(*) Mathematical Intuition of Decision Tree (*)

→ Decision Tree classifier :- Consider the following dataset :-

Day	outlook (f_1)	temperature (f_2)	humidity (f_3)	wind (f_4)	Decision (o/p)
1.	sunny	hot	high	weak	No
2.	sunny	hot	high	strong	No
3.	overcast	hot	high	weak	Yes
4.	rainfall	mild	high	weak	Yes
5.	rainfall	cool	normal	weak	Yes.
6.	rainfall	cool	normal	strong	No
7.	overcast	cool	normal	strong	Yes
8.	sunny	mild	high	weak	No
9.	sunny	cool	normal	weak	Yes.
10.	rainfall	mild	normal	weak	Yes.
11.	sunny	mild	normal	strong	Yes.
12.	overcast	mild	high	strong	Yes
13.	overcast	hot	normal	weak	Yes.
14.	rainfall	mild	high	strong	No

f_1, f_2, f_3, f_4 are independent features.

o/p → is our dependent feature.

Our independent features can be both numerical or categorical.

For Decision Tree classifier, our dependent feature (or o/p feature) will be categorical.

Let's talk about 2 Decision Tree classifiers:-

① ID_3

② CART → Classification & Regression Trees

ID_3 → Iterative Dichotomiser

ID_3

→ Iterative Dichotomiser

→ are talk about entropy

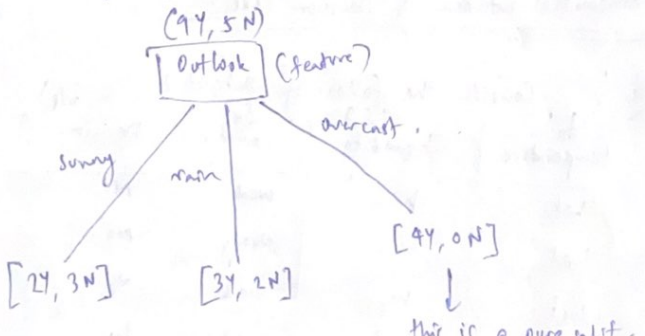
CART

→ Classification & Regression Trees

→ are talk about Gini impurity

Let's consider outlook feature as our root node.

Now in our target feature, we have 9 yes & 5 no.



this is a pure split.

We have only yes values & no no values.

This is a leaf node.

Our objective is to get the purity of the feature.

To get the purity of our features, we have 2 methods :-

① Entropy

② Gini impurity / Gini co-efficient

Entropy :-

$$\text{Entropy} = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i)$$

Gini - coeff :-

$$\text{Gini coeff} = 1 - \sum_{i=1}^{n} p_i^2$$

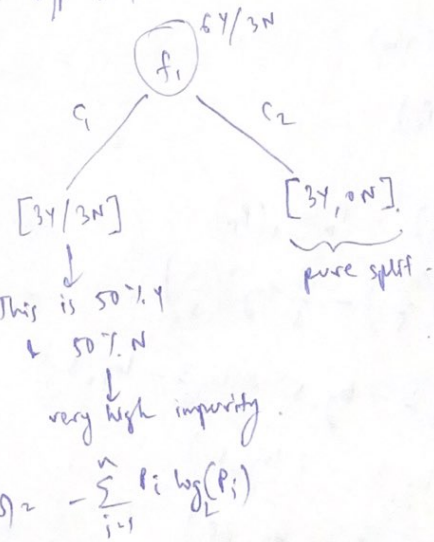
In our dataset, we have 2 classes → Yes & No. → binary classification.

$\therefore$ entropy = ~~$-p_Y \log(p_Y) - p_N \log$~~

$$= -p_Y \cdot \log_2(p_Y) - p_N \log_2(p_N)$$

$\therefore$ Gini impurity $= 1 - [p_Y^2 + p_N^2]$

Say we have a binary classification & based on one feature f_1 which has 2 categories & o/p = Y or N.



entropy = $H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$

$$= -p_Y \log_2(p_Y) - p_N \log_2(p_N)$$

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

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

$$= 1$$

So, for very impure split, we are getting entropy, $H(S) = 1$

Now let's see the entropy of the pure split.

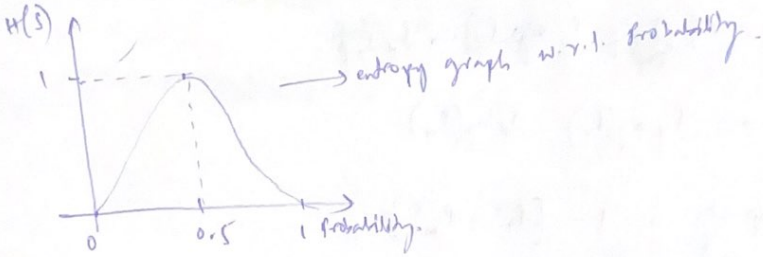
$$H(S) = -\sum_{i=1}^{n} P_i \log_2(P_i)$$

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

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

$$= 0.$$

∴ for pure split $H(S) = 0$



→ entropy graph w.r.t. probability.

Highest value of entropy = 1. → very impure split

$H(S) = 0$ → pure split.

Let's consider the split (2Y / 3N)

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

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

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

$$= -\left(\log_2 2 - \log_2 5 + \log_2 3 - \log_2 5\right)$$

$$= -\left(1 - \log_2 3 - 2\log_2 5\right)$$

$$= -\left(1 - \log_2 3 - \log_2 25\right)$$

$$= 0.97.$$

ⓧ Gini Coefficient or Gini impurity ⓧ

$$\text{Gini coeff} = 1 - \sum_{i=1}^{n} (P_i)^2$$

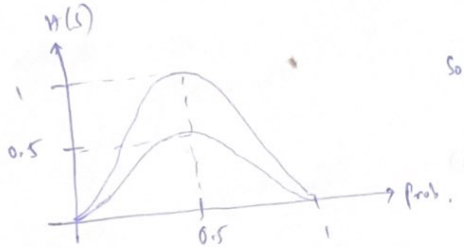
Now let's consider the three splits :-

① 3Y / 3N
② 3Y / 0N.
③ 2Y / 3N.

① 3Y / 3N

$$\text{Gini coeff} = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$$

$$= 1 - [0.25 + 0.25] = 0.5$$



So, range of H(s) → [0, 1]
range of Gini coeff → [0, 0.5]

② 3Y / 0N

$$\text{Gini coeff} = 1 - \left[\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right]$$

$$= 0.$$

100
− 68
32

③ consider → [4Y, 8N].

$$\text{Gini coeff} = 1 - \left[\left(\frac{4}{12}\right)^2 + \left(\frac{8}{12}\right)^2\right]$$

$$= 1 - \left[\frac{80}{144}\right] = 0.44$$

④ consider → [8Y / 2N]

$$\text{Gini coeff} = 1 - \left[\left(\frac{8}{10}\right)^2 + \left(\frac{2}{10}\right)^2\right]$$

$$= 1 - \left[\frac{64}{100} + \frac{4}{100}\right] = 1 - \frac{68}{100}$$

$$= \frac{100 - 68}{100} = \frac{32}{100} = 0.32$$

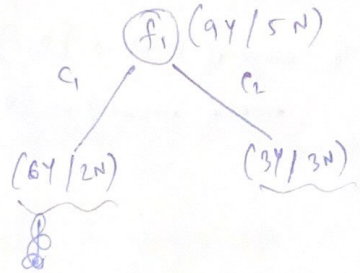
Now, say we have 3 features → feature 1, feature 2, feature 3.

Now, to see purity, if we find entropy & then the information gain ⟶ this is ID_3 approach.

If we find Gini impurity & then the information gain ⟶ CART approach.
(This is much faster & suitable for larger datasets)

$$\text{Gain}(S, f_1) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$$

Consider this scenario :-



$H(S)$ = root feature entropy.

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

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

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

$$= -\log_2 9 - \log_2 5$$

$$\approx 0.94$$

for 6Y / 2N

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

$$= -\log_2 6 - \log_2 2$$

$$= -\log_2 6 - 1$$

$$\approx 0.81$$

for 3Y / 3N

$$H(S) = 1$$

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

$$= 0.049$$

This is the ID_3 approach.

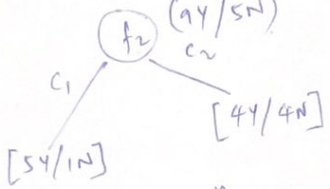
Similarly in the information gain formula, if we replace the entropy with Gini impurity → It will be the CART approach.

But, this is w.r.t. only one feature f_1.

~~Let's talk about feature~~

We have to choose the feature that has less impurity & more information.

Let's talk about feature 2 → f_2.



$$H(S) = \text{root entropy} = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

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

$$= -\log_2 9 - \log_2 5 \approx 0.94$$

$$\text{gain}_{(f_2)} = 0.94 - \left[\frac{6}{14} \times 0.65 + \frac{8}{14} \times 1\right]$$

$$= 0.09$$

Now, $\text{gain}(f_1) = 0.049$ & $\text{gain}(f_2) = 0.09$

Clearly, Inform gain of f_2 > info. gain of f_1.

So, we select f_2 as our root node. because it is greater & is providing more info.