Name / Mohamed Khaled Gallad

Sec / 2 , B.N. 115

Problem ① $\boxed{1.1}$ * Samples = 14

*** Entropy for the dataset E(S):**

* +ve = 8 , * -ve = 6

$$E(S) = -\left(\frac{8}{14}\log_2\frac{8}{14} + \frac{6}{14}\log_2\frac{6}{14}\right) = \underline{0.985}$$

*** Information gain for "Early registration":**

* 1 [4+, 2-] , * 0 [4+, 4-]

$$E(S_1) = -\left(\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right) = 0.918$$

$$E(S_0) = -\left(\frac{4}{8}\log_2\left(\frac{4}{8}\right) + \frac{4}{8}\log_2\left(\frac{4}{8}\right)\right) = 1$$

$$IG = E(S) - \frac{|S_1|}{|S|}E(S_1) - \frac{|S_0|}{|S|}E(S_0)$$

$$= 0.985 - \frac{6}{14} * 0.918 - \frac{8}{14} * 1 = \boxed{0.02}$$

*** IG for "Finished homework II":-**

* 1 [5+, 2-] , * 0 [3+, 4-]

$$E(S_1) = -\left(\frac{5}{7}\log_2\left(\frac{5}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right)\right) = 0.863$$

$$E(S_0) = -\left(\frac{4}{7}\log_2\left(\frac{4}{7}\right) + \frac{3}{7}\log_2\left(\frac{3}{7}\right)\right) = 0.985$$

$$IG = 0.985 - \frac{7}{14} * 6.863 - \frac{7}{14} * 0.985 = \boxed{0.061}$$

IG for "Senior":

1 [5+, 3-] , 0 [3+, 3-]

$E(S_1) = -\left(\frac{5}{8}\log_2\frac{5}{8} + \frac{3}{8}\log_2\frac{3}{8}\right) = 0.954$

$E(S_0) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1$

$IG = 0.985 - \frac{8}{14} * 0.954 - \frac{6}{14} * 1 = \boxed{0.011}$

IG for "Likes Coffee":

1 [3+, 1-] , 0 [5+, 5-]

$E(S_1) = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = 0.811$

$E(S_0) = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1$

$IG = 0.985 - \frac{4}{14} * 0.811 - \frac{10}{14} * 1 = \boxed{0.039}$

IG for "Liked The last homework":

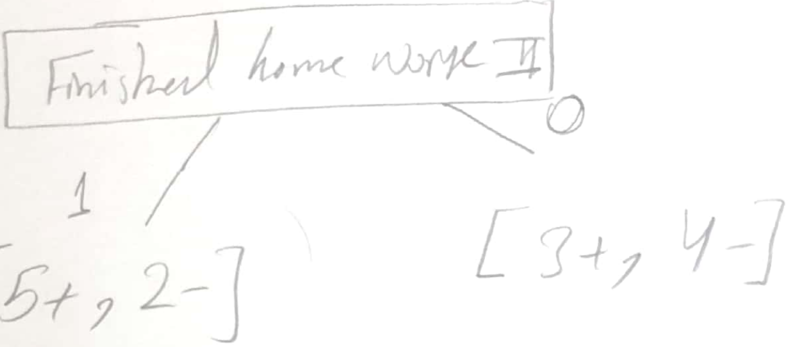
1 [5+, 4-] , 0 [3+, 2-]

$E(S_1) = -\left(\frac{5}{9}\log_2\frac{5}{9} + \frac{4}{9}\log_2\frac{4}{9}\right) = 0.991$

$E(S_0) = -\left(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right) = 0.971$

$IG = 0.985 - \frac{9}{14} * 0.991 - \frac{5}{14} * 0.971 = \boxed{1*14*10^{-3}}$

Depth '0'



Entropy for '1' (Finished home work II):

$$E(S) = -\left(\frac{5}{7}\log_2\frac{5}{7} + \frac{2}{7}\log_2\frac{2}{7}\right) = 0.863$$

Branch (1)

IG for "Early Registration":

1 [3+, 0-] , 0 [2+, 2-]

$$E(S_1) = -\left(\frac{3}{3}\log_2\left(\frac{3}{3}\right) + \frac{0}{3}\log_2\frac{0}{3}\right) = 0$$

$$E(S_0) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$$

$$IG = 0.863 - \frac{3}{7} * 0 - \frac{4}{7} * 1 = \boxed{0.292}$$

IG for "Senior":

1 [3+, 2-] , 0 [2+, 0]

$$E(S_1) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.971$$

$$E(S_0) = -\left(\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right) = 0$$

$$IG = 0.863 - \frac{5}{7} * 0.971 - \frac{2}{7} * 0 = \boxed{0.169}$$

* IG for 'Likes Coffee':

* 1 [1+, 1-] , * 0 [4+, 1-]

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

$E(S) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 0.722$

$IG = 0.863 - \frac{2}{7} * 1 - \frac{5}{7} * 0.722 = \boxed{0.062}$

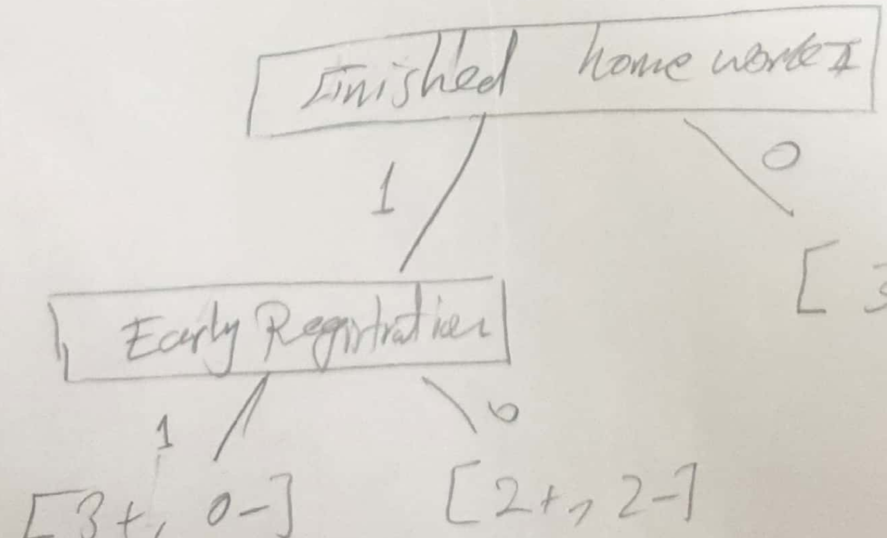
* IG for 'Liked the last homework'

* 1 [3+, 2-] , * 0 [2+, 0-]

$E(S_1) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.971$

$E(S_0) = -\left(\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right) = 0$

$IG = 0.863 - \frac{5}{7} * 0.971 - \frac{2}{7} * 0 = \boxed{0.169}$



Entropy for "0" (Didn't Finish homework II):

$$E(S) = \left[\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right] = 0.985$$

* IG for "Early Registration":

*1 [1+, 2-] , *0 [2+, 2-]

$$E(S_1) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.918$$

$$E(S_0) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$$

$$IG = 0.985 - \frac{3}{7} * 0.918 - \frac{4}{7} * 1 = \boxed{0.02}$$

*IG for "Senior":

*1 [2+, 1-] , *0 [1+, 3-]

$$E(S_1) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.918$$

$$E(S_0) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.811$$

$$IG = 0.985 - \frac{3}{7} * 0.918 - \frac{4}{7} * 0.811 = \boxed{0.128}$$

* IG for "Xmas Coffee":

*1 [2+, 0-] , *0 [1+, 4-]

$$E(S_1) = -\left(\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right) = 0$$

$$E(S_0) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$$

$IG = 0.985 - \frac{2}{7} * 0 - \frac{5}{7} * 0.722 = \boxed{0.469}$

IG for "Liked best homework":

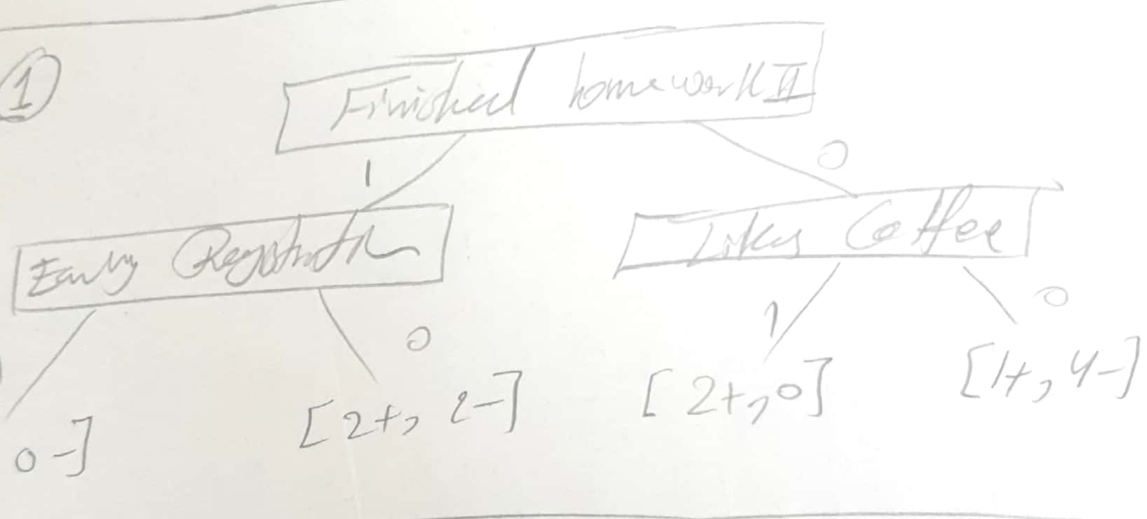
*1 [2+, 2−], *0 [1+, 2−]

$E(S_1) = 1$

$E(S_0) = 0.918$

$IG = 0.985 - \frac{4}{7} * 1 - \frac{3}{7} * 0.918 = \boxed{0.02}$

Depth ①



Starting Depth ②

* Entropy for (1, 1) (Finished & Early Registered)

$E(S) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0$

Decision? will get $\underline{A}$

* Entropy for (1, 0) (Finished & didn't Early)

$E(S) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$

* IG for " Senior":-

*1 [1+ , 2-] , *0 [1+ , 0]

$E(S_1) = 0.918$, $E(S_0) = 0$

$IG = 1 - \frac{3}{4} * 0.918 = \boxed{0.3115}$

* "IG For "Likes Coffee":-

*1 [1+, 1-] , *0 [1+ , 1-]

$E(S_1) = 1$, $E(S_0) = 1$

$IG = 1 - \frac{2}{4} * 1 - \frac{2}{4} * 1 = \boxed{0}$

* IG ~~for~~ "Linked the last homework":

*1 [1+ , 2-] , *0 [1+ , 0]

$E(S_1) = 0.918$, $E(S_0) = 0$

$IG = 1 - \frac{3}{4} * 0.918 - 0 = \boxed{0.3115}$

Either " senior " or " Liked last homework"

I'll choose senior

(Didn't Finish & Liked Coffee)

* Entropy for (0,1)

$E(S) = -\left(\frac{2}{2} \log_2 \frac{2}{2} + \frac{0}{2} \log_2 \frac{0}{2}\right) = 0$

Decision : will get A

*Entropy for (0,0) (Didn't Finish & didn't like [illegible])

$E(S) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$

* IG for "Early Registration":

*1 [0, 2-] , *0 [1+, 2-]

$E(S_1) = 0$, $E(S_0) = 0.918$

$IG = 0.722 - \frac{3}{5} * 0.918 = \boxed{0.1712}$

* IG for "Senior":

*1 [1+, 1-] , *0 [+0, 3-]

$E(S_1) = 1$, $E(S_0) = 0$

$IG = 0.722 - \frac{2}{5} * 1 = \boxed{0.322}$

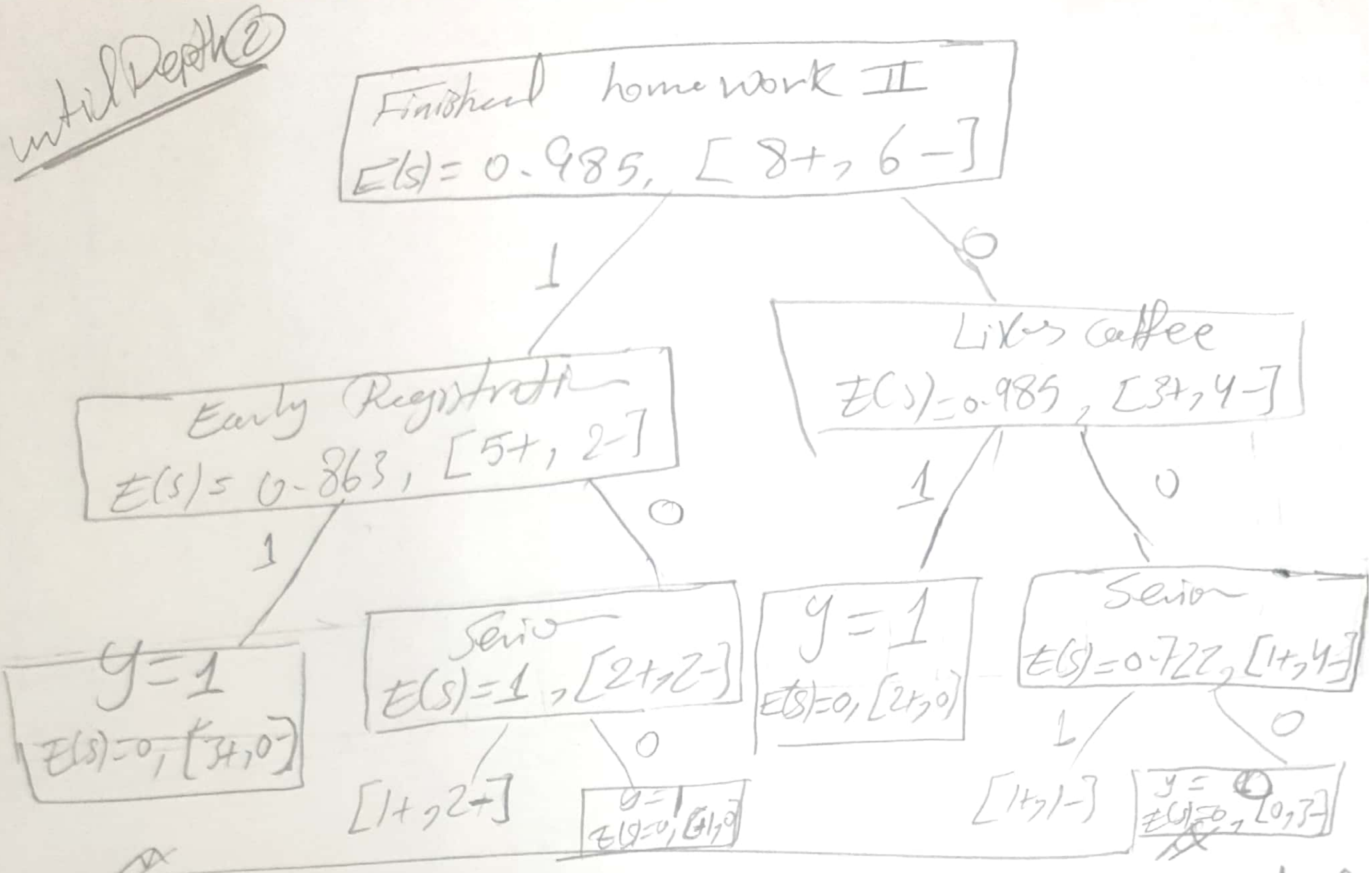
* IG for "Liked the last homework":

*1 [1+, 2-] , *0 [0, 2-]

$E(S_1) = 0.918$, $E(S_0) = 0$

$IG = 0.722 - \frac{3}{5} * 0.918 = \boxed{0.1712}$

So "Senior" will be our internal node



1.2 ID3 Continues to grow a tree until it makes no error over the set of training data in our case will grow until it really depth 3 which can cause over-fitting.

Using C4.5 decision tree can cause less deep tree then ID3 as C4.5 allows pruning which means we can remove branches that don't help by replacing them with leaf nodes, resulting in more robust tree to overfitting & less deeper.