COMP8118-A4

Monday, September 26, 2022 11:00 PM

Q1

a) We calculate the gains until reaching the conditions to stop the calculation. The highest gives us the top nod in the tree.

C4.5: $Gain(A,T) = (Info(T) - Info(A,T)) / SplitInfo(A)$

$SplitInfo(A) = -\sum_{v \in A} p(v) \log p(v)$

Step 1

$Info(T) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

$Info(T_{CS}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

$SplitInfo(CS,T) = \frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$ $\quad Gain(CS,T) = \frac{1-1}{1} = 0$

$Info(CS,T) = \frac{1}{2}\times 1 + \frac{1}{2}\times 1 = 1$

$Info(T_{old}) = Info(T_{middle}) = Info(T_{young}) = 1$

$Info(Age,T) = \frac{2}{8}\times 1 + \frac{2}{8}\times 1 + \frac{4}{8}\times 1 = 1$ $\quad Gain(Age,T) = 0$

$SplitInfo(Age,T) = -\frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{2}\log\frac{1}{2} = \frac{3}{2} = 1.5$

$Info(T_{fair}) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.812$

$Info(T_{high}) = -\frac{1}{4}\log\frac{1}{4} - 0\log 0 = 0$

$Info(T_{low}) = -\frac{3}{4}\log\frac{3}{4} - 0\log 0 = 0$

$SplitInfo(Income,T) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{8}\log\frac{1}{8} - \frac{3}{8}\log\frac{3}{8} = 1.4$

$Info(Income,T) = \frac{1}{2} Info(T_{fair}) = 0.4615$

$Gain(Income,T) = \frac{1 - 0.46}{1.4} = 0.39$

Step 2

No	CS-major	Age	Buy Bitcoin
1	Yes	old	Yes
2	Yes	middle	Yes
3	no	young	Yes
7	no	young	no

$Info(T) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.812$

$Info(T_{old}) = Info(T_{middle}) = -\frac{1}{4}\log\frac{1}{4} - 0\log 0 = 0$

$Info(T_{young}) = -\frac{1}{3}\log\frac{1}{3} - 1\log 1 = 0.53$

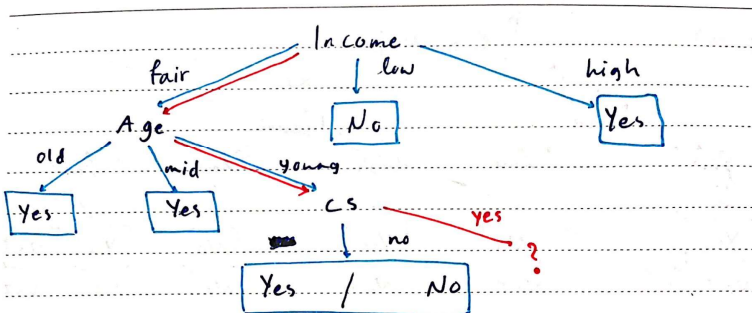
$SplitInfo(Age,T) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} = 1.5$

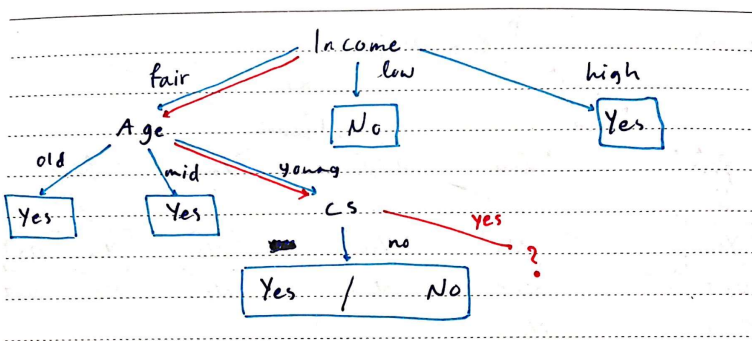
$Info(Age,T) = \frac{1}{2}\times 0.53 + 0 = 0.265$

$Gain(Age,T) = \frac{0.812 - 0.265}{1.5} = 0.364$

$Info(T_{CS\,yes}) = -1\log 1 - 0\log 0 = 0$

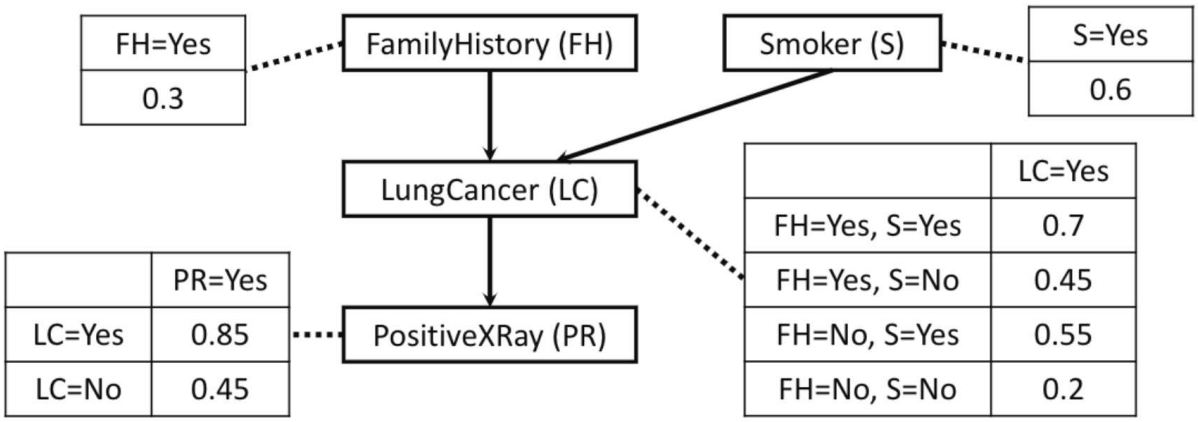
$Info(T_{CS\,No}) = -\frac{1}{1}\log\frac{1}{1} - \frac{1}{1}\log\frac{1}{1} = 1$





b) Looking at the tree, the arrows colored in red show the information about a young CS computer student with fair income. There is no previous example like this in the table, resulting in difficulty in predicting whether the answer is yes or no. But, according to the gain values and the probability of the young people who have a fair income, we can assume the prediction to be "Yes"

Q2



a)

P(LC = Yes) = P(LC = Yes | FH = Yes, S = Yes)*P(FH = Yes, S = Yes)
+ P(LC = Yes | FH = Yes, S = No)*P(FH = Yes, S = No)
+ P(LC = Yes | FH = No, S = Yes)*P(FH = No, S = Yes)
+ P(LC = Yes | FH = No, S = No)*P(FH = No, S = No)
= 0.3 * 0.6 * 0.7 + 0.3 * 0.4 * 0.45 + 0.7 * 0.6 * 0.55 + 0.7 * 0.4 * 0.2 = 0.467

b)

P(PR = Yes | FH = Yes, S = Yes) = P(PR = Yes | LC = Yes) P(LC = Yes | FH = Yes, S = Yes)
+ P(PR = Yes | LC = No) P(LC = No| FH = Yes, S = Yes)
= 0.85 * 0.7 + 0.45 * 0.3 = 0.73

c)

P(LC = Yes | PR = Yes, FH = Yes, S = Yes) =

$$= \frac{P(PR = Yes \mid LC = Yes, FH = Yes, S = Yes)\, P(LC = Yes \mid FH = Yes, S = Yes)}{P(PR = Yes \mid FH = Yes, S = Yes)}$$

$$= \frac{P(PR = Yes \mid LC = Yes) P(LC = Yes \mid FH = Yes, S = Yes)}{P(PR = Yes \mid LC = Yes) P(LC = Yes \mid FH = Yes, S = Yes) + P(PR = Yes \mid LC = No) P(LC = No \mid FH = Yes, S = Yes)}$$

= 0.85 * 0.7 / (0.85 * *0.7 + (0.45* * 0.3)) = 0.8151