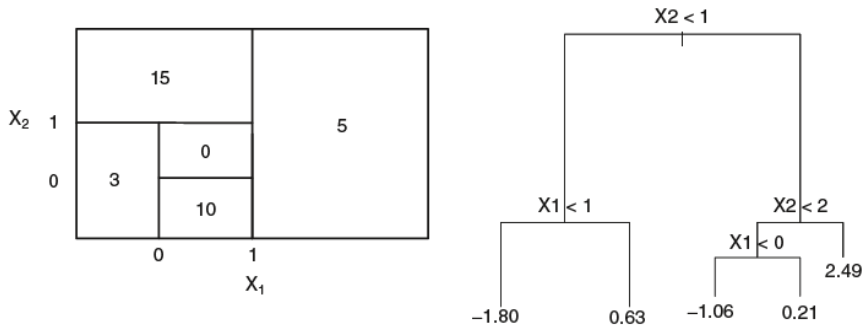
Data Mining HW 3

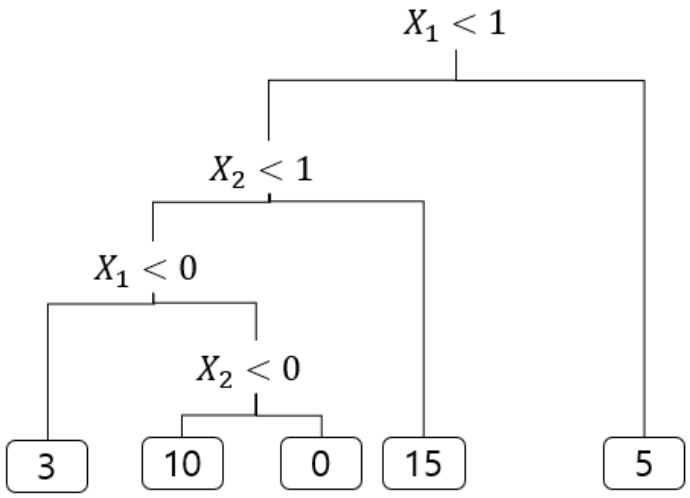
Due: 2022.05.23 24:00

Exercises for Tree

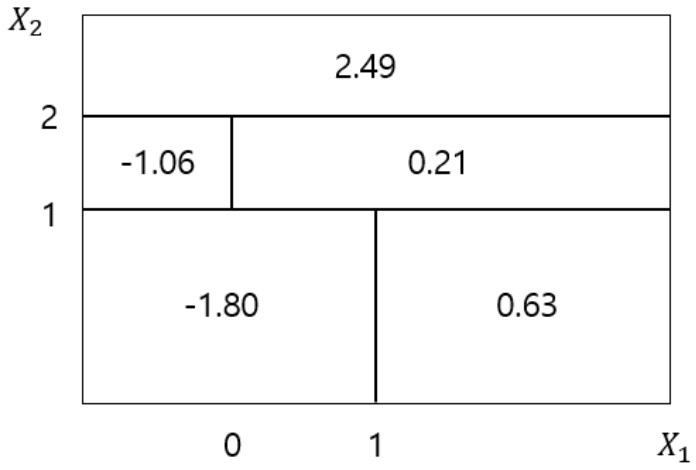


1. 위의 그림을 보고, 아래의 물음에 답하여라.

   (a) 왼쪽 그림을 보고, 오른쪽과 같은 tree를 그려라. 왼쪽 그림의 박스안의 숫자는 각 영역 내의 Y의 평균이다.



(b) 오른쪽 그림을 보고 왼쪽과 유사한 다이어그램을 생성하여라. 예측 변수 공간을 올바른 영역으로 나누고, 각 영역의 평균을 표시해야 함.

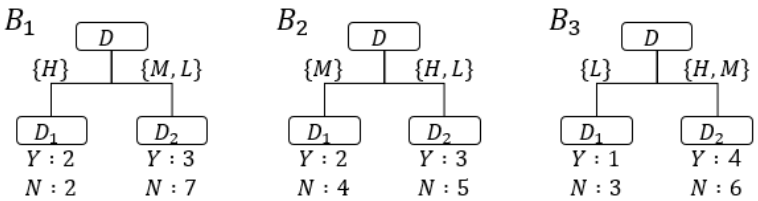


2. Tree 강의노트 p.22의 데이터를 이용하여, 불순도 측도로 Gini 지수를 사용했을 때, 첫번째 분리 규칙을 찾아라.

	age	income	student	credit_rating	buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	exellent	no
3	31…40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	exellent	no
7	31…40	low	yes	exellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	exellent	yes
12	31…40	medium	no	exellent	yes
13	31…40	high	yes	fair	yes
14	>40	medium	no	exellent	no

- Age $= \{\leq 30, 30 \ldots 40, > 40\}$
  - $\Delta Gini(A_1) = 0.459 - 0.394 = 0.065$
  - $\Delta Gini(A_2) = 0.459 - 0.357 = 0.102$
  - $\Delta Gini(A_3) = 0.459 - 0.457 = 0.002$

  $\Delta Gini(A_2)$ 값이 가장 크기 때문에 Age 변수에 대한 분리 규칙은 $A_2 = \{30 \ldots 40\}, \{\leq 30, > 40\}$ 로 선택
- Income $= \{H, M, L\}$



  - $B_1 = \{H\}, \{M, L\}$

$$Gini(D_1) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$Gini(D_2) = 1 - \left(\frac{3}{10}\right)^2 - \left(\frac{7}{10}\right)^2 = 0.42$$

$$Gini_{B_1}(D) = \frac{4}{14} \times 0.5 + \frac{10}{14} \times 0.42 = 0.443$$

$$\Rightarrow \Delta Gini(B_1) = 0.459 - 0.443 = 0.016$$

- $B_2 = \{M\}, \{H, L\}$

$$Gini(D_1) = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = 0.444$$

$$Gini(D_2) = 1 - \left(\frac{3}{8}\right)^2 - \left(\frac{5}{8}\right)^2 = 0.469$$

$$Gini_{B_2}(D) = \frac{6}{14} \times 0.444 + \frac{8}{14} \times 0.469 = 0.458$$

$$\Rightarrow \Delta Gini(B_2) = 0.459 - 0.458 = 0.001$$

- $B_3 = \{L\}, \{H, M\}$

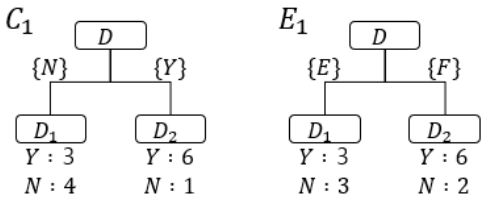
$$Gini(D_1) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$$

$$Gini(D_2) = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = 0.480$$

$$Gini_{B_3}(D) = \frac{4}{14} \times 0.375 + \frac{10}{14} \times 0.480 = 0.450$$

$$\Rightarrow \Delta Gini(B_3) = 0.459 - 0.450 = 0.009$$

$\Delta Gini(B_1)$ 값이 가장 크기 때문에 Age 변수에 대한 분리 규칙은 $B_1 = \{H\}, \{M, L\}$ 로 선택



- Student $= \{N, Y\}$
  - $C_1 = \{Y\}, \{N\}$

$$Gini(D_1) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.490$$

$$Gini(D_2) = 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0.245$$

$$Gini_{C_1}(D) = \frac{7}{14} \times 0.490 + \frac{7}{14} \times 0.245 = 0.367$$

$$\Rightarrow \Delta Gini(C_1) = 0.459 - 0.367 = 0.092$$

- Credit rating $= \{F, E\}$
  - $E_1 = \{E\}, \{F\}$

$$Gini(D_1) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$$

$$Gini(D_2) = 1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = 0.375$$

$$Gini_{E_1}(D) = \frac{6}{14} \times 0.5 + \frac{8}{14} \times 0.375 = 0.428$$

$$\Rightarrow \Delta Gini(E_1) = 0.459 - 0.428 = 0.031$$

$\Rightarrow$ 각 변수의 분리규칙 중 Age의 지니불순도 감소량이 가장 크기 때문에 첫번 때 분리규칙은 A_2이다.