assn2.3

Daniel Loi
dtloi@ucsc.edu 1547401

Kevin Arellano
kcarella@ucsc.edu PUT ID NUMBER HERE

Claudio Sangeroki
csangero@ucsc.edu PUT ID NUMBER HERE

October 2019

1. (a)

$$AI = [70, 75, 80, 85, 85, 90, 100, 150]$$

$$a = [72.5, 77.5, 82.5, 85, 87.5, 95, 125]$$

Because AI is a continuous variable, we can partition it into ranges of the form AI $< a$ and AI $\geq a$, where a is the midpoint between the AI[i], where AI[i] is the current value of AI we are considering, and AI[i+1] values. Therefore, if we have N AI values, then we only need to consider $N-1$ partition values , or a values. In this case, $N = 8$, $a = 7$.

(b) Let $p_+ = P(B) = \frac{1}{2}$, $p_- = P(H) = \frac{1}{2}$, and $S =$ Preference. Therefore,

$$Entropy(S) = \frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = \frac{1}{2}$$

(c) In order to find the optimal root, we will have to calculate the information gain between Gender vs. AI.

i. In order to find out if we should split on Gender, we will calculate

$$Gain(S, Gender) = E(S) - (E(S_1) + E(S_2))$$

where $S_1 = M$ and $S_2 = F$.

$$E(S_1) = \frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) = 0.92$$

$$E(S_2) = \frac{2}{5}log_2(\frac{2}{5}) - \frac{3}{5}log_2(\frac{3}{5}) = 0.97$$

Therefore,

$$Gain(S, Gender) = 1 - ((\frac{3}{8})0.92 + (\frac{5}{8})0.97) = 0.04875$$

ii. In order to find out if we should split on $a = 72.5$, we will calculate

$$Gain(S, a) = E(S) - (E(S_1) + E(S_2))$$

where $S_1 = a < 72.5$ and $S_2 = a \geq 72.5$.

$$E(S_1) = \frac{1}{1}log_2(\frac{1}{1}) - \frac{0}{1}log_2(\frac{0}{1}) = 0$$

$$E(S_2) = \frac{3}{7}log_2(\frac{3}{7}) - \frac{4}{7}log_2(47) = 0.99$$

Therefore,

$$Gain(S, a) = 1 - ((\frac{1}{8})0 + (\frac{7}{8})0.99) = 0.13375$$

iii. In order to find out if we should split on $a = 77.5$, we will calculate

$$Gain(S, a) = E(S) - (E(S_1) + E(S_2))$$

where $S_1 = a < 77.5$ and $S_2 = a \geq 77.5$.

$$E(S_1) = \frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$$

$$E(S_2) = \frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$$

Therefore,

$$Gain(S, a) = 1 - ((\frac{2}{8})1 + (\frac{6}{8})1) = 0$$

iv. In order to find out if we should split on $a = 82.5$, we will calculate

$$Gain(S, a) = E(S) - (E(S_1) + E(S_2))$$

where $S_1 = a < 82.5$ and $S_2 = a \geq 82.5$.

$$E(S_1) = \frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) = 0.91$$

$$E(S_2) = \frac{2}{5}log_2(\frac{2}{5}) - \frac{3}{5}log_2(\frac{3}{5}) = 0.97$$

Therefore,

$$Gain(S, a) = 1 - ((\frac{3}{8})0.91 + (\frac{5}{8})0.97) = 0.05250$$

v. In order to find out if we should split on $a = 85$, we will calculate

$$Gain(S, a) = E(S) - (E(S_1) + E(S_2))$$

where $S_1 = a < 85$ and $S_2 = a \geq 85$.

$$E(S_1) = \frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) = 0.91$$

$$E(S_2) = \frac{2}{5}log_2(\frac{2}{5}) - \frac{3}{5}log_2(\frac{3}{5}) = 0.97$$

Therefore,

$$Gain(S, a) = 1 - ((\frac{3}{8})0.91 + (\frac{5}{8})0.97) = 0.05250$$

vi. In order to find out if we should split on $a = 87.5$, we will calculate

$$Gain(S, a) = E(S) - (E(S_1) + E(S_2))$$

where $S_1 = a < 87.5$ and $S_2 = a \geq 87.5$.

$$E(S_1) = \frac{2}{5}log_2(\frac{2}{5}) - \frac{3}{5}log_2(\frac{3}{5}) = 0.97$$

$$E(S_2) = \frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) = 0.91$$

Therefore,

$$Gain(S, a) = 1 - ((\frac{5}{8})0.97 + (\frac{3}{8})0.91) = 0.05250$$

vii. In order to find out if we should split on $a = 95$, we will calculate

$$Gain(S, a) = E(S) - (E(S_1) + E(S_2))$$

where $S_1 = a < 95$ and $S_2 = a \geq 95$.

$$E(S_1) = \frac{4}{6}log_2(\frac{4}{6}) - \frac{2}{6}log_2(\frac{2}{6}) = 0.91$$

$$E(S_2) = \frac{1}{1}log_2(\frac{1}{1}) - \frac{0}{1}log_2(\frac{0}{1}) = 0$$

Therefore,

$$Gain(S, a) = 1 - ((\frac{6}{8})0.91 + (\frac{2}{8})0) = 0.3175$$

viii. In order to find out if we should split on $a = 125$, we will calculate

$$Gain(S, a) = E(S) - (E(S_1) + E(S_2))$$

where $S_1 = a < 125$ and $S_2 = a \geq 125$.

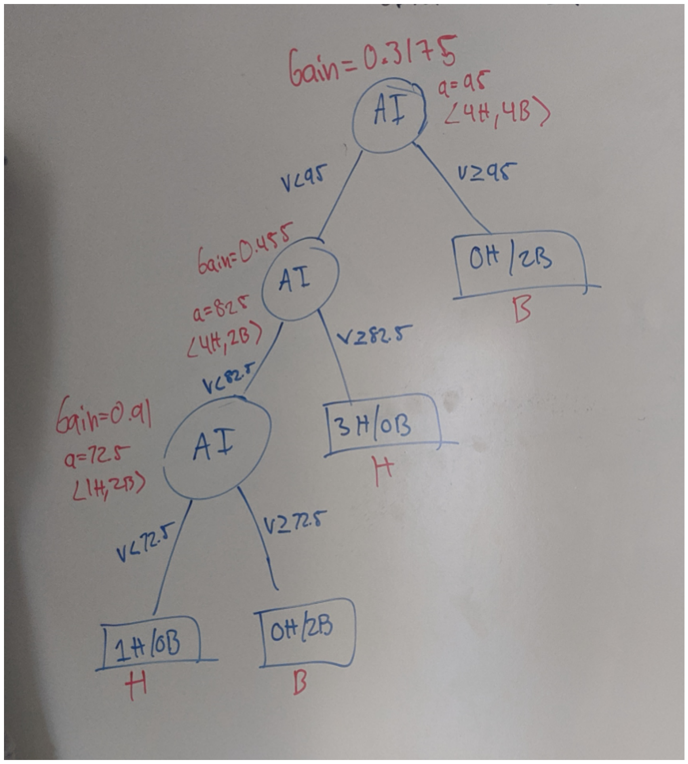
$$E(S_1) = \frac{3}{7}log_2(\frac{3}{7}) - \frac{4}{7}log_2(\frac{4}{7}) = 0.99$$

$$E(S_2) = \frac{2}{2}log_2(\frac{2}{2}) - \frac{0}{2}log_2(\frac{0}{2}) = 0$$

Therefore,

$$Gain(S, a) = 1 - ((\frac{7}{8})0.99 + (\frac{1}{8})0) = 0.13375$$

Since $a = 95$ gave us the highest information gain, it is the optimal root node to split on.

(d) 
Gain=0.3175
AI
a=95
<4H,4B>
V<95
V≥95
Gain=0.455
AI
a=82.5
<4H,2B>
V<82.5
V≥82.5
0H/2B
B
Gain=0.91
a=72.5
<1H,2B>
AI
3H/0B
H
V<72.5
V≥72.5
1H/0B
H
0H/2B
B