Desiderio Pilla
9/18/19
CISC 684 Homework #1

1.a) Four attributes: X_1, X_2, X_3, X_4

$E(p_0, p_1) = -p_0 \log_2 p_0 - p_1 \log_2 p_1$

Initial Entropy: $E(.5, .5) = 1$

X_1: $n_{00} = 4$ $n_{01} = 2$ $(0.6)E(4/6, 2/6) = 0.55$
$n_{10} = 1$ $n_{11} = 3$ $(0.4)E(1/4, 3/4) = 0.10$
Gain $= 1 - 0.55 - 0.10 = 0.35$

X_2: $n_{00} = 5$ $n_{01} = 2$ $(0.7)E(5/7, 2/7) = 0.604$
$n_{10} = 0$ $n_{11} = 3$ $(0.3)E(3/3) = 0$
Gain $= 1 - 0.604 - 0 = 0.395$

X_3: $n_{00} = 2$ $n_{01} = 2$ $(0.4)E(2/4, 2/4) = 0.4$
$n_{10} = 3$ $n_{11} = 3$ $(0.6)E(3/6, 3/6) = 0.6$
Gain $= 1 - 0.4 - 0.6 = 0$

X_4: $n_{00} = 3$ $n_{01} = 3$ $(0.6)E(3/6, 3/6) = 0.6$
$n_{10} = 2$ $n_{11} = 2$ $(0.4)E(2/4, 2/4) = 0.4$
Gain $= 1 - 0.6 - 0.4$

X_2 has the largest Gain, so it is the first node.

$X_2 = 0$, $E(5, 2) = 0.863$

X_1: $n_{00} = 4$ $n_{01} = 1$ $(\frac{5}{7})E(4/5, 1/5) = 0.516$
$n_{10} = 1$ $n_{11} = 1$ $(\frac{2}{7})E(1/2, 1/2) = 0.286$
Gain $= 0.863 - 0.516 - 0.286 = 0.061$

X_3: $n_{00} = 2$ $n_{01} = 1$ $(3/7)E(2/3, 1/3) = 0.394$
$n_{10} = 3$ $n_{11} = 1$ $(4/7)E(3/4, 1/4) = 0.464$
Gain $= 0.863 - 0.394 - 0.464 = 0.006$

X_4: $n_{00} = 3$ $n_{01} = 1$ $(4/7)E(3/4, 1/4) = 0.464$
$n_{10} = 2$ $n_{11} = 1$ $(3/7)E(2/3, 1/3) = 0.394$
Gain $= 0.863 - 0.464 - 0.394 = 0.006$

X_1 has the largest gain, so it is the next node.

$X_2 = 1$: $E(0,3) = 0$

The gain is already zero, so no more nodes are needed

$X_2 = 0 \rightarrow X_1 = 0$: $E(4,1) = 0.722$

X_3 : $n_{00} = 1$ $n_{01} = 1$ $(2/5)\,E(1/2, 1/2) = 0.4$

$n_{10} = 3$ $n_{11} = 0$ $(3/5)\,E(1) = 0$

Gain = $0.722 - 0.4 - 0 = 0.322$

X_4 : $n_{00} = 2$ $n_{01} = 0$ $(2/5)\,E(2) = 0$

$n_{10} = 2$ $n_{11} = 1$ $(3/5)\,E(2/3, 1/3) = 0.551$

Gain = $0.722 - 0 - 0.551 = 0.171$

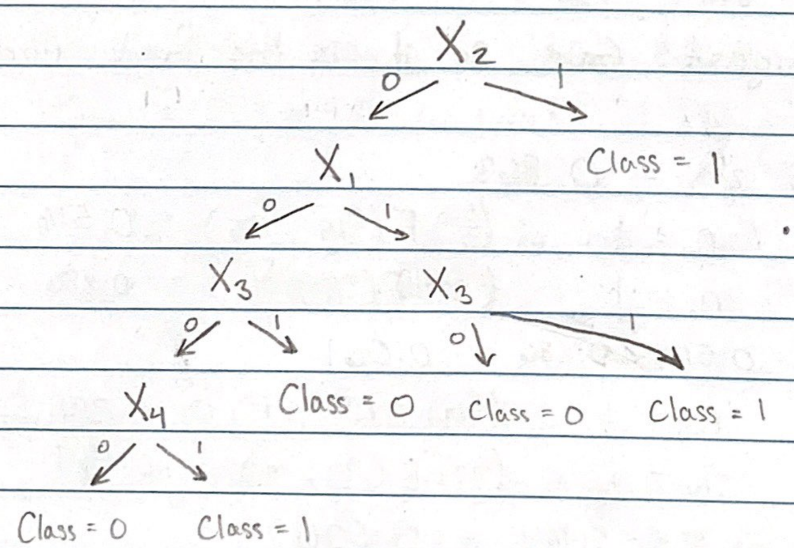
X_3 has the largest gain, so it is the third node, and X_4 is the last node (for $X_3 = 0$ only)

$X_2 = 0 \rightarrow X_1 = 1$: $E(1,1) = 0.5$

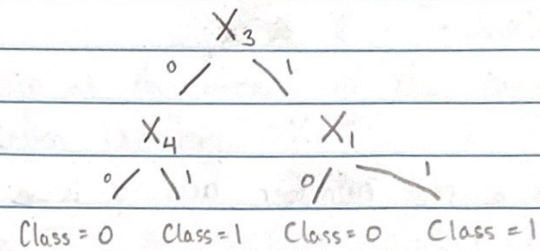
X_3 : $n_{00} = 1$ $n_{01} = 0$ $(1/2)\,E(1) = 0$

$n_{10} = 0$ $n_{11} = 1$ $(1/2)\,E(1) = 0$

Gain = $0.5 - 0 - 0 = 0.5$ = max possible gain, no further nodes needed.



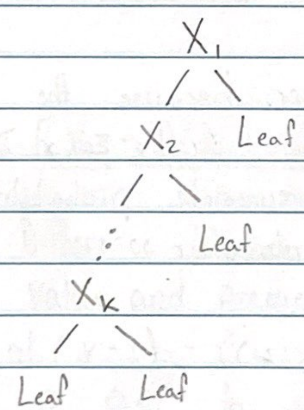
b) 4 leafs, 3 internal nodes, depth of 2



This tree also matches the data with 100% accuracy

c) The decision tree in (b) should be preferred because it is shorter and has less leaves.

2. X is a vector of n Booleans $\{X_1, X_2, \dots X_n\}$
k is an integer that is less than n.
f_k is a target concept which is a disjunction consisting of K literals



The smallest possible consistent decision tree for f_k would have a depth of K, with a leaf at every depth along the branch.