Homework 6

Name: Yuan Li
ID: N19728558 netID yl6606
Email: foxerlee1@gmail.com

Part I: Written Exercises

1.

Answer:

For entropy criterion, the gain should be:

$$\begin{aligned} g(entropy, x_1) &= H(S) - \frac{20}{30}H(S_l) - \frac{10}{30}H(S_r) \\ &= -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}) - \frac{1}{3}(-log_2 1 - 0log_2 0) \\ &= log_2 3 - \frac{4}{3} \\ g(entropy, x_2) &= 0 \end{aligned}$$

So we would choose feature x_1 for the root.
For Gini criterion, the gain should be:

$$\begin{aligned} g(gini, x_1) &= G(S) - \frac{20}{30}G(S_l) - \frac{10}{30}G(S_r) \\ &= \frac{2}{3}(1-\frac{2}{3}) + \frac{1}{3}(1-\frac{1}{3}) - \frac{2}{3}(\frac{1}{2}(1-\frac{1}{2}) + \frac{1}{2}(1-\frac{1}{2})) - \frac{1}{3}(1*(1-1) + 0*(1-0)) \\ &= \frac{1}{9} \\ g(gini, x_2) &= \frac{2}{3}(1-\frac{2}{3}) + \frac{1}{3}(1-\frac{1}{3}) - (1-\frac{5}{9}) = 0 \end{aligned}$$

So we would choose feature x_1 for the root.
For misclassification criterion, the gain should be:

$$\begin{aligned} g(misclassification, x_1) &= M(S) - \frac{20}{30}M(S_l) - \frac{10}{30}M(S_r) \\ &= (1-\frac{2}{3}) - \frac{2}{3}*(1-\frac{1}{2}) - \frac{1}{3}*(1-1) \\ &= 0 \\ g(misclassification, x_2) &= 0 \end{aligned}$$

So we would choose x_1 randomly.
Consider the above result, we would choose x_1 for the root.

2. (a)

Answer:

A dataset with 4 positive examples and 5 negative examples would have a higher entropy.

2. (b)

Answer:

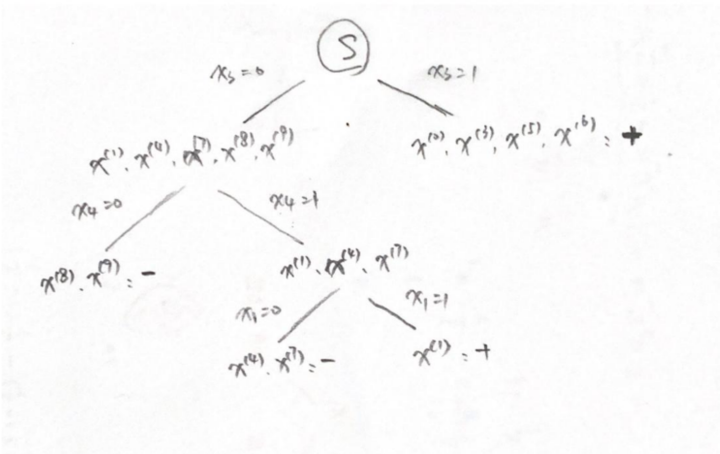
For first split, x_1 to x_4 gain information will be:

$$
\begin{aligned}
H(S) &= -\frac{4}{9}log_2\frac{4}{9} - \frac{5}{9}log_2\frac{5}{9} \\
g(x_1) &= H(S) - \frac{5}{9}H(S_l, x_1) - \frac{4}{9}H(S_r, x_1) \\
&= \frac{7}{3}log_2 3 - \frac{10}{9}log_2 5 - \frac{10}{9} \\
g(x_2) &= H(S) - \frac{6}{9}H(S_l, x_2) - \frac{3}{9}H(S_r, x_2) \\
g(x_3) &= H(S) - \frac{4}{9}H(S_l, x_3) - \frac{5}{9}H(S_r, x_3) \\
g(x_4) &= H(S) - \frac{5}{9}H(S_l, x_4) - \frac{4}{9}H(S_r, x_4)
\end{aligned}
$$

Because $g(x_3)$ is largest, I choose x_3 to split for the root.

The same as the root, I calculate x_1, x_2, x_4 gain information for the second node. As $g(x_2)$ and $g(x_4)$ are the same result, I randomly choose x_4.

For the last node, $g(x_1)$ is largest, I choose x_1. The whole tree should be as follow:



2. (c)

Answer:

$$
\begin{aligned}
H(Y) &= -\frac{4}{9}log_2\frac{4}{9} - \frac{5}{9}log_2\frac{5}{9} \\
H(Y|X) &= \frac{5}{9}(-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5}) + \frac{4}{9}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) \\
&= -\frac{1}{3}log_2\frac{3}{5} - \frac{2}{9}log_2\frac{2}{5} - \frac{2}{9}*(-1) - \frac{2}{9}*(-1) \\
&= -\frac{1}{3}log_2\frac{3}{5} - \frac{2}{9}log_2\frac{2}{5} + \frac{4}{9}
\end{aligned}
$$

So $H(Y) - H(Y|X)$ should be:

$$
\begin{aligned}
H(Y) - H(Y|X) &= -\frac{4}{9}log_2\frac{4}{9} - \frac{5}{9}log_2\frac{5}{9} + \frac{1}{3}log_2\frac{3}{5} + \frac{2}{9}log_2\frac{2}{5} - \frac{4}{9} \\
&= \frac{7}{3}log_2 3 - \frac{10}{9}log_2 5 - \frac{10}{9}
\end{aligned}
$$

2. (d)

Answer:

Set dataset S has n data. So the result should be:

$$\begin{aligned} H(S) &= -(\frac{n}{zn} log_2 \frac{n}{zn}) * z \\ &= -log_2 \frac{1}{z} \end{aligned}$$