06/11/22.

Mathematical Intuition of DT :-

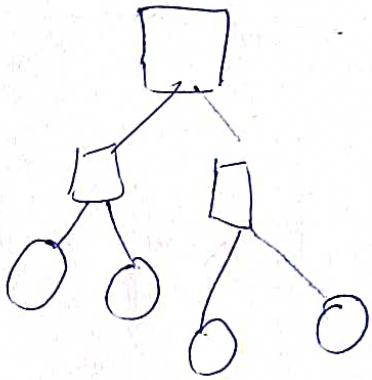
- Decision tree Classifier :- DT (regression) → O/p → f1 (cat, num)
- Decision tree regressor → O/p - (cat, num)

Classification And regression trees

- IDB

Iterative Dicotomiser

- entropy
- gini impurity

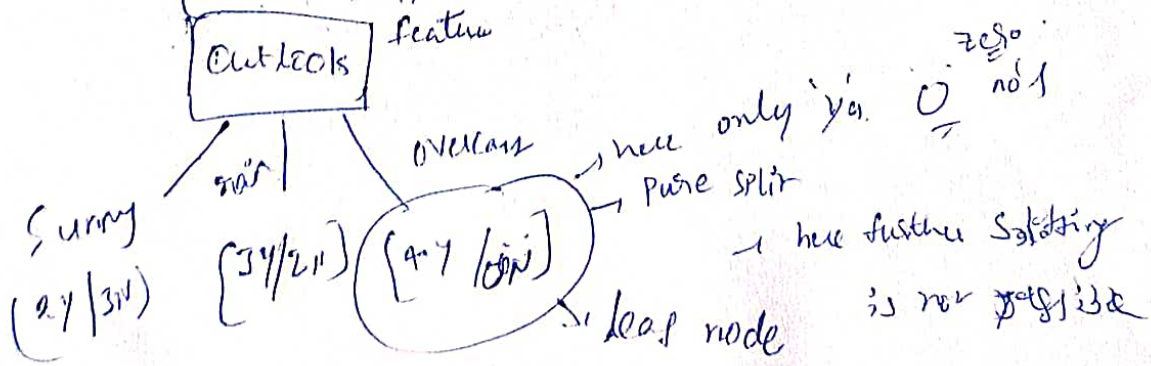


Day	f1 Outlook	f2 temperature	f3 humidity	f4 wind	→ O/p Decision
	Sunny	hot	high	weak	No
	Sunny	hot	high	Strong	No
	rain	hot	hi	weak	yes
	overcast	hot		weak	yes
	Sunny	mild		weak	yes
	rainfall	Cool		Strong	No
				weak	yes
				weak	yes

[9Y/5N] O/p NO

Outlook feature

Outlook → feature



Pure split :- only one parameter ; here further splitting is not possible

leaf node :- further splitting is not possible here.

→ to check the purity of the feature, 2 ways are there

Purity → 1. Entropy

2. Gini - coefficient, gini impurity

Entropy :- $H(S) = -\sum_{i=1}^{n} P_i \times \log_2(P_i)$

Gini - Coefficient → $1 - \sum_{i=1}^{n} P_i^2$

for binary $1 - [P_Y^2 + P_N^2]$

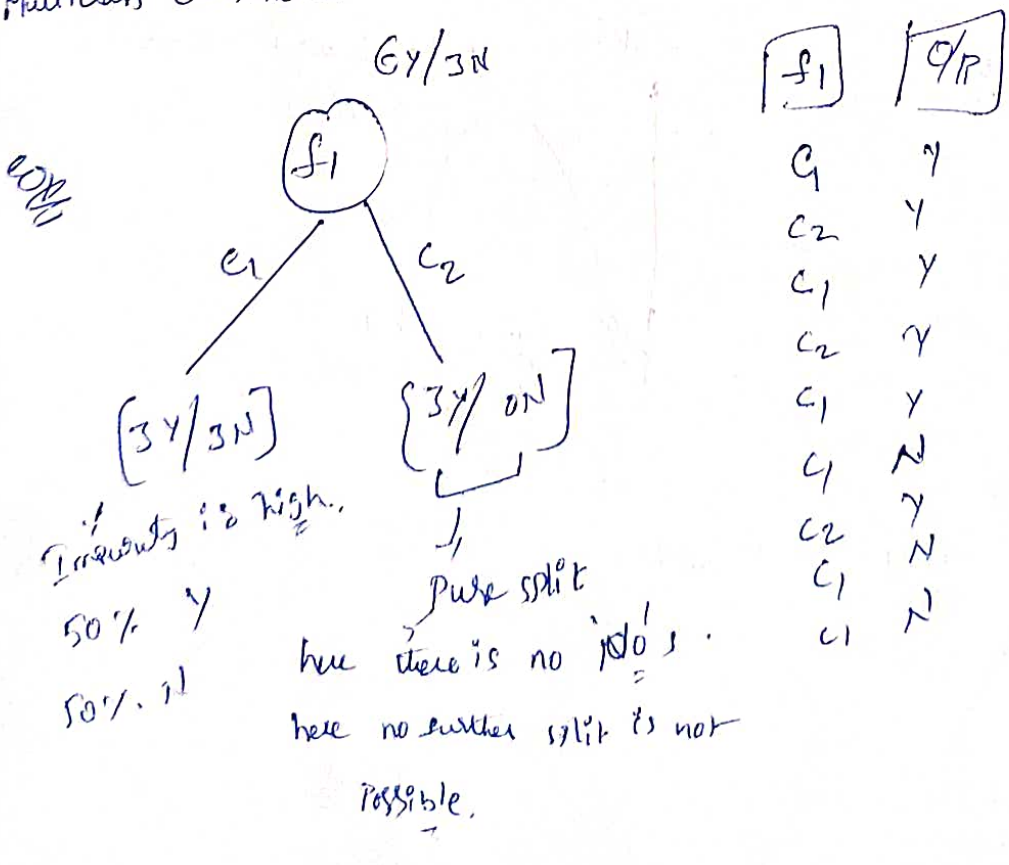
2 class → Y/N.

Binary classification → $-P_Y \log(P_Y) - P_N \log(P_N)$

for 3 class - C_1, C_2, C_3

$-P_{C_1} \log(C_1) - P_{C_2} \log(C_2) - P_{C_3} \log(C_3)$

Multiclass classification



f1	O/P
C_1	Y
C_2	Y
C_1	Y
C_2	Y
C_1	Y
C_1	N
C_2	Y
C_2	N
C_1	N
C_1	N

Entropy $H(S) = -\sum_{i=1}^{n} P_i \log(P_i)$

$= -P_Y \log_2(P_Y) - P_N \log_2(P_N)$

from (3Y/3N)

$= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$

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

$= -\frac{1}{2}\left[\log_2(1) - \log_2(2)\right] - \frac{1}{2}\left[\log_2(1) - \log_2(2)\right]$

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

$= \frac{1}{2} + \frac{1}{2} = 1.$

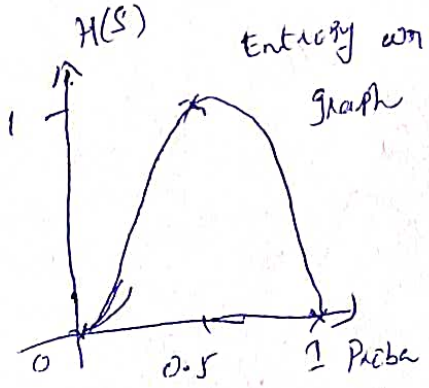
Entropy $H(S) = 1$ ⇒ It is for impurity data = 1.

whenever there is high impurity

Eg 3Y/0N ⇒ for pure split entropy = 0. , entropy value is 1:

$-\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)$

Entropy = 0



Entropy with Probability graph

(50%)

$H(S) = 1$ ⇒ Very impure split

$H(S) = 0$ ⇒ Pure split

Eg. 2 Yes / 3 No.

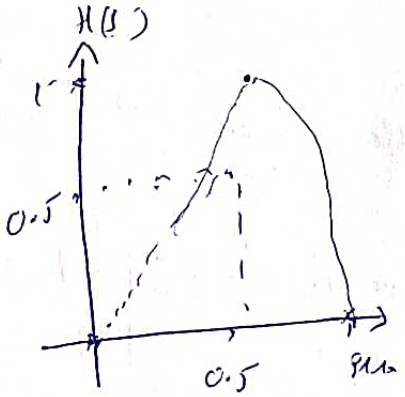
$$H(S) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= 0.97$$

(2) Gini coefficient / Gini impurity

$$1 - \sum_{i=1}^{n} P_i^2$$

1. 3Y/3N → Entropy → H(S) = 1 (very impure split)
2. 3Y/0N =>
3. 2Y/3N =>



Gi:: $= 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$

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

$= 1 - \left[\frac{1}{4} + \frac{1}{4}\right]$

$= 1 - \left[\frac{2}{4}\right] \Rightarrow 1 - \frac{1}{2} = \frac{1}{2}$

$= 0.5$

Gi:: 4Y/8N

$= 1 - \left\{\left(\frac{4}{12}\right)^2 + \left(\frac{8}{12}\right)^2\right\}$

$1 - \left\{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right\}$

$1 - \left\{\frac{1}{9} + \frac{4}{9}\right\}$

$= 1 - \left[\frac{5}{9}\right] = \frac{9-5}{9} = \frac{4}{9}$

$= 0.44$

Gi:: 8Y/2N

$\Rightarrow 1 - \left\{\left(\frac{8}{10}\right)^2 + \left(\frac{2}{10}\right)^2\right\}$

$1 - \left\{\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right\}$

$= 1 - \left\{\frac{16}{25} + \frac{1}{25}\right\}$

$= 1 - \frac{17}{25} = \frac{25-17}{25}$

$= 0.32$

3Y/0N

$1 - \left\{\left(\frac{3}{3}\right)^2 + \left(\frac{0}{10}\right)^2\right\}$

$= 1 - 1$

$= 0$

Information Gain :- ID3 approach → entropy

CART → Gini impurity → Binary classification

feature 1, feature 2, feature 3

→ Priority

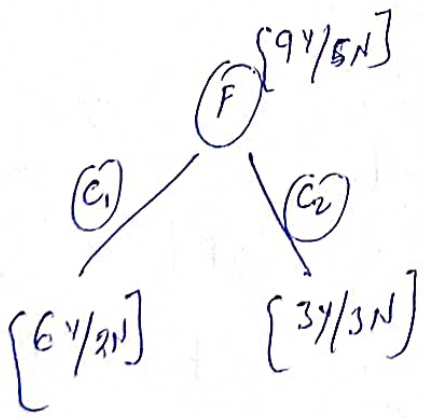
ID3 approach { 1. entropy, Information } Gini impurity, Info gain } CART → binary split

Y / N

Gini impurity is very fast, when compare with entropy mathematically.

IG :- Information Gain :- → we can cal using entropy & Gini impurity

$$Gain(S, f_1) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$$ → w.r.t to entropy



V = value.

feature f1 = (9Y/5N)

$H(S)$ = root entropy

$= -P_Y \log_2 P_Y - P_N \log_2 P_N$

$= -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right)$

$= -0.64 \log(0.64) - 0.35 \log_2(0.35)$

$H(S) = 0.94$

root feature entropy

0.4 (0.19) - 0.35 · 0.96

- 0.1216 - 0.033

$c_1 = 6Y/2N \Rightarrow$

$c_1 = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$

$= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$

$0.75 \cdot (0.12) - 0.25 \cdot 0.60$

$= 0.81$ $\quad$ 0.13

$=$

$c_2 = 3Y/3N = 1.$

$H(S) = 1.$

$(6+2) \rightarrow$ first split $\frac{8}{14} \rightarrow$ total value

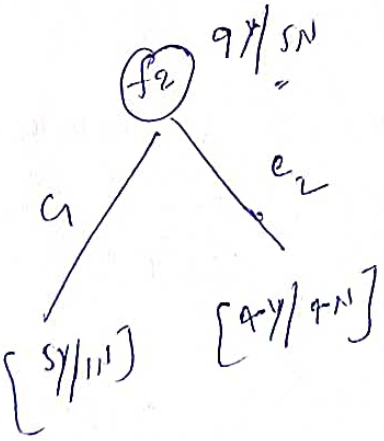
2nd split $(3+3)$.

$$Gain(S, f_1) = 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

$$= 0.94 - [0.462 + 0.42]$$

$$f_1 = 0.049.$$

$f_2 =$



$IG_1 = ?$

$$Gain = 0.94 - \left(\frac{6}{14}\right)(0.65) - \frac{8}{14} \times 1$$

$=$

$f_2 \quad H(S) = 0.94$

$Gain(f_2) = 0.04$

$c_1 = 5Y/1N$

$= -\frac{5}{6}\log_2\left(\frac{5}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right)$

$IG(f_2) > IG(f_1)$

This is greater and it is providing more info.

$\therefore 0.65.$

$c_2 = 4Y/4N = -\left\{\frac{4}{8}\log\frac{4}{8} + \frac{4}{8}\log\frac{4}{8}\right\}$

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

$= 1.$

Day	Outlook	temperature	humidity	wind	Decision
	Sunny	hot	high	weak	No
Sp	Sunny	hot	high	Strong	No
	Overcast	hot	high	weak	YES
	rainfall	mild	high	weak	YES
	Rainfall	cool	normal	weak	YES
	rainfall	cool	normal	Strong	No
	Overcast	cool	normal	Strong	YES
	Sunny	mild	high	weak	No
	Sunny	cool	normal	weak	YES
	rainfall	mild	normal	weak	YES
	Sunny	mild	normal	Strong	YES
	Overcast	mild	high	Strong	YES
	Overcast	hot	normal	weak	YES
	rainfall	mild	high	Strong	No

(S) [9Y/5N]

Outlook [top feature]

Sunny (2Y/3N) | over (4Y/0N) | rain (3Y/2N)

w, t, h

w

t, h | t, h

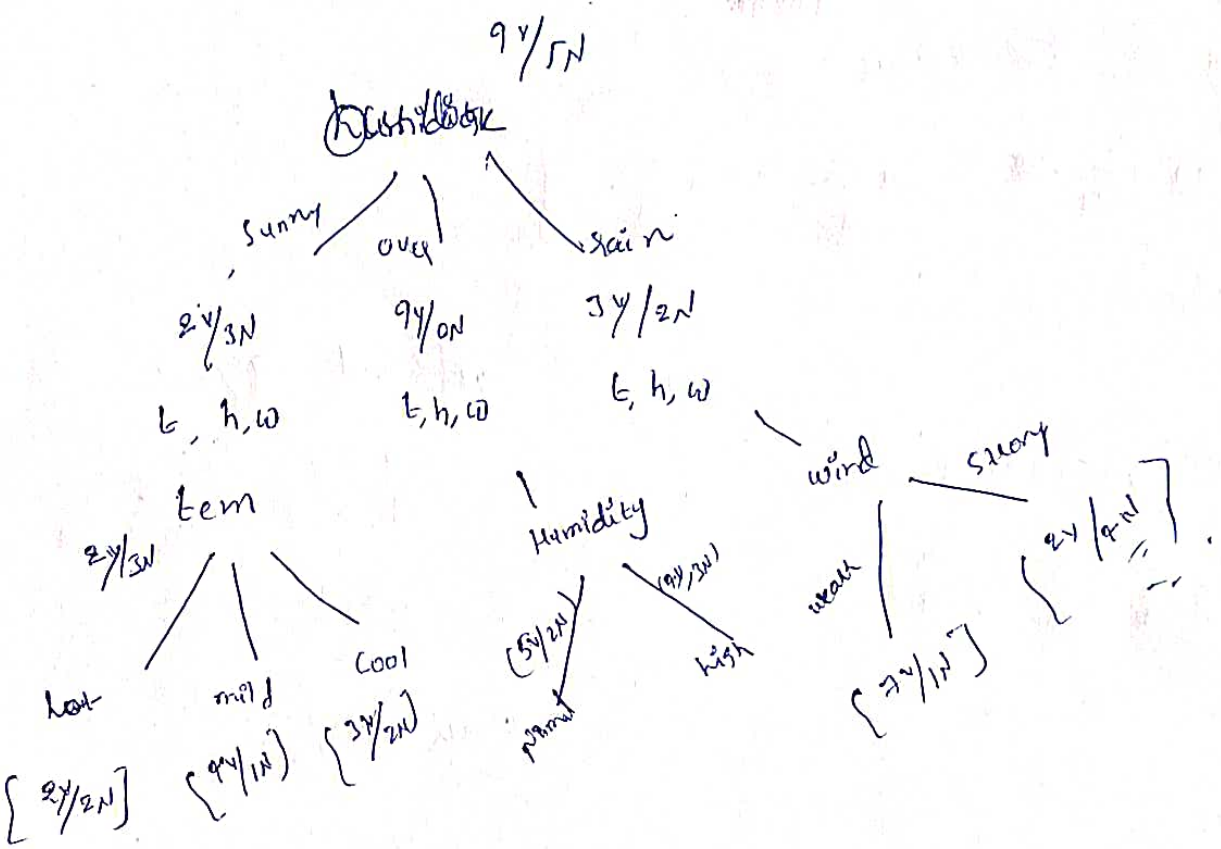
Outlook

(T, H, W) | (T, H, W)

Outlook

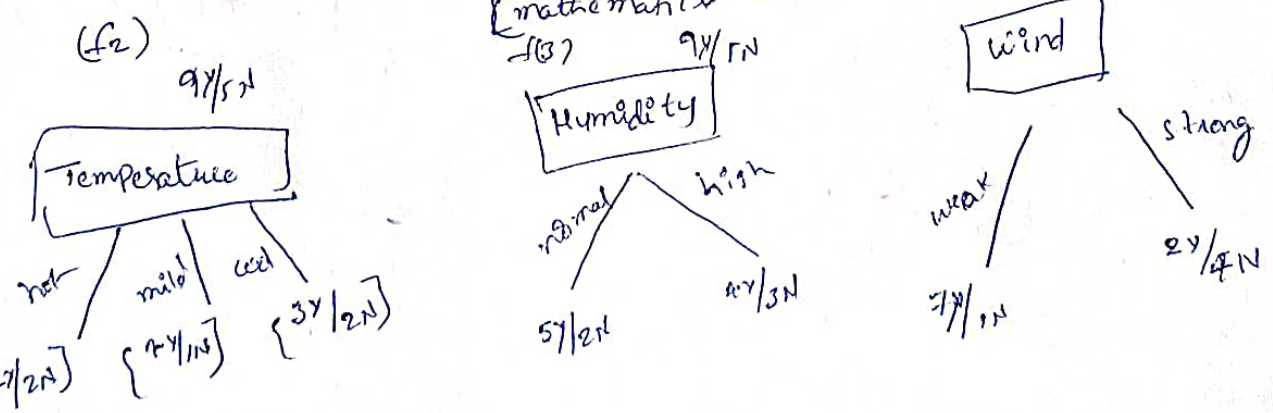
temp | h

[H, W] [H, W]



- Entropy vs gini-impurity [5 dif]
- Take this dataset [tennis play] build DT from scratch [mathematical representation]

Eg1)



f2 ⇒ Temperature

9Y/5N =

$H(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$ (0.1233, 0.1576)

$H(S) = 0.94$

Cool (3Y/2N)

$= -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right)$ =

Hot: 2Y/2N.

$H(S_v) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\frac{2}{4}$

$H(S_v) = 1$

Mild ⇒ 4Y/1N

$H(S_v) = -\frac{4}{5}\log\frac{4}{5} - \frac{1}{4}\log\frac{1}{5}$

=

IG f1 f2 (Temperature)

$= 0.94 \left[\frac{4}{14} \log_2 \quad + \frac{5}{14} \cdot \quad + \frac{5}{14} \cdot \right]$

f_3 (Humidity) = 9Y/5N

$-\frac{9}{14} \log_2(\frac{9}{14}) - \frac{5}{14} \log_2(\frac{5}{14})$

=

Normal :- 8Y/2N

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

=

High :- 4Y/3N

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

=

Gain (f3) = $0.94 \left[\frac{7}{14} \cdot \quad + \frac{7}{14} \cdot \right]$

f_4(wind) = 9Y/5N

$-\frac{9}{14} \log_2(\frac{9}{14}) - \frac{5}{14} \log_2(\frac{5}{14})$

= 0.94

Weak : 7Y/1N

$= -\frac{7}{8} \log_2(\frac{7}{8}) - \frac{1}{8} \log_2(\frac{1}{8})$

=

Strong :- Strong :- 2Y/4N

$-\frac{2}{6} \log_2(\frac{2}{6}) - \frac{4}{6} \log_2(\frac{4}{6})$

=

IG (f4) :- $0.94 \left[\frac{7}{14} \cdot \quad + \frac{1}{14} \cdot \right]$