLES - METODOS DE AGREGACIÓN

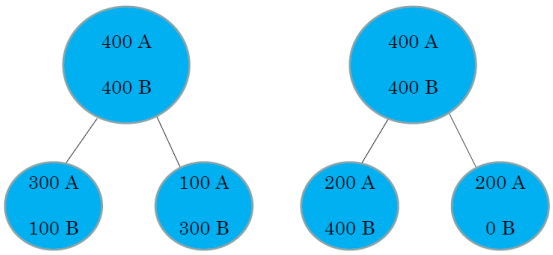
1. Demostrar que si X es categórica con m clases entonces hay $2^{m-1}-1$ posibles divisiones.
2. Demostrar que las tres expresiones del índice de Gini son iguales.
3. Considerar el siguiente conjunto de datos:

x_1	a	a	b	a	a	b	b	b
x_2	b	a	a	a	a	b	b	b
y	1	1	1	1	-1	-1	-1	-1

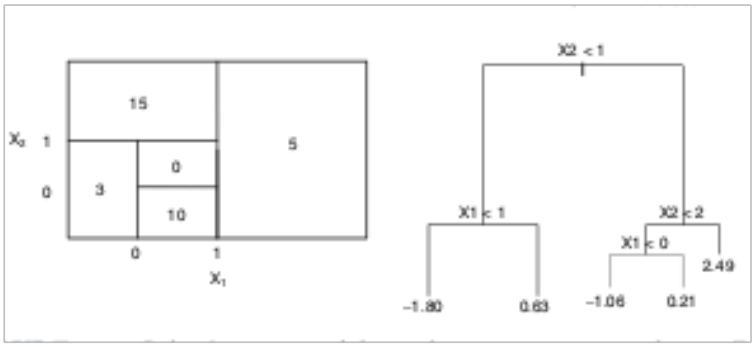
   a) Construir un árbol de clasificación, con el índice de Gini, a partir de esta muestra

   b) Comparar con el árbol obtenido en R. ¿Qué argumento de la función rpart debería cambiar para obtener el mismo árbol?

4. Calcular $\Delta i(t,s)$ para estas dos particiones utilizando el error de clasificación, el índice de Gini y la entropía.



5. Consideramos la siguiente figura:



   a) Dibujar el árbol correspondiente a la partición del espacio ilustrada en la parte izquierda de la figura.

   b) Crear un diagrama similar al del lado izquierdo de la figura, utilizando el árbol ilustrado en el panel derecho de la misma figura.

6. Se desea construir un árbol de decisión para clasificar dos variedades de tomates: Cherry y Perita. Se usarán dos atributos: Dulzor y Tamaño. Se dispone de un conjunto de entrenamiento con 90 tomates: 45 Cherry y 45 Perita. Los datos se resumen en las siguientes tablas:

   a) Dibujar los dos árboles de decisión posibles, resultantes de dividir el nodo raíz por uno de los dos atributos. Cada árbol debe incluir la siguiente información:

      1) Número de observaciones en el nodo raíz y hojas.

Dulzor	No tomates	Cherry	Perita
Alto	55	30	25
Bajo	35	15	20

Tamaño	No tomates	Cherry	Perita
Pequeño	60	40	20
Grande	30	5	25

2) La distribución de las etiquetas (y) en el nodo raíz y hojas.

3) Impureza de Gini $H(y;S) = 1 - \sum_c p_c^2$ en el nodo raíz y hojas.

4) En el nodo raíz: la pregunta realizada.

5) En las hojas: la predicción.

b) Calcule la impureza de Gini esperada

$$\frac{N(t_L)}{N(t)} i_{t_L} + \frac{N(t_R)}{N(t)} i_{t_R}$$

asociada a cada una de las dos divisiones de la parte anterior.

c) ¿Qué pregunta elegiría en el nodo raíz? Justificar en base a las partes anteriores.

d) Para el árbol elegido en la parte anterior, calcular el error

7. Consideramos 3 muestras bootstrap de un mismo conjunto de datos ue contiene clases blancas y negras. A continuación, aplicamos un árbol de clasificación a cada muestra bootstrap y, para un nuevo valor $\mathbf{x_0}$ obtenemos 3 estimaciones de $\mathbb{P}(\text{negro}|\mathbf{x_0})$: 0,1, 0,1 y 0,9.¿Cuál es la predicción final según el criterio del voto mayoritario? ¿Cuál es la predicción final si promediamos las probabilidades?

8. Consdideramos el conjunto de datos Carseats de la biblioteca rpart.

   a) Dividir el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba (2/3 -1/3) con set.seed(2024)).

   b) Ajustar un árbol de regresión al conjunto de entrenamiento. Dibujar el árbol e interpretar los resultados. ¿Qué tasas de error sobre la muestra de entrenamiento y sobre la muestra test se obtienen?

   c) Utilizar la validación cruzada para determinar el nivel óptimo de complejidad del árbol. ¿La poda del árbol mejora la tasa de error de la prueba?

   d) Utilizar el método Bagging para analizar estos datos. ¿Qué tasa de error sobre la muestra test se obtiene?

   e) Utilice Random Forest para analizar estos datos. ¿Qué tasa de error sobre la muestra test se obtiene? Utilice la función importance() para determinar qué variables son las más importantes. Describa el efecto de m, el número de variables consideradas en cada división, sobre la tasa de error obtenida.

   f) Contestar las mismas preguntas si la variable Sales se discretiza de la siguiente manera: 1 si la variable Sales es superior a 8, 0 en caso contrario.

9. En el método Adaboost para dos clases probar que

   a) si $\varepsilon < 1/2$ entonces $\alpha > 0$ y la actualización de los pesos es correcta

   b) $\text{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) = \underset{y\in\{-1,1\}}{\text{Argmax}} \left(\sum_{t=1}^{T} \alpha_t \mathbb{1}_{\{h_t(x)=y\}}\right)$

   c) justificar los cálculos del ejemplo de juguete de Freund y Schapire en las transparencias