Exercises

Part 1 : Section 3.7

1.1) Exercise 6

6.a) Impurity measures for binary classification problem:

Gini index for parent node:

Total count $= 7 + 3 = 10$

$7 \in class\, 0$, $3 \in class\, 1$

$$I_g = 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2$$

$$= 1 - 0.49 - 0.09$$

$$= 0.42$$

Misclassification error rate:

$$= 1 - max\left[\frac{7}{10}, \frac{3}{10}\right]$$

$$= 1 - 0.7$$

$$= 0.3$$

6.b) **Gini index** of the child nodes:

Node C1: Total count $= 3$

$$3 \in class\, 0,\ 0 \in class\, 1$$

$$I_g = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2$$

$$= 1 - 1 - 0$$

$= 0$

Node C2: Total count $= 4 + 3 = 7$

$$4 \in class\, 0,\ 3 \in class\, 1$$

$$I_g = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2$$

$$= 1 - 0.32 - 0.18$$

$$= 0.5$$

Weighted Gini Index of child nodes $= \frac{3}{10} \cdot 0 + \frac{7}{10} \cdot 0.5 = 0.35$

Gini Measure $= 0.42 - 0.35 \; = \; 0.07$

Yes, I would consider this attribute test condition because of the significant drop in Gini Measure.

6.c) **Misclassification** of child nodes

Node C1:

$$= 1 \; - \; max\left[\frac{3}{3}, \frac{0}{3}\right]$$

$$= 1 \; - \; 1$$

$$= 0$$

Node C2:

$$= 1 - \; max\left[\frac{4}{7}, \frac{3}{7}\right]$$

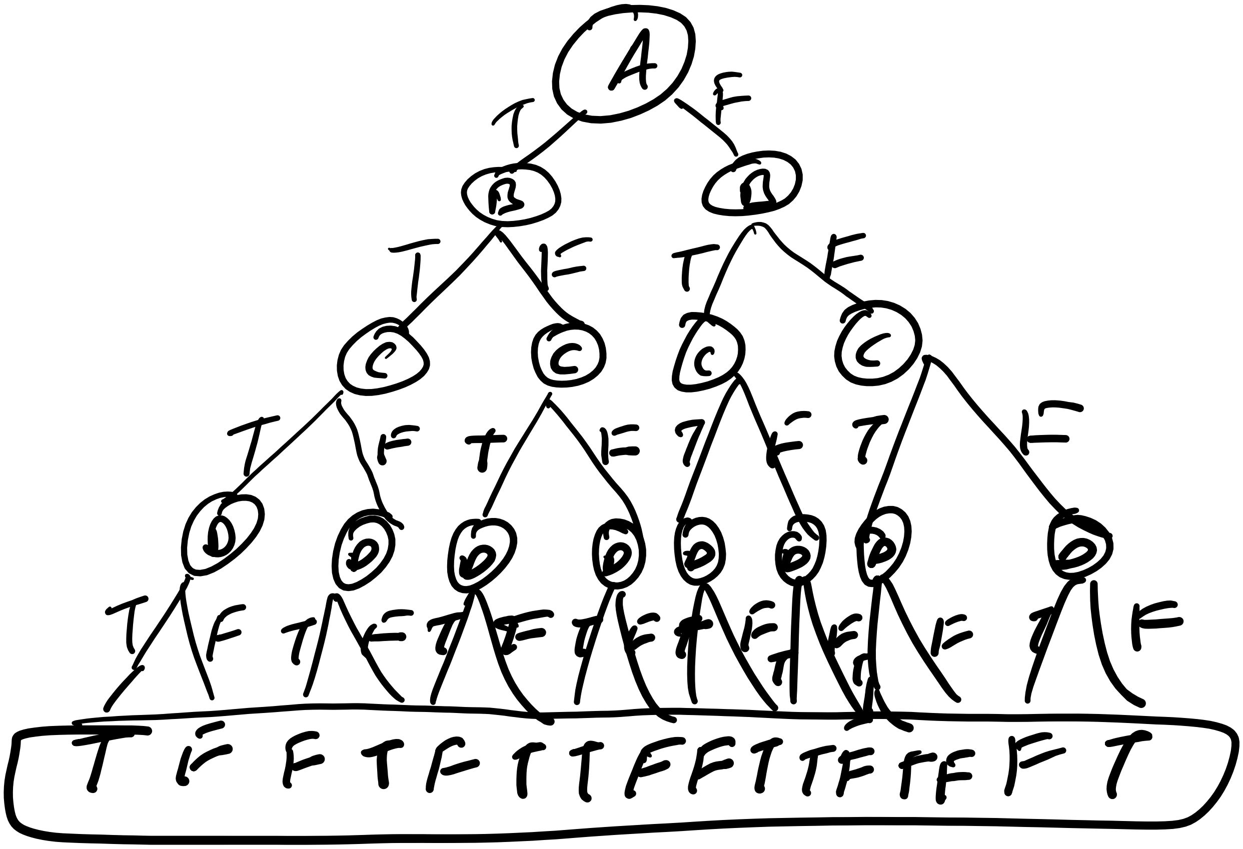
$$= 0.43$$

Weighted Misclassification of child nodes $= \frac{3}{10} \cdot 0 \; + \; \frac{7}{10} \cdot 0.43 \; = \; 0.30$

Misclassification Measure $= \; 0.30 \; - \; 0.30 \; = \; 0$

No, I would not consider this attribute test condition because there is no difference between weighted child node and parent node Misclassification error.

1.2) Exercise 1



The decision tree is already simplified.

1.2) Exercise 3

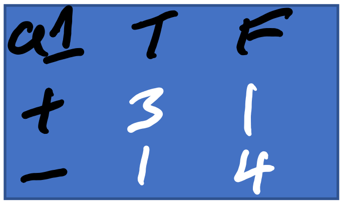
3.a) Total examples = 9, Positive = 4, Negative = 5

Therefore, P+ $= \frac{4}{9}$ and P- $= \frac{5}{9}$

$$\text{Entropy} = -\frac{4}{9}\cdot \log_2\left(\frac{4}{9}\right) - \frac{5}{9}\cdot \log_2\left(\frac{5}{9}\right)$$

$$= 0.9911$$

3.b) Entropy for a1:



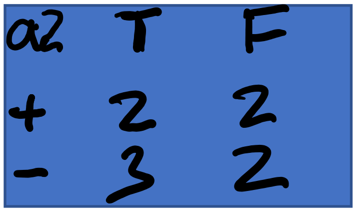
a1	T	F
+	3	1
−	1	4

$$= \frac{4}{9}\left(-\frac{3}{4}\cdot \log_2\frac{3}{4} \;-\frac{1}{4}\cdot \log_2\frac{1}{4}\;\right) + \frac{5}{9}\left(-\frac{1}{5}\cdot \log_2\frac{1}{5}\;-\frac{4}{5}\cdot \log_2\frac{4}{5}\right)$$

$$= 0.7618$$

Information gain for a1 $= 0.9911 - 0.7618 = 0.2292$

Entropy for a2:



a2	T	F
+	2	2
-	3	2

$$= \frac{5}{9}\left(-\frac{2}{5}\cdot\log_2\frac{2}{5} - \frac{3}{5}\cdot\log_2\frac{3}{5}\right) + \frac{4}{9}\left(-\frac{2}{4}\cdot\log_2\frac{2}{4} - \frac{2}{4}\cdot\log_2\frac{2}{4}\right)$$

$$= 0.9839$$

Information gain for a2 $= 0.9911 - 0.9839 = 0.0072$

3.c)

a3	*Class*	*Split point*	*Entropy*	*Gain*
1	+	2.0	0.8484	0.1427
3	-	3.5	0.9885	0.0026
4	+	4.5	0.9183	0.0728
5	-			
5	-	5.5	0.9839	0.0072
6	+	6.5	0.9728	0.0183
7	+			
7	-	7.5	0.8889	0.1022

3.d) According to information gain, a1 produces the best split, since it's the highest.

3.e) The best split a1 since its error rate is lower, $\left(\frac{2}{9} < \frac{4}{9}\right)$.

3.f) Gini Index for a1 is lower therefore a1 is better split.

Gini index for a1 $= 0.3444$

Gini index for a2 $= 0.4889$