Datamining H.W:-1

SaiParthish Mandumula
(1002022847)

1. Let Anthony be represented as 'a'

Brutus as 'b'

Caeser as 'c_1'

Calpurnia as 'c_2'

Cleopetra as 'c_3'

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

* log weighted

	Anthony & Cleopetra	Julius Caesar	The tempest	Hamlet
a →	3.195	2.86	2.69	3
b →	1.60	3.19	2	1.30
c_1 →	3.36	1.30	3.35	3
c_2 →	0	2	0	0
c_3 →	2.75	0	0	2.77

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

	IDF
a	0.30
b	1.30
c_1	0.22
c_2	1.30
c_3	1.69

Tf - IDF vector representation

Anthony & cleopetra (A)

$\rightarrow 0.977a + 2.02b + 0.73c_1 + 0c_2 + 4.64c_3$

Julius Caeser (J) $\Rightarrow 0.86a + 4.15b + 0.29c_1 + 26c_2 + 0c_3$

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

Hamlet (H) $\rightarrow 0.9a + 1.69b + 0.66c_1 + 0c_2 + 4.84c_3$

$||A|| = 5.236$

$||J|| = 4.921$

$||T|| = 2.82$

$||H|| = 4.341$

$\therefore$ Normalized Tf - IDF vector representation.

$0.12a + 0.46 + 0.14c + 0c_2 + 0.89c_3$

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

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

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

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

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

$A - J = 0.371$

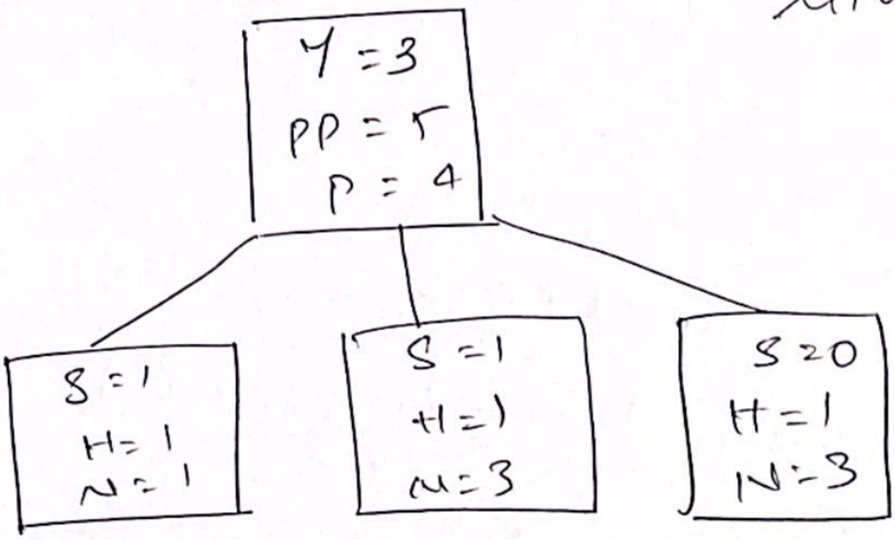
$A - T = 0.4566$

$A - H = 0.998$

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

Sai Prathish . M

② ① Age as the root



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

$$= \underline{0.569}$$

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

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

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

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

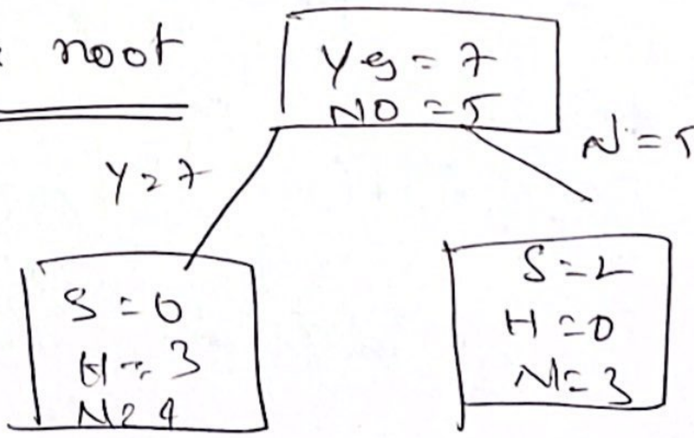
$$= 0.525$$

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

$$= 0.569 - 0.525$$

$$= \underline{0.044}$$

② Astigmation as the root



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

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

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

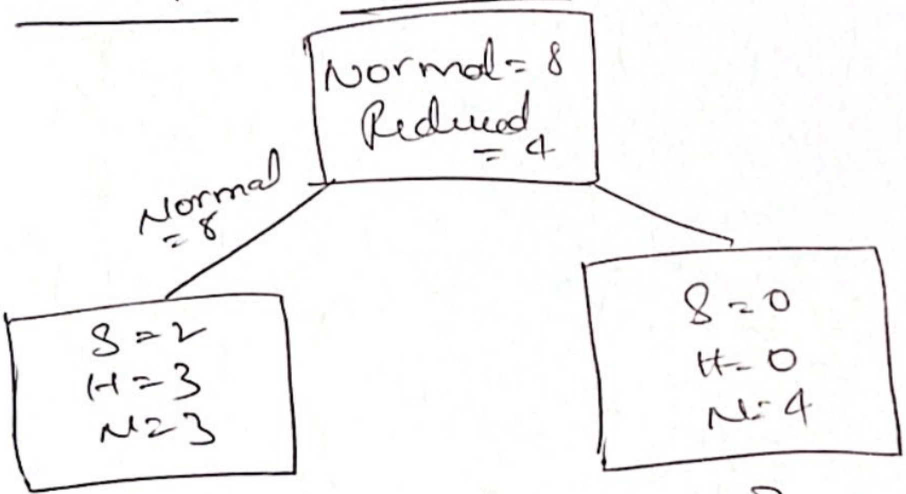
$$\text{Gain} = GINI(\text{Parent}) - GINI(\text{Child})$$

$$= 0.569 - 0.48525$$

$$= 0.08375$$

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

(3) Tear Production Rate as the root



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

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

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

Scanned with CamScanner

Sri · Parthish · Mandeemula

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

$= 0.569 - 0.437$

$= 0.132$

Since we get the highest gain with the ~~age~~ Tear Production as root, we will proceed with ~~Age~~ Tear production 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 = ~~N~~Y) = $\frac{3}{9}$

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

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

$P($~~windy~~ windy $= True / Play\ Golf = N) = 3/5$

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

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

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

~~P(windy~~

$P(play\ Golf = Yes / A) =$

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

$P(play\ Golf = No / A) =$

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

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
1002022847

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

$\Rightarrow 2w_1 + 2w_2 - b = 1$ —— 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

$\boxed{w_1 = 1/3 \quad w_2 = 1/3}$

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

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

ⓒ $w_1x_1 + w_2x_2 + b$

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

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

So, the more than 0, it will lie ~~above~~ below the Hyperplane, Hence, it will be of the ~~pos class~~ negative class.