Problem 1

Part 1

Early registration	Finished homework II	Senior	Likes Coffee	Liked The Last homework	A
1	1	0	0	1	1
1	1	1	0	1	1
0	0	1	0	0	0
0	1	1	0	1	0
0	1	1	0	0	1
0	0	1	1	1	1
1	0	0	0	1	0
0	1	0	1	1	1
0	0	1	0	1	1
1	0	0	0	0	0
1	1	1	0	0	1
0	1	1	1	1	0
0	0	0	0	1	0
1	0	0	1	0	1

$$Entropy = -P_+log_2 - P\_log_2$$

$$Gain = E_s - \sum \frac{|s_v|}{|s|} E_{s_v}$$

Entropy for data

- Number of instences=14
- Number of postive(+)=8
- Number of negative(-)=6
- $$E = -\frac{8}{14}log_2\frac{8}{14} - \frac{6}{14}log_2\frac{6}{14} = 0.985$$

Early Registration

- Number of ones=6

> 4(+) And 2(-)

- Number of zeros=8

> 4(+) And 4(-)

$$E(s_1) = -\frac{4}{6}log_2\frac{4}{6} - \frac{2}{6}log_2\frac{2}{6} = 0.9182$$

$$E(s_0) = -\frac{4}{8}log_2\frac{4}{8} - \frac{4}{8}log_2\frac{4}{8} = 1$$

$$gain = 0.985 - (\frac{6}{14} * 0.981 + \frac{8}{14} * 1) = 0.02$$

Finished HomeWork

- Number of ones=7

> 5(+) And 2(-)

- Number of zeros=7

> 3(+) And 4(-)

$$E(s_1) = -\frac{5}{7}log_2\frac{5}{7} - \frac{2}{7}log_2\frac{2}{7} = 0.863$$

$$E(s_0) = -\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7} = 0.9852$$

$$gain = 0.985 - (\frac{7}{14} * 0.863 + \frac{7}{14} * 0.9852) = 0.06$$

Senior

- Number of ones=8

> 5(+) And 3(-)

- Number of zeros=6

> 3(+) And 3(-)

$$E(s_1) = -\frac{5}{8}log_2\frac{5}{8} - \frac{3}{8}log_2\frac{3}{8} = 0.9544$$

$$E(s_0) = -\frac{3}{6}log_2\frac{3}{6} - \frac{3}{6}log_2\frac{3}{6} = 1$$

$$gain = 0.985 - (\frac{8}{14} * 0.9544 + \frac{6}{14} * 1) = 0.01$$

Likes Coffee

- Number of ones=4

> 3(+) And 1(-)

- Number of zeros=10

> 5(+) And 5(-)

$$E(s_1) = -\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4} = 0.81127$$

$$E(s_0) = -\frac{5}{10}log_2\frac{5}{10} - \frac{5}{10}log_2\frac{5}{10} = 1$$

$$gain = 0.985 - (\frac{4}{14} * 0.81127 + \frac{10}{14} * 1) = 0.03$$

Liked The Last homework

- Number of ones=9

> 5(+) And 4(-)

- Number of zeros=5

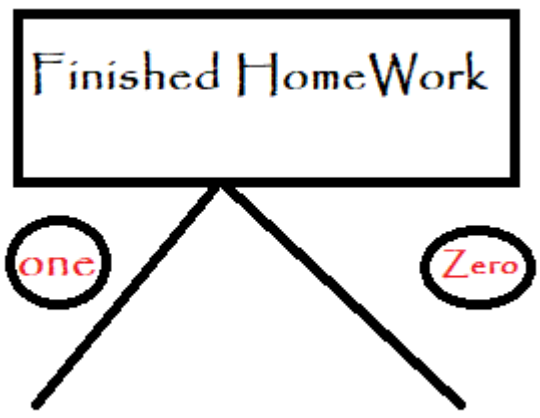
> 3(+) And 2(-)

$$E(s_1) = -\frac{5}{9}log_2\frac{5}{9} - \frac{4}{9}log_2\frac{4}{9} = 0.991$$

$$E(s_0) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.9709$$

$$gain = 0.985 - (\frac{9}{14} * 0.991 + \frac{5}{14} * 0.9709) = 0.001$$

Root Is Finished HomeWork



Branch of ONE

Finished homework II	Early registration	Senior	Likes Coffee	Liked The Last homework	A
1	1	0	0	1	1
1	1	1	0	1	1
1	0	1	0	1	0
1	0	1	0	0	1
1	0	0	1	1	1
1	1	1	0	0	1
1	0	1	1	1	0

Entropy for data

- Number of instences=7
- Number of postive(+)=5
- Number of negative(-)=2

$$E = -\frac{5}{7}log_2\frac{5}{7} - \frac{2}{7}log_2\frac{2}{7} = 0.863$$

Early Registration

- Number of ones=3

> 3(+) And 0(-)

- Number of zeros=4

> 2(+) And 2(-)

$$E(s_1) = -\frac{3}{3}log_2\frac{3}{3} - \frac{0}{3}log_2\frac{0}{3} = 0$$

$$E(s_0) = -\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4} = 1$$

$$gain = 0.863 - (\frac{3}{7} * 0 + \frac{4}{7} * 1) = 0.291$$

Senior

- Number of ones=5

> 3(+) And 2(-)

- Number of zeros=2

> 2(+) And 0(-)

$$E(s_1) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.9709$$

$$E(s_0) = -\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2} = 0$$

$$gain = 0.863 - (\frac{5}{7} * 0.9709 + \frac{2}{7} * 0) = 0.4469$$

Likes Coffee

- Number of ones=2

> 1(+) And 1(-)

- Number of zeros=4

> 4(+) And 1(-)

$$E(s_1) = -\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} = 1$$

$$E(s_0) = -\frac{4}{5}log_2\frac{4}{5} - \frac{1}{5}log_2\frac{1}{5} = 0.7219$$

$$gain = 0.863 - (\frac{2}{7} * 1 + \frac{5}{7} * 0.7219) = 0.06$$

Liked The Last homework

- Number of ones=5

> 3(+) And 2(-)

- Number of zeros=2

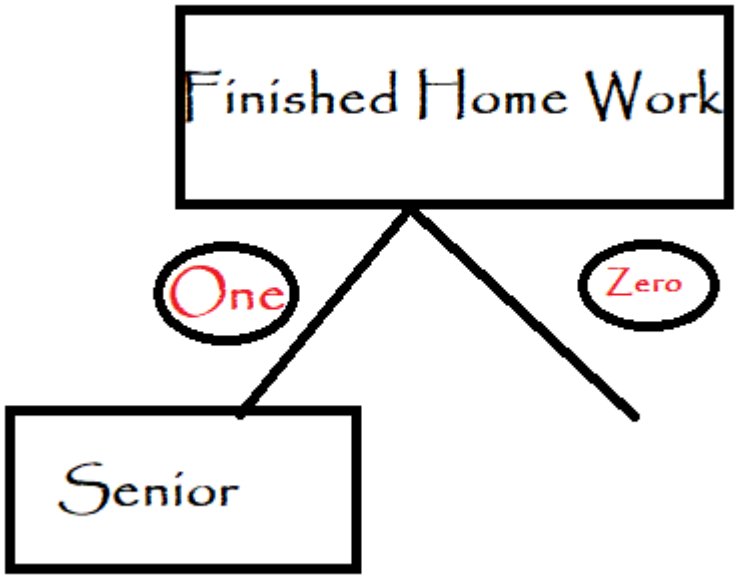
> 2(+) And 0(-)

$$E(s_1) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.9709$$

$$E(s_0) = -\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2} = 0$$

$$gain = 0.863 - (\frac{5}{7} * 0.9709 + \frac{2}{7} * 0) = 0.0169$$

Second Node in Branch of 1 Is Senior



Branch of Zero

Finished homework II	Early registration	Senior	Likes Coffee	Liked The Last homework	A
0	0	1	0	0	0
0	0	1	1	1	1
0	1	0	0	1	0
0	0	1	0	1	1
0	1	0	0	0	0
0	0	0	0	1	0
0	1	0	1	0	1

Entropy for data

- Number of instences=7
- Number of postive(+)=3
- Number of negative(-)=4

$$E = -\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7} = 0.9852$$

Early Registration

- Number of ones=3

> 1(+) And 2(-)

- Number of zeros=4

> 2(+) And 2(-)

$$E(s_1) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.9182$$

$$E(s_0) = -\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4} = 1$$

$$gain = 0.985 - (\frac{3}{7} * 0.9182 + \frac{4}{7} * 1) = 0.02$$

Senior

- Number of ones=3

> 2(+) And 1(-)

- Number of zeros=4

> 1(+) And 3(-)

$$E(s_1) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.918$$

$$E(s_0) = -\frac{1}{4}log_2\frac{1}{4} - \frac{3}{4}log_2\frac{3}{4} = 0.811$$

$$gain = 0.985 - (\frac{3}{7} * 0.9182 + \frac{4}{7} * 0.811) = 0.128$$

Likes Coffee

- Number of ones=2

> 2(+) And 0(-)

- Number of zeros=5

> 1(+) And 4(-)

$$E(s_1) = -\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2} = 0$$

$$E(s_0) = -\frac{1}{5}log_2\frac{1}{5} - \frac{4}{5}log_2\frac{4}{5} = 0.7219$$

$$gain = 0.985 - (\frac{2}{7} * 0 + \frac{5}{7} * 0.7219)$$

Liked The Last homework

- Number of ones=4

> 2(+) And 2(-)

- Number of zeros=3

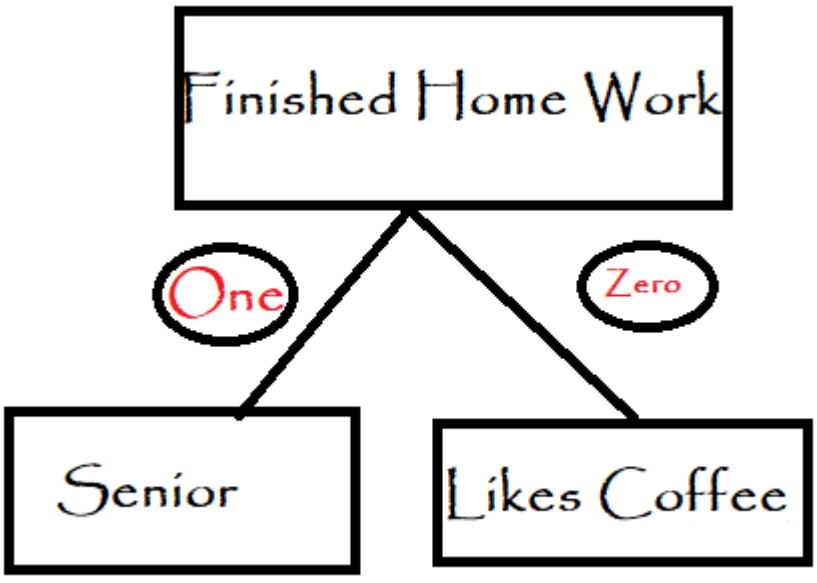
> 1(+) And 2(-)

$$E(s_1) = -\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4} = 1$$

$$E(s_0) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.9182$$

$$gain = 0.985 - (\frac{4}{7} * 1 + \frac{3}{7} * 0.9182) = 0.02$$

`Second Node in Branch of 0 Is Senior`



Part 2

- I Think Each Algorithm has advantage and disadvantages
- C4.5 Cause or construct empty branches or over fitting so it is worse than ID3 to create tree eith less deep
- ID3 Cause Over_fitting or over classified if small sample is tested
- May be CART is better because it can handle numerical and catagorical data , it can identify significant values and eliminate non-significant

> each one of them use different criteria to be created such as ID3 use `Information Gain` and C4.5 use `Gain Ratio` and CART use `Gini Impurity`

> I think too if we used Random forest concept, may can build many tree with less deep and companine them with some technique to make better decision