4033/5033 Assignment: Decision Tree

Sasank Sribhashyam

In this assignment, we will construct a decision tree by hand. Table 1 contains the records of ten students, each described by five features (F1, F2, F3, F4, F5) and one binary label 'GPA'.

Table 1. Student Data Set

Student ID	F1	F2	F3	F4	F5	GPA (A/B)
01	1	1	0	0	1	B
02	1	0	0	1	1	B
03	1	0	1	0	0	B
04	0	0	1	1	0	A
05	1	1	0	1	0	B
06	0	1	0	1	1	A
07	0	1	1	1	1	B
08	1	1	0	0	1	B
09	1	1	1	0	1	A
10	1	0	1	0	0	A

Task 1. Construct a decision tree by hand using Table 1 as training data. Here are some specific criteria to follow when generating the tree – combine entropies of two child nodes by selecting the smallest one (see lecture note for an example) – each feature is only used to split one node. – always first split node of the largest entropy – stop splitting a node when it has zero entropy or its depth becomes 2 (or no more feature to use). Draw the constructed tree (using any proper software; do not hand-draw it) and show it in Figure 1. Put name of a feature inside a node if this feature is used to split that node, and put name of a class inside a leaf node if this class is used to label that node. Break ties based on alphabetical/numerical order e.g., pick A for a tie between A and B, or pick F2 for a tie between F2 and F3.

Task 2. Estimate the expected classification error (defined in lecture note) of your constructed tree based on Table 1. You need to properly elaborate on the estimation process instead of just giving a number.

$$N1 = (2 \text{ As and } 1 \text{ B})$$
$$N2 = (1 \text{ A and } 1 \text{ B})$$
$$N3 = (1 \text{ A and } 4 \text{ B's})$$

The probability of $N1, N2, N3$:

$$P(X \in N1) = \frac{3}{10}$$
$$P(X \in N2) = \frac{2}{10}$$
$$P(X \in N3) = \frac{5}{10}$$

The error of the nodes $N1, N2, N3$:

$$\text{error1} = P(Y \neq B | X \in N1) = \frac{1}{3}$$
$$\text{error2} = P(Y \neq A | X \in N2) = \frac{1}{2}$$
$$\text{error3} = P(Y \neq A | X \in N3) = \frac{1}{5}$$

The expected classification error is :

$$P(X \in N1) \cdot \text{error1} + P(X \in N2) \cdot \text{error2} + P(X \in N3) \cdot \text{error3} = \frac{3}{10} \cdot \frac{1}{3} + \frac{2}{10} \cdot \frac{1}{2} + \frac{5}{10} \cdot \frac{1}{5} = \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = \frac{3}{10} = 0.3.$$

Task 3. Prune your constructed tree by merging two child nodes into their parent. You should pick the two child nodes that lead to a pruned tree which has the lowest (estimated) expected classification error. Draw the pruned tree in Figure 2 and show its expected classification error.

$$N1 = (3A \text{ and } 2B)$$
$$N2 = (1A \text{ and } 4B)$$
$$P(X \in N1) = \frac{5}{10}$$
$$P(X \in N2) = \frac{5}{10}$$
$$\text{error1} = P(Y \neq B | X \in N1) = \frac{1}{5}$$
$$\text{error2} = P(Y \neq A | X \in N2) = \frac{2}{5}$$
$$\begin{aligned} \text{Expected Classification Error} &= P(X \in N1) \times \text{error1} + P(X \in N2) \times \text{error2} \\ &= \frac{5}{10} \times \frac{1}{5} + \frac{5}{10} \times \frac{2}{5} = \frac{1}{10} + \frac{2}{10} = \frac{3}{10} = 0.3 \end{aligned}$$

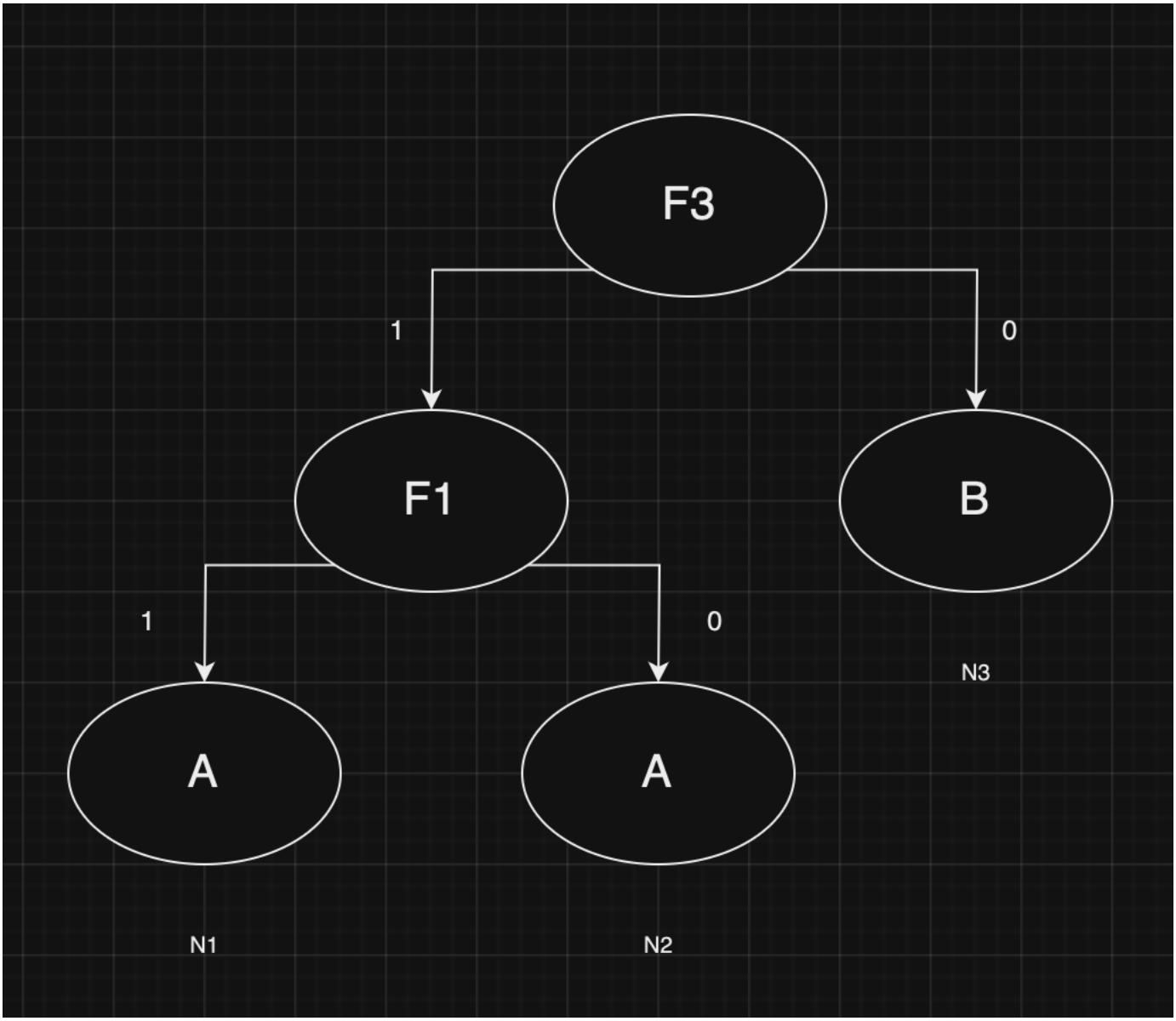

Fig. 1. Constructed Decision Tree Classifier

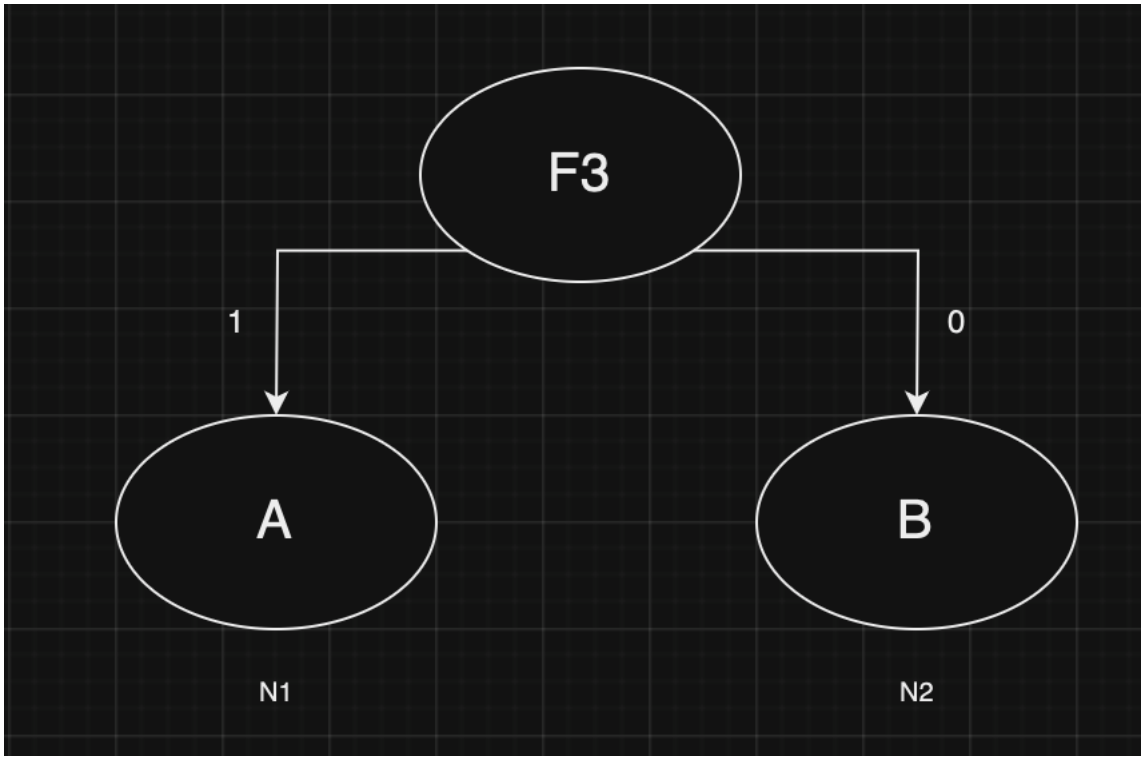

Fig. 2. Pruned Decision Tree Classifier