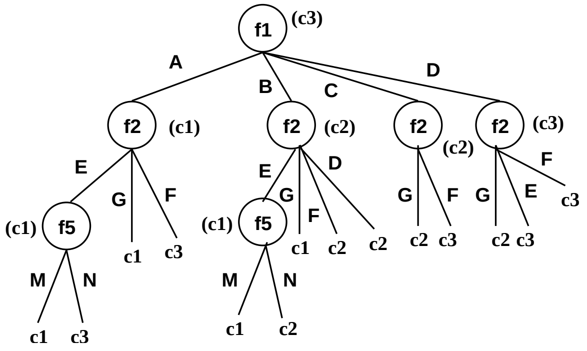
CS 4804 Homework 5
Solution Sketches

1. **(60 points)** The induced tree is given below:



You will notice that two of the test data can be easily classified: (B, G, I, K, N) into c1; and (B, F, J, K, M) into c2. The remaining cannot be classified by the tree unless you have default classifications (denoted in the figure by the classes in parentheses). In this case, the classifications will be: (C, D, J, L, M) into c2; and (C, D, J, L, N) into c2.

2. **(20 points)** Notice that the decision tree is really presenting a disjunction of the boolean variables. The simplest dataset that satisfies the desired conditions is:

f1	f2	f3	f4	c
0	0	0	0	0
1	0	0	0	1
0	1	0	0	1
0	0	1	0	1
0	0	0	1	1
1	1	1	1	1

where the first four columns are the attributes and the last column is the classification. There are many other answers, which would be more elaborate than this dataset.

3. **(20 points)** Using the textbook's terminology, we have to prove that the only case when Gain, given by,

$$I(\frac{p}{p+n}, \frac{n}{p+n}) - \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i})$$

is zero is when the ratios $\frac{p_i}{p_i+n_i}$ are the same. Now, if these ratios are indeed the same, we get

$$\frac{p_i}{p_i+n_i} = \frac{p}{p+n}$$

and similarly,

$$\frac{n_i}{p_i+n_i} = \frac{n}{p+n}$$

In which case, we can write the Gain as:

$$I(\frac{p}{p+n}, \frac{n}{p+n}) - \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(\frac{p}{p+n}, \frac{n}{p+n})$$

or

$$I(\frac{p}{p+n}, \frac{n}{p+n})(1 - \frac{\sum_{i=1}^{v}(p_i + n_i)}{p+n})$$

which is zero (using the fact that $p = \sum p_i$ and $q = \sum q_i$).