Homework 3

Id	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	Over 100K	Yes
2	No	Married	Under 100K	No
3	Yes	Single	Over 100K	Yes
4	Yes	Married	Over 100K	Yes
5	No	Single	Over 100K	No
6	Yes	Single	Under 100K	No
7	No	Married	Over 100K	No
8	No	Single	Over 100K	No
9	Yes	Single	Under 100K	No
10	Yes	Married	Over 100K	Yes

지니계수(Gini index)를 사용하여 최대 깊이 분류 트리(maximum depth classification tree)를 찾으세요.

Sol) Gini index를 사용하여 최대 깊이 분류 트리를 만들려면 CART 알고리즘을 이용하면 된다.

$Gini_{refund} = \frac{6}{10}\left(1-\left(\frac{4}{6}\right)^2-\left(\frac{2}{6}\right)^2\right) + \frac{4}{10}\left(1-\left(\frac{0}{4}\right)^2-\left(\frac{4}{4}\right)^2\right) \approx 0.267$

($P(Refund=Yes)$, $P(Refund=No)$, $P(Cheat=Yes \mid Refund=Yes)$, $P(Cheat=No \mid Refund=Yes)$, $P(Cheat=Yes \mid Refund=No)$, $P(Cheat=No \mid Refund=No)$)

← $Gini_{refund=Yes} = Gini_{refund=No} = Gini_{refund}$ 이므로 따로 계산할 필요는 없다. (∵ Refund가 Binary data 이기 때문)

$Gini_{marital\ status} = \frac{6}{10}\left(1-\left(\frac{2}{6}\right)^2-\left(\frac{4}{6}\right)^2\right) + \frac{4}{10}\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right) \approx 0.467$

$Gini_{taxable\ income} = \frac{7}{10}\left(1-\left(\frac{6}{7}\right)^2-\left(\frac{1}{7}\right)^2\right) + \frac{3}{10}\left(1-\left(\frac{0}{3}\right)^2-\left(\frac{3}{3}\right)^2\right) \approx 0.343$

$Gini_{refund}$가 가장 작으므로 Refund를 기준으로 분류한다.

Refund=No인 경우 모두 Cheat=No 이므로 분류를 마친다.

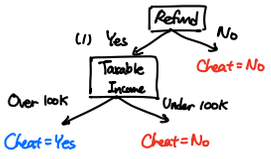
(1) Refund=Yes인 경우 위와 같은 방법으로 계속 분류한다.

$Gini^{(1)}_{marital} = \frac{4}{6}\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right) + \frac{2}{6}\left(1-\left(\frac{2}{2}\right)^2-\left(\frac{0}{2}\right)^2\right) \approx 0.333$

($P(Cheat=Yes \mid Material\ Status=Single,\ Refund=Yes)$, $P(Cheat=No \mid Material\ Status=Single,\ Refund=No)$)

$Gini^{(1)}_{taxable\ income} = \frac{4}{6}\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right) + \frac{2}{6}\left(1-\left(\frac{0}{2}\right)^2-\left(\frac{2}{2}\right)^2\right) = 0$

$Gini^{(1)}_{taxable\ income}$이 가장 작으므로 Taxable Income을 기준으로 분류한다.



Taxable Income = Over 100K 인 경우 모두 Cheat = Yes 이므로 분류를 마친다.

Taxable Income = Under 100K인 경우 모두 Cheat = No 이므로 분류를 마친다.

최종적인 최대 깊이 분류 의사 트리다 놓다. □