Data Mining HW-1

SaiParthish Mandumula
(1002022847)

1. Let Anthony be represented as 'a'

Brutus as 'b'

Caeser as 'c1'

Calpurnia as 'c2'

Cleopatra as 'c3'

∴ log weighted tf → $\begin{cases} 1 + \log_{10} tf, & tf > 0 \\ 0, & \text{Otherwise} \end{cases}$

* log weighted

	Anthony & Cleopatra	Julius Caesar	The tempest	Hamlet
a	3.195	2.80	2.69	3
b	1.60	3.19	2	1.30
c1	3.36	1.30	3.35	3
c2	0	2	0	0
c3	2.75	0	0	2.77

* log weighted IDF → $\log_{10}\left(\frac{N}{df}\right)$

	IDF
a	0.30
b	1.30
c1	0.22
c2	1.30
c3	1.69

Tf - IDF vector representation

Anthony & Cleopatra (A)

$\rightarrow 0.977a + 2.07b + 0.73c_1 + 0c_2 + 4.64c_3$

Julius Caeser (J) $\Rightarrow 0.86a + 4.15b + 0.29c_1 + 2.6c_2 + 0c_3$

The tempest (T) $\rightarrow 0.81a + 2.6b + 0.74c_1 + 0c_2 + 0c_3$

Hamlet (H) $\rightarrow 0.9a + 1.69b + 0.66c_1 + 0c_2 + 3.84c_3$

$||A|| = 5.236$

$||J|| = 4.921$

$||T|| = 2.82$

$||H|| = 4.341$

$\therefore$ Normalized Tf - IDF vector representation.

$A \rightarrow 0.12a + 0.46 + 0.14c + 0c_2 + 0.89c_3$

$J \rightarrow 0.17a + 0.83b + 0.06c_1 + 0.52c_2 + 0c_3$

$T \rightarrow 0.29a + 0.92b + 0.26c_1 + 0c_2 + 0c_3$

$H \rightarrow 0.21a + 0.39b + 0.15c_1 + 0c_2 + 0.80c_3$

* Cosine Similarities b/w 'Anthony & Cleopatra and other

$A - J = 0.371$

$A - T = 0.4566$

$A - H = 0.998$

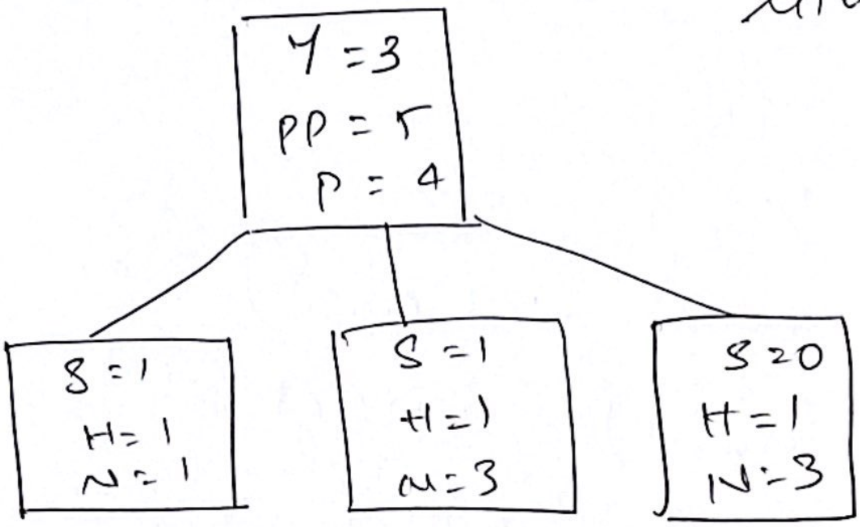
$\Rightarrow$ Anthony & Cleopatra' is most similar to 'Hamlet'.

Sai Parthish . M

② ① Age as the root

$$Gini(Parent) = 1 - \left[\frac{2^2+3^2+7^2}{12^2}\right]$$

$= 0.569$



$$GINI(Y) = 1 - \left[\frac{1^2+1^2+1}{3^2}\right] = 0.67$$

$$GINI(PP) = 1 - \left[\frac{1^2+1^2+3^2}{5^2}\right] = 0.56$$

$$GINI(P) = 1 - \left[\frac{1^2+3^2}{4^2}\right] = 0.375$$

Child split $= 3/12\,(0.67) + 5/12\,(0.56) + 4/12\,(0.375)$

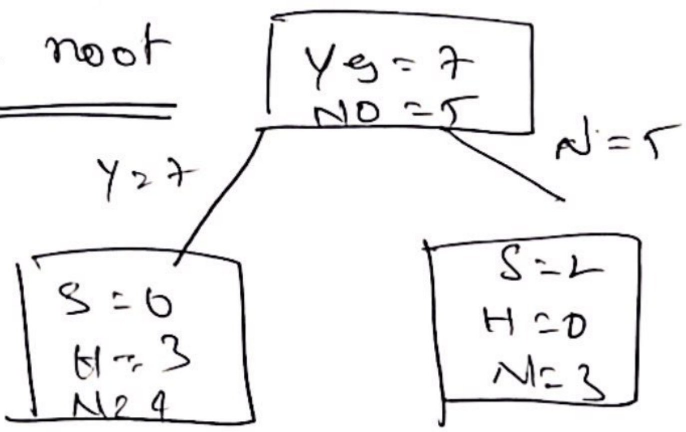
$= 0.525$

Gain = GINI(Parent) - Gini(Child)

$= 0.569 - 0.525$

$= 0.44$

② Astigmation as the root



$$GINI(Y) = 1 - \left[\frac{3^2 + 4^2}{7^2}\right] = 0.489$$

$$GINI(N) = 1 - \left[\frac{2^2 + 3^2}{5^2}\right] = 0.48$$

$$\text{Child Split} = 7/12\,(0.489) + \frac{5}{12}\,(0.48) = 0.48525$$

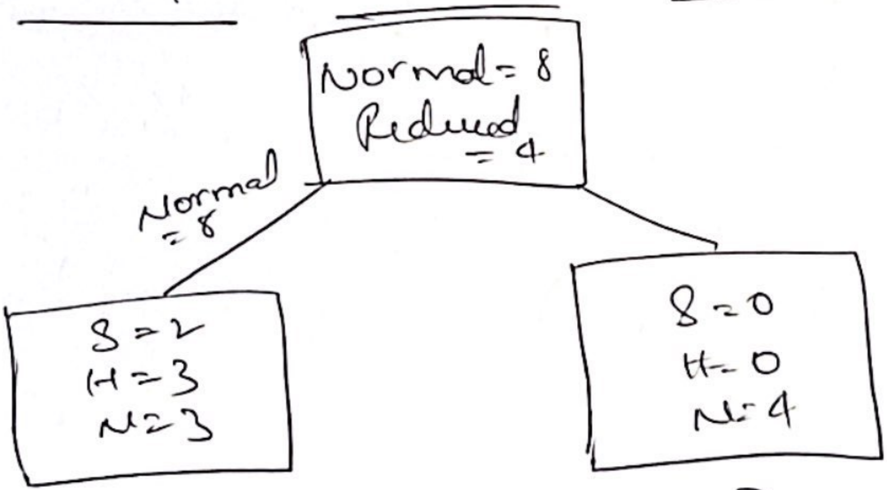
$$\text{Gain} = GINI(\text{Parent}) - GINI(\text{Child})$$

$$= 0.569 - 0.48525$$

$$= 0.08375$$

$$\begin{array}{r} 0.569 \\ 0.48525 \\ \hline .08375 \end{array}$$

③ Tear Production Rate as the root



$$GINI(N) = 1 - \left[\frac{2^2 + 3^2 + 3^2}{8^2}\right] = 0.656$$

$$GINI(R) = 1 - \left[\frac{4^2}{4^2}\right] = 0$$

$$\text{Child Split} = 8/12\,[0.656] + \frac{4}{12}\,[0] = 0.437$$

Sri. Parthish. Mandlemula

$\therefore$ Gain = GINI(Parent) - GINI(child)

$= 0.569 - 0.437$

$= 0.132$

Since we get the highest gain with the age as root, we will proceed with 'Age' as the root node.

(3) Test Instance attribute

->

Outlook	Temperature	Humidity	windy
Rainy	Cool	High	True.

P(play Golf = Y/A) = P(A/Play Golf = Yes) . P(Play Golf = Yes)

P(Play Golf = No/A) = P(A/play Golf = NO) . P(playGolf = NO)

we calculate the following

P(play Golf = Yes) = 9/14

P(play Golf = NO) = 5/14 -> P(outlook = Rainly/play golf = Y) = 2/9

P(Temperature = cool/play Golf = ~~Y~~) = $\frac{3}{9}$

P(Temperature = cool/play Golf = N) = 1/5

P(windy = True/Play Golf = N) = 3/9

$P(\text{Windy} = \text{True} / \text{Play Golf} = N) = 3/5$

$P(\text{outlook}) = \text{Rainy} / \text{Play Golf} = N) = 3/5$

$P(\text{Humidity} = \text{High} / \text{play Golf} = Y) = 3/9$

$P(\text{Humidity} = \text{High} / \text{play Golf} = N) = 4/5$

~~P(total~~

$P(\text{play Golf} = \text{Yes} / A) =$

$$\left(2/9\right)\left(3/9\right)\left(3/9\right)\left(3/9\right)\left(9/14\right) = 5.29 \times 10^{-3}$$

$P(\text{play Golf} = \text{NO} / A) =$

$$\left(3/5\right)\left(1/5\right)\left(4/5\right)\left(3/5\right)\left(5/14\right) = 20 \times 10^{-3}$$

Test . attributes . class $\rightarrow$ play golf - No

4) a Keeping the Samples in mind,

Support Vector (+ve class) $\rightarrow (0,2), (2,0)$

Support Vector (-ve class) $\rightarrow (-2,-2)$

For the . +ve hyperplane ; $\vec{w} \cdot \vec{x} + b = 1$

For the -ve hyperplane , $\vec{w} \cdot \vec{x} + b = -1$

$\rightarrow$ $0w_1 + 2w_2 + b = 1$ —— ①

$2w_1 + 0w_2 + b = 1$ —— ②

Scanned with CamScanner

Sai Parthish . M
L002022847

$-2w_1 - 2w_2 + b = -1$

$\Rightarrow 2w_1 + 2w_2 - b = 1$ —— 3

By 1 in 3, and 2 in 3

$(1-b) + (1-b) - b = 1$

$2 - 3b = 1$

$\boxed{b = 1/3}$ —— 4

with 4 in 1 and 2, we get

$$\boxed{\begin{aligned} w_1 &= 1/3 \\ w_2 &= 1/3 \end{aligned}}$$

(b) margin $= \frac{2}{\|w\|} = \frac{2}{0.471} = 4.242$

$\|w\| = \sqrt{(1/3)^2 + (1/3)^2}$
$= 0.471$

(c) $w_1x_1 + w_2x_2 + b$

$(1/3\, x - 1) + (1/3\, x - 1) + 1/3$

$= -1/3 - 1/3 + 1/3 = -1/3$

So, the more than -1, it will lie above the Hyperplane, Hence, it will be of the +ve class.