Karasu Naveen Ajay

ID: 11646981

CSCE 5380 - Homework - 2

1. a)

Attributes	Yes Count	No Count
1. Tobacco Smoking	5	5
2. Radon Exposure	2	8
3. Chronic Cough	7	3
4. Weight Loss	5	5
5. Lung Cancer	5	5

· Calculate Initial Entropy:

we have 5 +ve & 5 -ve example of lung Cancer, where N=10

$$E(S) = -\sum_{i=1}^{n} P_i \log_2 (P_i)$$

$$= -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10}$$

$$= 1$$

Attributes:

Tobacco Smoking:

Values → [Yes, No]

S_{Yes} [+4, -1] $E(S_{Yes}) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}$

$$= 0.7219$$

S_{NO} → [+1, -4] ∴ $E(S_{No}) = -\frac{1}{5}\log_2 1/5 - \frac{4}{5}\log_2\frac{4}{5}$

$$= 0.7219$$

$$\therefore Gain(s) = E - \sum_{v \in r} \frac{|S_r|}{|S|} E(S_r)$$

$$= E - \frac{5}{10} E(S_{yes}) - \frac{5}{10} E(S_{No})$$

$$= 1 - \frac{5}{10}(0.7219) - \frac{5}{10}(0.7219)$$

$$Gain(S, \text{Tobacco Smoking}) = \underline{\underline{0.2781}}$$

Attribute: <u>Radon Exposure:</u> Values → [Yes, No]

$S_{yes} \leftarrow [+2, -0]$ $E(S_{yes}) = -\frac{2}{2}\log_2 \frac{2}{2} - \frac{0}{2}\log_2 \frac{0}{2}$

$= -1 \times (0) - 0$

$= 0 - 0$

$= 0$

$S_{No} \rightarrow [+3, -5]$

$E(S_{No}) = -\frac{3}{8}\log_2 \frac{3}{8} - \frac{5}{8}\log_2 \frac{5}{8}$

$= 0.954$

$$Gain(S, \text{Radon}) = E - \sum_{v \in (yes, No)} \frac{|S_r|}{S} E(S_r)$$

$$= 1 - \frac{2}{10} E(S_{yes}) - \frac{8}{10} E(S_{No})$$

$$= 1 - \frac{2}{10}(0) - \frac{8}{10}(0.954)$$

$$Gain(S, \text{Radon}) = \underline{0.2368}$$

Attribute : Chronic Cough :

$S_{Yes} \rightarrow [+4, -3]$ $\quad E(S_{Yes}) = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}$

$= 0.9852$

$S_{No} \rightarrow [+1, -2]$ $\quad E(S_{No}) = -\frac{1}{3}\log_2 1/3 - \frac{2}{3}\log_2 2/3$

$= 0.9182$

$$Gain(S, Chronic) = E - \sum_{v \in (yes, no)} \frac{|S_v|}{S} E(S_v)$$

$$= 1 - \frac{7}{10}E(S_{Yes}) - \frac{3}{10}E(S_{No})$$

$$= 1 - \frac{7}{10}(0.9852) - \frac{3}{10}(0.9182)$$

$$Gain(S, Chronic) = 0.0349$$

Attribute :- Weight loss

S_{Yes} $[+3, -2]$ $\quad E(S_{Yes}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.9709$

S_{No} $[+2, -3]$ $\quad E(S_{No}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.9709$

$$Gain(S, weight) = E - \frac{5}{10}E(S_{Yes}) - \frac{5}{10}E(S_{No})$$

$$= 1 - \frac{5}{10}(0.9709) - \frac{5}{10}(0.9709)$$

$$Gain(S, weight) = 0.0291$$

From the information, we will Consider Tobacco Smoking as a root has it have max. information Gain.

As there is One 'No' for 'Yes' in Cancer. we will the repeat process

=> Tobacco Smoking

Yes | NO

Exposure	Chronic	Weightloss	Cancer
Yes	Yes	NO	Yes
NO	Yes	No	Yes
NO	Yes	Yes	Yes
NO	Yes	Yes	Yes
NO	No	No	No

Exposure	chronic	Weightloss	Cancer.
Yes	No	Yes	Yes
NO	Yes	No	No
No	Yes	Yes	No
No	Yes	No	No
No	NO	Yes	No

In "Yes" $E = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.7219$

Attribute Exposure:-

$S_{Yes} \rightarrow [+1, -0]$ $E(S_{Yes}) = -\frac{1}{1}\log_2\frac{1}{1} - 0 = 0$

$S_{NO} \rightarrow (+3, -1)$ $E[S_{NO}] = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2 1/4$

$= 0.8112$

$Gain(S, Exposure) = 0.7219 - \frac{1}{5}(0) - \frac{4}{5}(0.8112) = 0.07294$

Attribute Cough:-

$S_{Yes} \rightarrow [+4, 0]$ $E(S_{Yes}) = -\frac{4}{4}\log_2\frac{4}{4} - 0 = 0$

$S_{NO} \rightarrow [0, -1]$ $E(S_{NO}) = 0 - \frac{1}{1}\log_2\frac{1}{1} = 0$

$\therefore$ Gain (s, lough) = 0.7219 − 0 − 0 = 0.7219

Attribute: weight loss:

S_{Yes} [+2, −0] $\therefore E(S_{Yes}) = -\frac{2}{2}\log_2 \frac{2}{2} - 0 = 0$

S_{NO} [+2, −1] $\therefore E(S_{No}) = -\frac{2}{3}\log_2 \frac{2}{3} - \frac{1}{3}\log_2 \frac{1}{3}$

$= 0.9182$

Gain (s, Weight) $= 0.7219 - \frac{2}{5}(0) - \frac{3}{5}(0.9182) = 0.1709$

⇒ ~~Max Gain = Gain (S, weight) = 0.1709~~

~~from all three.~~

In NO:-

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

Attribute Exposure:

S_{Yes} [+1, −0] $E(S_{Yes}) = 0$

S_{NO} [+0, −4] $E(S_{No}) = 0$

Gain (S, Radon Exposure) = 0.7219

Attribute Chronic:

S_{Yes} [+0, −3] $E(S_{Yes}) = 0$

S_{NO} [+1, −1] $E(S_{No}) = -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2}$

$= 1$

$$Gain(S, Chronic) = 0.7219 - 0 - \frac{2}{5}(1) = 0.3219$$

Attribute :- Weight loss:-

S_{yes} [+1, -2] $\quad E_y(S_{yes}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$= 0.9182$

S_{NO} [+0, -2] $\quad E(S_{NO}) = 0$

$$Gain(S, weight) = 0.7219 - \frac{3}{5}(0.9182) - 0$$

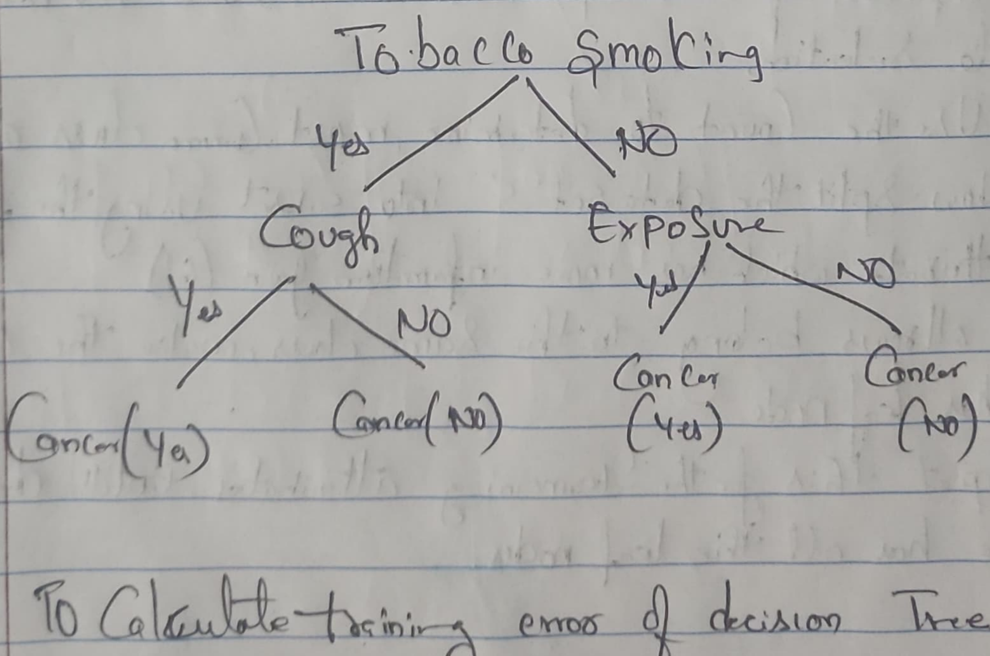
$$= 0.1709$$

Here: $Gain(S, Random) = 0.7219$

$Gain(S, Chronic) = 0.3219$ $\quad$ ($\because$ Max. Value of Gain)

$Gain(S\ weight) = 0.1709$

$\therefore$ The tree will be



(b) To Calculate training error of decision Tree, we are Considering the Confusion matrix

	NO Cancer	Cancer
NO Cancer	5	0
lung Cancer	0	5

NO Cancel   lung Cancre

$$Accuracy = TP+TN / TP + TN + FP + FN$$

$$\cancel{= 5+5/0 + 0+5+0}$$

$$\cancel{= 5+5 = 10}$$

$$= \frac{5+5}{5+0+5+0}$$

$$= \frac{10}{10} = 1$$

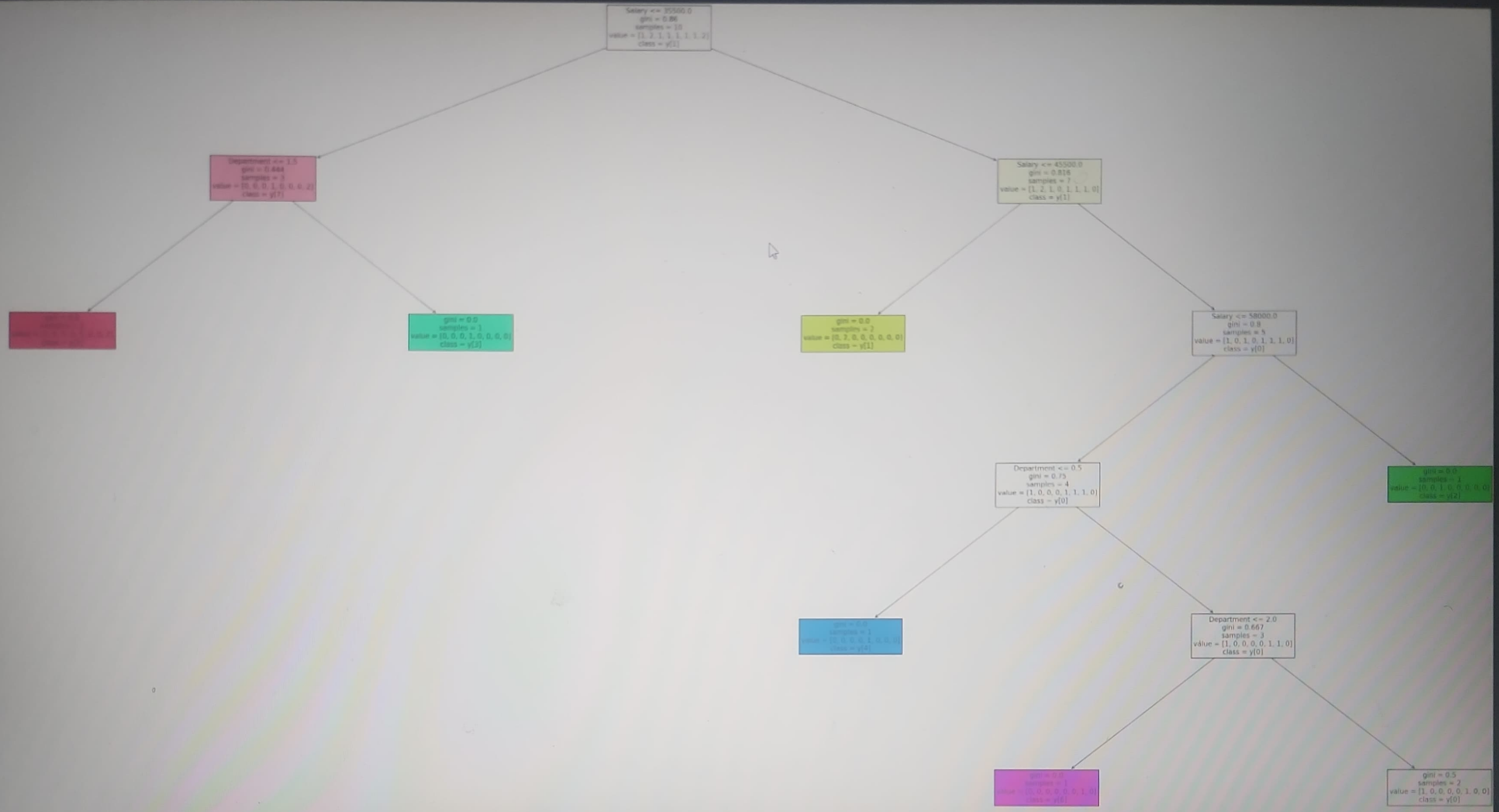
Training error =

$$\frac{FP + FN}{TP+TN+FP+FN} = \frac{0+0}{0+5+0+5} = \frac{0}{10}$$

$$= 0$$

2 a) The basic decision tree algorithm should be modified as follows to taken into Consideration the Count of each Generalised data tuple.

- The Count of each tuple must be included into Calculation of attribute Selection,
- Use the Count to determine, most Common class among tupels
- Now Split the data set 'S' into subset using the feature / attribute which has more information Gain (G)
- If all rows belong to the Same class, make the Current node has leaf node
- Repeat this for the Remaining attributes, Untill decision Tree has all the leaf nodes.


Salary <= 35500.0
gini = 0.86
samples = 10
value = [1, 2, 1, 1, 1, 1, 1, 2]
class = y[1]
Department <= 1.5
gini = 0.444
samples = 3
value = [0, 0, 0, 1, 0, 0, 0, 2]
class = y[7]
gini = 0.0
samples = 1
value = [0, 0, 0, 1, 0, 0, 0, 0]
class = y[3]
Salary <= 45500.0
gini = 0.816
samples = 7
value = [1, 2, 1, 0, 1, 1, 1, 0]
class = y[1]
gini = 0.0
samples = 2
value = [0, 2, 0, 0, 0, 0, 0, 0]
class = y[1]
Salary <= 58000.0
gini = 0.8
samples = 5
value = [1, 0, 1, 0, 1, 1, 1, 0]
class = y[0]
Department <= 0.5
gini = 0.75
samples = 4
value = [1, 0, 0, 0, 1, 1, 1, 0]
class = y[0]
gini = 0.0
samples = 1
value = [0, 0, 1, 0, 0, 0, 0, 0]
class = y[2]
Department <= 2.0
gini = 0.667
samples = 3
value = [1, 0, 0, 0, 0, 1, 1, 0]
class = y[0]
gini = 0.5
samples = 2
value = [1, 0, 0, 0, 0, 1, 0, 0]
class = y[0]

2 c) We first estimate prior probabilities for "status" class labels

$\therefore P(\text{Senior}) = 5/11$ & $P(\text{Junior}) = 6/11$

=> Now Conditional probablities

$P(\text{department} / \text{status})$

Class	Sales	Systems	Marketing	Secretary
Senior	1/5	2/5	1/5	1/5
Junior	1/3	1/3	1/6	1/6

$P(\text{age} / \text{status})$

Class	21-25	26-30	31-35	36-40	41-45	46-50
Senior	0	0	2/5	1/5	1/5	1/5
Junior	1/6	1/2	1/3	0	0	0

$P(\text{Salary} / \text{status})$

Class	26k-30k	31-35k	36-40k	41-45k	46-50k	66k-70k
Senior	0	0	1/5	0	2/5	2/5
Junior	1/3	0	0	1/6	1/3	0

$\therefore$ So far test instance

$$V_{NB} = \arg\max \; P(v_j) \prod_i P(a_i / v_j)$$

v_j (Yes, No)

$\therefore$ (system, "26 - 30", "46 - 50k")

Two labels → Senior, → Junior

$$P(\text{Senior}/a) = P(\text{Senior}) * P(\text{system}/\text{senior}) * P(26\text{-}30/\text{senior}) * P(46k\text{-}50k/\text{senior})$$

$$= \frac{5}{11} \times \frac{2}{5} \times 0 \times \frac{2}{5}$$

$$= 0$$

$$\boxed{P\left(\frac{\text{senior}}{a}\right) = 0}$$

$$P(\text{Junior}/a) = P(\text{Junior}) \times P(\text{system}/\text{Junior}) \times P(26\text{-}30/\text{Junior}) \times P(46k\text{-}50k/\text{Junior})$$

$$= \frac{6}{11} \times \frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{33}$$

$$\boxed{\therefore P(\text{Junior}/a) = 0.030}$$

∴ By Comparing the two values & the max value is Junior.

Hence, the label for this instance (person) is Junior

$$\Rightarrow \quad P(\text{Senior}|a) < P(\text{Junior}|a)$$