Ayberk Aydin

Task 2:

a) $Training\ error\ rate = \frac{\#\ incorrect\ predictions}{total\ \#\ predictions} = \frac{6+2+5+5+5+6}{14+5+6+7+2+10+8+6+5+17+15+5} = 0.29$

b) Let - and + be represented as 0 and 1, respectively. We wish to calculate the individual probabilities of these two classes. We first define three branches and calculate these probabilities in each branch:

- $A = 0 \rightarrow B$
  $B = 1 \rightarrow E$
  $E = 0 \rightarrow p(0) = \frac{10}{12} = \frac{5}{6}$
  Since $p(1) = 1 - p(0)$, we have $p(1) = \frac{2}{12} = \frac{1}{6}$
- $C = 1 \rightarrow p(0) = \frac{5}{20} = \frac{1}{4} \rightarrow p(1) = \frac{3}{4}$
- $D = 1 \rightarrow p(0) = \frac{7}{13} \rightarrow p(1) = \frac{6}{13}$

The denominator of the three weights is $12 + 20 + 13 = 45$. Thus, the weights are: $\frac{12}{45}$, $\frac{20}{45}$, and $\frac{13}{45}$, giving us a weighted probability for 1 of $\frac{23}{45} = 0.511$.

Task 3:

Q1: $-0.4 \log_2 0.4 - 0.6 \log_2 0.6 = 0.971$

Q2: T and F entropies:

- T (7/10): $-\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} = 0.985$
- F (3/10): $0 - \log_2\frac{3}{3} = 0$

Gain in entropy = entropy before splitting – (T entropy + F entropy) = 0.971 – ((7/10)*0.985 + (3/10)*0) = 0.281

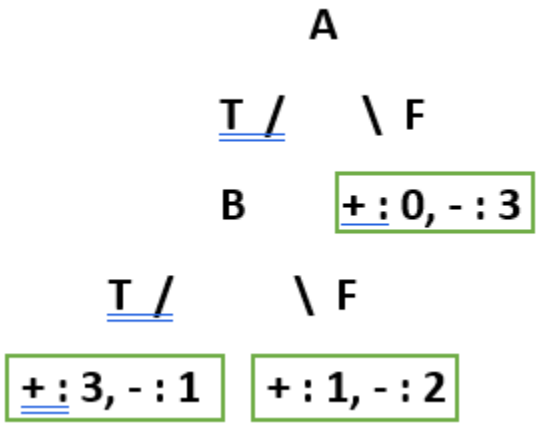
Q3: T and F entropies:

- T (4/10): $-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.811$
- F (6/10): $-\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} = 0.650$

Gain in entropy = entropy before splitting – (T entropy + F entropy) = 0.971 – ((4/10)*0.811 + (6/10)*0.65) = 0.256

Q4: Since $0.281 > 0.256$, the decision tree would choose the A attribute since it has a higher information gain.

Q5:



Task 4:

Q1: No, they are non-linear classifiers like neural networks.

Q2: Small changes in data create large structural changes in decision trees which can cause instability. Decision trees take a relatively long time to train models. Decision trees are insufficient when trying to predict continuous values or for applying regression (K, 2020).

Q3: No, the Gini index is better because misclassification error is not sensitive enough to differences in class probabilities (Peixeiro, 2020).

References:

[i] K, D. (2020, August 22). Top 5 advantages and disadvantages of Decision Tree Algorithm. Retrieved September 17, 2020, from https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a

[ii] Peixeiro, M. (2020, May 19). Everything You Need to Know About Decision Trees. Retrieved September 17, 2020, from https://towardsdatascience.com/everything-you-need-to-know-about-decision-trees-8fcd68ecaa71

Associated code can be found here.