Homework 2

September 27, 2021

Jose Carlos Munoz

3.2c)

$$Gini_{Male} = 1 - (\frac{4}{10})^2 - (\frac{6}{10})^2 = 0.48 \quad (1)$$

$$Gini_{Female} = 1 - (\frac{4}{10})^2 - (\frac{6}{10})^2 = 0.48 \quad (2)$$

$$Gini_{Gender} = \frac{10}{20} * Gini_{Male} + \frac{10}{20} * Gini_{Female} = 0.48 \quad (3)$$

The Gini for Male is as showin in(1)
The Gini for Female is as showin in(2)
The Gini for gender is as showin in(3)

3.2d)

$$Gini_{Family} = 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 0.375 \quad (1)$$

$$Gini_{Sports} = 1 - (\frac{8}{8})^2 - (\frac{0}{8})^2 = 0.00 \quad (2)$$

$$Gini_{Luxury} = 1 - (\frac{1}{8})^2 - (\frac{7}{8})^2 = 0.21875 \quad (3)$$

$$Gini_{Cars} = \frac{4}{20} * Gini_{Family} + \frac{8}{20} * Gini_{Sports} + \frac{8}{20} * Gini_{Luxury} = 0.1625 \quad (4)$$

The Gini for Family is as showin in(2)
The Gini for Sports is as showin in(3)
The Gini for Luxury is as showin in(4)
The Gini for Cars is as showin in(**??**)

3.2e)

$$Gini_{Small} = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = .48 \quad (1)$$

$$Gini_{Medium} = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = \frac{24}{49} \quad (2)$$

$$Gini_{Large} = 1 - (\frac{3}{4})^2 - (\frac{1}{4})^2 = 0.5 \quad (3)$$

$$Gini_{Extra_Large} = 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5 \quad (4)$$

$$Gini_{Shirt_size} = \frac{5}{20} * Gini_{Small} + \frac{7}{20} * Gini_{Medium} + \frac{4}{20} * Gini_{Large} + \frac{4}{20} * Gini_{Extra_Large} = 0.4914 \quad (5)$$

The Gini for Small is as showin in(1)
The Gini for Medium is as showin in(2)
The Gini for Large is as showin in(3)
The Gini for Extra Large is as showin in(4)
The Gini for Shirt Size is as showin in(5)

3.2f)
The Car type because it has the lowest Gini Index.

3.5a)

$$E_{orig} = -\frac{4}{10} * log(\frac{4}{10}) - \frac{6}{10} * log(\frac{6}{10}) = .9710 \quad (1)$$

The overall Entropy before the split is shown in (1)

$$\begin{aligned} E_T &= -\frac{4}{7} * log(\frac{4}{7}) - \frac{3}{7} * log(\frac{4}{7}) \\ E_F &= -\frac{3}{3} * log(\frac{3}{3}) - \frac{0}{3} * log(\frac{0}{0}) \\ \Delta E &= Eorig - \frac{7}{10} * E_T - \frac{3}{10} * E_F \\ &= 0.2813 \end{aligned} \quad (2)$$

The data gain from the splitting for A is show in (2)

$$\begin{aligned} E_T &= -\frac{3}{4} * log(\frac{3}{4}) - \frac{1}{4} * log(\frac{1}{4}) \\ E_F &= -\frac{1}{6} * log(\frac{1}{6}) - \frac{5}{6} * log(\frac{5}{6}) \\ \Delta E &= Eorig - \frac{4}{10} * E_T - \frac{6}{10} * E_F \\ &= 0.2565 \end{aligned} \quad (3)$$

The data gain from the splitting for B is show in (3)

3.7a)
To find best greedy split we find which of the options gives us the least errors

$$\begin{bmatrix} X & C1 & C2 \\ 0 & 60 & 60 \\ 1 & 40 & 40 \end{bmatrix} \quad (1)$$

As seen from (1),when X is 0, we see that there is a min of 60 errors and when X is 1, there is a min of 40 errors. So for X, it has an error rate of $frac60 + 40200;.5$

$$\begin{bmatrix} Y & C1 & C2 \\ 0 & 40 & 60 \\ 1 & 60 & 40 \end{bmatrix} \quad (2)$$

As seen from (2), when Y is 0, we see that there is a min of 40 errors and when X is 1, there is a min of 40 errors. So for X, it has an error rate of $frac40 + 40200;.4$

$$\begin{bmatrix} Z & C1 & C2 \\ 0 & 30 & 70 \\ 1 & 70 & 30 \end{bmatrix} \quad (3)$$

As seen from (3), when Z is 0, we see that there is a min of 30 errors and when X is 1, there is a min of 30 errors. So for X, it has an error rate of $frac30 + 30200;.3$
We split first at Z because it has the lowest error rate

Z=0

$$\begin{bmatrix} X & C1 & C2 \\ 0 & 15 & 45 \\ 1 & 15 & 25 \end{bmatrix} \quad (1)$$

As seen from (1),when X is 0, we see that there is a min of 15 errors and when X is 1, there is a min of 15 errors. So for X, it has an error rate of $frac15 + 15100;.3$

$$\begin{bmatrix} Y & C1 & C2 \\ 0 & 15 & 45 \\ 1 & 15 & 25 \end{bmatrix} \quad (2)$$

As seen from (1),when Y is 0, we see that there is a min of 15 errors and when X is 1, there is a min of 15 errors. So for X, it has an error rate of $frac15 + 15100;.3$
Since both are about the same, the node split can be choosen arbirtarly

Z=1

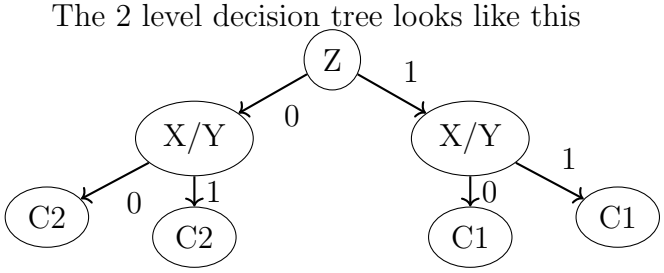
$$\begin{bmatrix} X & C1 & C2 \\ 0 & 25 & 15 \\ 1 & 25 & 15 \end{bmatrix} \quad (1)$$

As seen from (1),when X is 0, we see that there is a min of 15 errors and when X is 1, there is a min of 15 errors. So for X, it has an error rate of $frac15 + 15100;.3$

$$\begin{bmatrix} Y & C1 & C2 \\ 0 & 45 & 15 \\ 1 & 45 & 15 \end{bmatrix} \quad (2)$$

As seen from (2),when Y is 0, we see that there is a min of 15 errors and when X is 1, there is a min of 15 errors. So for X, it has an error rate of $frac15 + 15100$;.3
Since both are about the same, the node split can be choosen arbirtarly

The 2 level decision tree looks like this



This Decision tree has an error of $frac15 + 15 + 15 + 15200$, 0.3
3.7b)
If we start with X instead of Z then this is how it would played out
X=0

$$\begin{bmatrix} Y & C1 & C2 \\ 0 & 5 & 55 \\ 1 & 55 & 55 \end{bmatrix} \quad (1)$$

As seen from (1),when Y is 0, we see that there is a min of 5 errors and when Y is 1, there is a min of 5 errors. So for Y, it has an error rate of $\frac{5+5}{100}$;.1

$$\begin{bmatrix} Z & C1 & C2 \\ 0 & 15 & 45 \\ 1 & 45 & 15 \end{bmatrix} \quad (2)$$

As seen from (2),when Z is 0, we see that there is a min of 15 errors and when Z is 1, there is a min of 15 errors. So for X, it has an error rate of $\frac{15+15}{100}$;.3
Since Y is has the lowest error rate, we split at Y

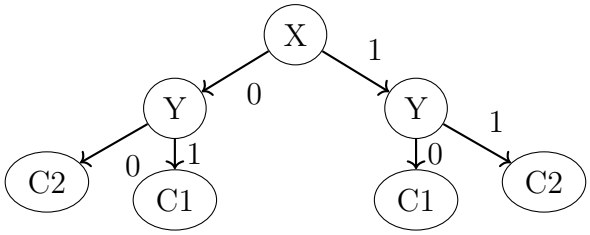
X=1

$$\begin{bmatrix} Y & C1 & C2 \\ 0 & 35 & 5 \\ 1 & 5 & 35 \end{bmatrix} \quad (1)$$

As seen from (1),when Y is 0, we see that there is a min of 5 errors and when Y is 1, there is a min of 5 errors. So for X, it has an error rate of $\frac{5+5}{100}$;.1

$$\begin{bmatrix} Z & C1 & C2 \\ 0 & 15 & 25 \\ 1 & 25 & 15 \end{bmatrix} \quad (2)$$

As seen from (2),when Z is 0, we see that there is a min of 15 errors and when Z is 1, there is a min of 15 errors. So for X, it has an error rate of $\frac{15+15}{100}$;.3
Since Y has the lowest rate of error, we split at Y
The 2 level decision tree will now looks like this



This Decision tree has an error of $frac5 + 5 + 5 + 5200$, 0.1
3.7c)
We see that the decision tree for answer 3.7b has a lower error rate. This demonstrates that a greedy split is not always hueristic
3.8a)
From the table we first find the original Entropy, E_{orig}. This can be found as this

$$\begin{aligned} E_{orig} &= 1 - max(\frac{50}{100}, \frac{50}{100}) \\ &= 1 - \frac{50}{100} \\ &= \frac{1}{2} \end{aligned} \tag{1}$$

Now we can find teh entropy for each of the possible splits
for ΔE_A

$$\begin{bmatrix} A & + & - \\ T & 25 & 0 \\ F & 25 & 50 \end{bmatrix} \tag{2}$$

$$\begin{aligned} E_{A=T} &= 1 - max(\frac{25}{25}, \frac{0}{25}) \\ &= 1 - \frac{25}{25} \\ &= 0 \end{aligned} \tag{3}$$

$$\begin{aligned} E_{A=F} &= 1 - max(\frac{25}{75}, \frac{50}{75}) \\ &= 1 - \frac{50}{75} \\ &= \frac{1}{3} \end{aligned} \tag{4}$$

$$\begin{aligned} \Delta E &= E_{orig} - \ * E_{A=T} - \ * E_{A=F} \\ &= \frac{1}{2} - \frac{25}{100} * 0 - \frac{75}{100} * \frac{1}{3} \\ &= \frac{1}{4} \end{aligned} \tag{5}$$

for ΔE_B

$$\begin{bmatrix} B & + & - \\ T & 30 & 20 \\ F & 20 & 30 \end{bmatrix} \tag{6}$$

$$
\begin{aligned}
E_{B=T} &= 1 - max(\frac{30}{50}, \frac{20}{50}) \\
&= 1 - \frac{30}{50} \\
&= \frac{2}{5}
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
E_{B=F} &= 1 - max(\frac{20}{50}, \frac{30}{50}) \\
&= 1 - \frac{30}{50} \\
&= \frac{2}{5}
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
\Delta E &= E_{orig} - \frac{50}{100} * E_{B=T} - \frac{50}{100} * E_{B=F} \\
&= \frac{1}{2} - \frac{50}{100} * \frac{2}{5} - \frac{50}{100} * \frac{2}{5} \\
&= \frac{1}{5}
\end{aligned} \tag{9}
$$

for ΔE_C

$$
\begin{bmatrix} C & + & - \\ T & 25 & 25 \\ F & 25 & 25 \end{bmatrix} \tag{10}
$$

$$
\begin{aligned}
E_{C=T} &= 1 - max(\frac{25}{50}, \frac{25}{50}) \\
&= 1 - \frac{25}{50} \\
&= \frac{1}{2}
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
E_{C=F} &= 1 - max(\frac{25}{50}, \frac{25}{50}) \\
&= 1 - \frac{25}{50} \\
&= \frac{1}{2}
\end{aligned} \tag{13}
$$

$$
\begin{aligned}
\Delta E &= E_{orig} - \frac{50}{100} * E_{B=T} - \frac{50}{100} * E_{B=F} \\
&= \frac{1}{2} - \frac{50}{100} * \frac{1}{2} - \frac{50}{100} * \frac{1}{2} \\
&= 0
\end{aligned} \tag{14}
$$

As we can see from (5), (9), and (14) we can conlude to use A as our first splitting point. This is because it has the highest information gain

3.8b)

We can not Split when A = T because it has an E of 0, which means it is a pure node. There is no point of splitting pure nodes.

So we try splitting A=F, the new E_{orig} is $E_{A=F}$

for ΔE_B

$$\begin{bmatrix} B & + & - \\ T & 25 & 20 \\ F & 0 & 30 \end{bmatrix} \tag{1}$$

$$\begin{aligned} E_{B=T} &= 1 - max(\frac{25}{45}, \frac{20}{45}) \\ &= 1 - \frac{25}{45} \\ &= \frac{20}{45} \end{aligned} \tag{2}$$

$$\begin{aligned} E_{B=F} &= 1 - max(\frac{0}{30}, \frac{30}{30}) \\ &= 1 - \frac{30}{30} \\ &= 0 \end{aligned} \tag{3}$$

$$\begin{aligned} \Delta E &= E_{orig} - \frac{45}{75} * E_{B=T} - \frac{30}{100} * E_{B=F} \\ &= \frac{1}{3} - \frac{45}{75} * \frac{20}{45} - \frac{30}{75} * 0 \\ &= \frac{5}{75} \end{aligned} \tag{4}$$

for ΔE_C

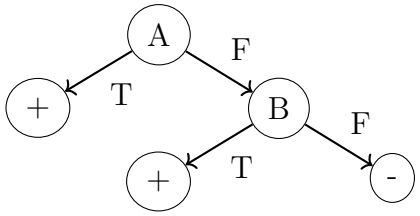
$$\begin{bmatrix} C & + & - \\ T & 0 & 25 \\ F & 25 & 25 \end{bmatrix} \tag{5}$$

$$\begin{aligned} E_{C=T} &= 1 - max(\frac{0}{25}, \frac{25}{25}) \\ &= 1 - \frac{25}{25} \\ &= 0 \end{aligned} \tag{6}$$

$$\begin{aligned} E_{C=F} &= 1 - max(\frac{25}{50}, \frac{25}{50}) \\ &= 1 - \frac{25}{50} \\ &= \frac{1}{2} \end{aligned} \tag{7}$$

$$\begin{aligned} \Delta E &= E_{orig} - \frac{25}{75} * E_{B=T} - \frac{50}{75} * E_{B=F} \\ &= \frac{1}{3} - \frac{25}{75} * 0 - \frac{50}{75} * \frac{1}{2} \\ &= 0 \end{aligned} \tag{8}$$

From what we see from (**??**) and (**??**), we will split B because it is the one with the highest information gain.
3.8c)



The error rate now can be calculate to be .2 because of all the error each of the end nodes total is 20 of 100.