HW2

Task 1

Jupyter notebook code in Github link below.

Task 2

(a) What is the training error rate for the tree? Explain how you get the answer?

I calculated the misclassification error for each leaf in the tree and averaged across all to get:

0.31221

(b) Given a test instance T={A=0, B=1, C=1, D=1, E=0}, what class would the decision tree above assign to T? Explain how you get the answer?

The class would be (-) negative. If you use the instance and follow the path down the tree from root, you reach a leaf where the P(-) is 10/12.

Task 3

Q1: What is the overall entropy before splitting?

$$= -\frac{4}{10} log\left(\frac{4}{10}\right) - \frac{6}{10} log(\frac{6}{10})$$

$$= 0.97095$$

Q2: What is the gain in entropy after splitting on A?

$$Entropy(A = T) = -\frac{4}{7} log\left(\frac{4}{7}\right) - 0log(0) = .46134$$

$$Entropy(A = F) = -0log(0) - \left(\frac{3}{3}\right) log\left(\left(\frac{3}{3}\right)\right) = 0$$

$$Entropy_{gain} = 0.97095 - 0.46134 = 0.509604$$

Q3: What is the gain in entropy after splitting on B:

$$Entropy(B = T) = -\frac{3}{4}log\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)log\left(\left(\frac{1}{4}\right)\right) = 0.8112$$

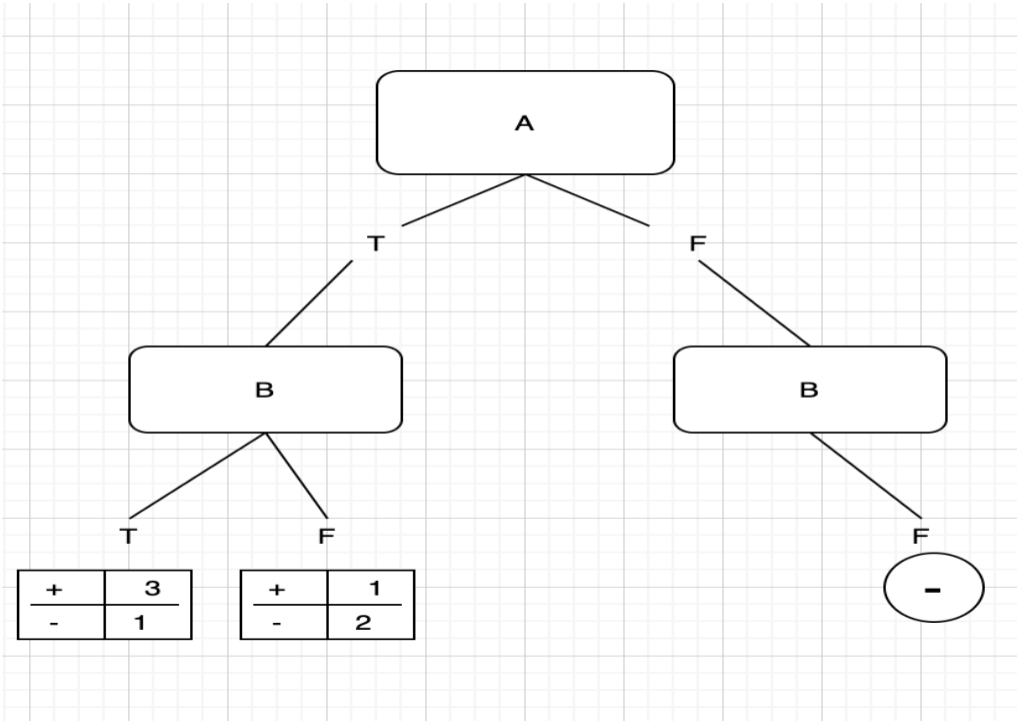
$$Entropy(B = F) = -\frac{1}{6}log\left(\frac{1}{6}\right) - \left(\frac{5}{6}\right)log\left(\left(\frac{5}{6}\right)\right) = 0.65002$$

$$Entropy_{gain} = 0.97095 - 0.8112 - 0.65002 = 0$$

Q4: Which attribute would the decision tree choose?

It would choose attribute A since it has the most information gain.

Q5: Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations. (We want to split first on the variable which maximizes the information gain until there are no nodes with two class labels.)



<u>Task 4: Please answer and explain.</u>

Q1: Are decision trees a linear classifier?

No, decision trees are non-linear because they have the ability to classify non-linear data.

Q2: What are the weaknesses of decision trees?

Some weaknesses of decision trees are their sensitivity to noise in the data and it is expensive to train (time complexity).

Q3: Is Misclassification errors better than Gini index as the splitting criteria for decision trees?

No, Gini Index covers more area (information gain) than the misclassification criteria.

LINK TO GITHUB REPO: https://github.com/CamachoBry/CAP5610-ML/tree/main/Homeworks/HW2

Additional Questions:

•Approximately how many hours did you spend on this assignment?

About 6 hours.

•Which aspects of this assignment did you find most challenging? Were there any significant stumbling blocks?

The most challenging was Task 2, due to not understanding the question (if training error rate meant classification error?)

•Which aspects of this assignment did you like? Is there anything you would have changed?

I enjoyed the coding part, doing feature selection using sklearn and validating. I would have maybe made the rest of the tasks a bit more code oriented.