06-NOV-2022

Decision Tree

Mathematical Intuition

1. Decision Tree classifier
2. Decision Tree Regressor

classification Problem

Dependent features

	f_1	f_2	f_3	f_4	
Dataset					↓
Day	Outlook	Temperature	Humidity	Wind	Decision
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	~~high~~ normal	Strong	yes
12	Overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainfall	mild	high	strong	NO

Independent Features

- Here f_1, f_2, f_3, f_4 are categorical features.
- Based on the output feature we decide whether the problem is about classification or Regression.

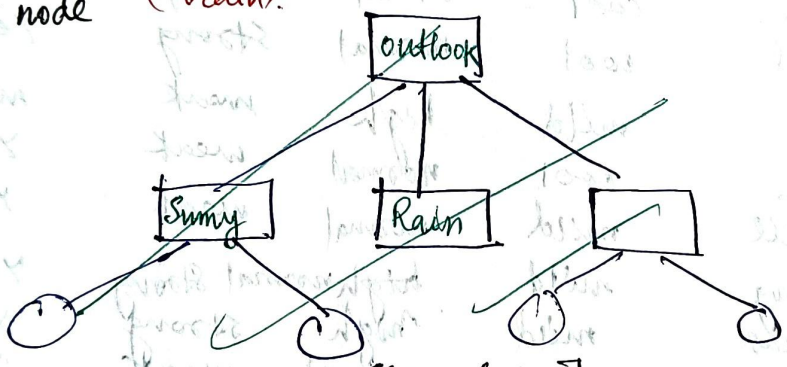
{ I/P (Cat or Num) — O/P (Cat or Num) } Decision Tree classifier

{ I/P (Cat or Num) — O/P (Numerical) } Decision Tree Regressor

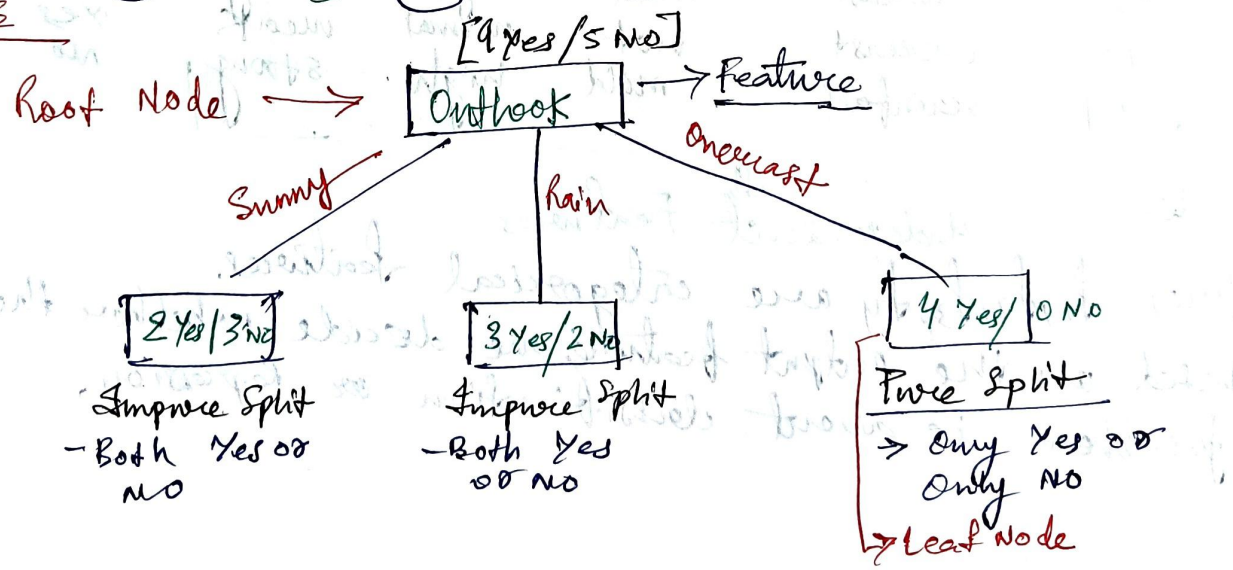
① ID3
- [Iterative Dichotomiser]
- Deal with Entropy

② CART
- [Classification and Regression Trees]
- Deal with Gini Impurity

→ Choose Outlook as parent node (We can decide based on entropy or Information Gain).



TREE



Impure Split
- Both Yes or No

Impure Split
- Both Yes or No

Pure Split
→ Only Yes or Only No
→ Leaf Node

→ Try to find out pure split.

*Check Purity using below methods

Two methods:

① Entropy

② Gini Impurity or Gini coefficients

Purity

→ How much information one feature is providing

① Entropy

- Randomness
- We can check purity of the feature using Entropy

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

② Gini Impurity

- check Purity using G.I.

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

For 2 class

$$\text{Entropy} = -P_Y \log_2(P_Y) - P_N \log_2 P(N)$$

Y → Yes } 2 class
N → NO }

→ log base 2

For 3 class → (C_1, C_2, C_3)

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

For 2 class (Yes and No)

$$G.I = 1 - [P_Y^2 + P_N^2]$$

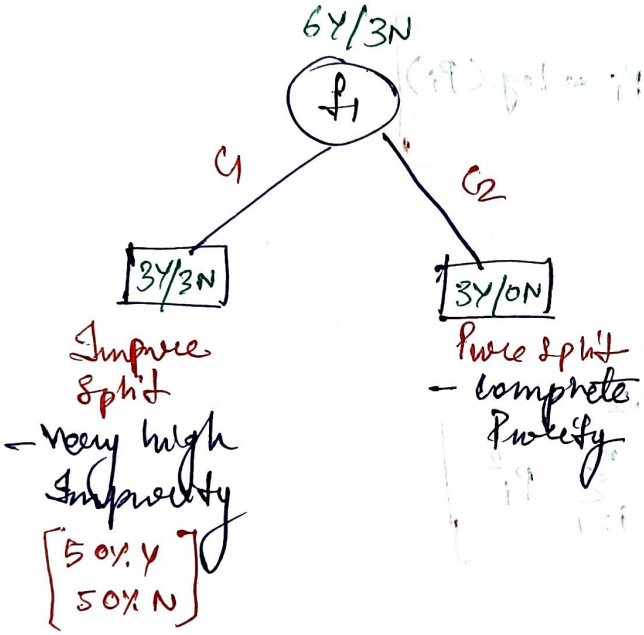
For 3 class (C1, C2, C3)

$$G.I = 1 - (P_{C_1}^2 + P_{C_2}^2 + P_{C_3}^2)$$

Q Why are we calculating the G.I and entropy?

Ans To check the impurity and randomness of the feature.

Example



Dataset

f1	O/P
C1	Y
C2	Y
C1	Y
C2	Y
C1	Y
C1	N
C2	Y
C1	N
C1	N

Checking impurity for f1

① Entropy (For 3Y/3N split) ($P_Y = 3/6$, $P_N = 3/6$)

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

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

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

$$= -\log_2\left(\frac{3}{6}\right) = -\log(2^{-1})$$

$$= 1$$

If entropy = 0 → Pure split
If entropy = 1 → very Impure split

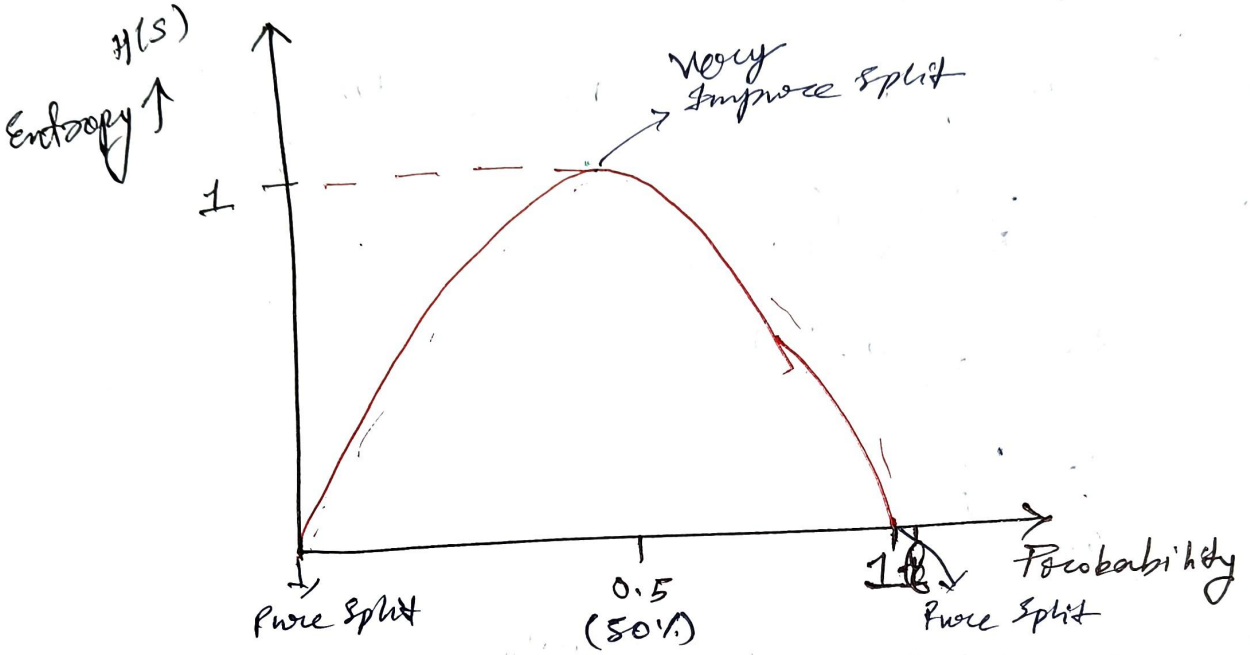
(For 3Y/0N split) [$P_Y = \frac{3}{3} = 1$, $P_N = 0$]

Entropy $(H(S)) = -P_Y \log_2(P_Y) - P_N (\log(P_N))$

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

$= 0$

Entropy graph w.r.t Probability



(For 2Y/3N) $(P_Y = \frac{2}{5}, P_N = \frac{3}{5})$

Entropy $(H(S)) = -P_Y \log_2(P_Y) - P_N \log_2(P_N)$

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

$= 0.97$

② Gini coefficients / Gini Impurity

For (3Y/3N Split) $\{P_Y = \frac{3}{6}, P_N = \frac{3}{6}\}$

$$G.I = 1 - \sum_{i=1}^{n} (P)^2 = 1 - \left\{\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right\}$$
$$= 1 - \left(\frac{1}{4} + \frac{1}{4}\right)$$
$$= \frac{1}{2} = 0.5$$

If $H(S) = 1$ (3Y/3N) How will you decide?
→ We use G.I in that case

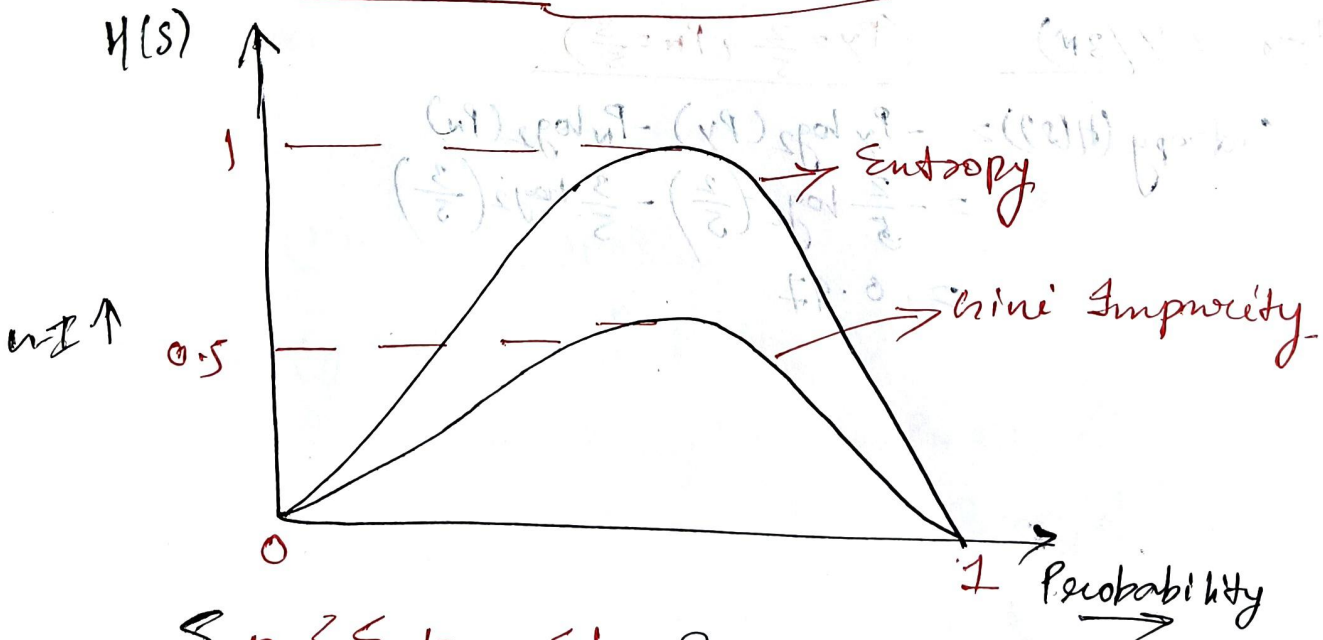
For (3Y/0N Split) $\{P_Y = \frac{3}{3}, P_N = 0\}$

$$G.I = 1 - \left\{\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right\}$$
$$= 1 - 1 = 0$$

For (2Y/3N Split) $\{P_Y = \frac{2}{5}, P_N = \frac{3}{5}\}$

$$G.I = 1 - \left\{\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)\right\}$$
$$= 1 - \left\{\frac{4}{25} + \frac{9}{25}\right\}$$
$$= 1 - \frac{13}{25}$$
$$= \frac{12}{25}$$

Entropy and G.I wrt Probability



$\{0 \leq \text{Entropy} \leq 1$
$0 \leq G.I \leq 0.5\}$

For (4Y/8N split) $\{P_Y = \frac{4}{12}, P_N = \frac{8}{12}\}$

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

$$= 1 - \left\{\frac{16}{144} + \frac{64}{144}\right\} = 1 - \frac{80}{141}$$

$$= 0.444$$

For (6Y/2N split) $\{P_Y = \frac{8}{10}, P_N = \frac{2}{10}\}$

$$G.I = 1 - \left\{\left(\frac{8}{10}\right)^2 + \left(\frac{2}{10}\right)^2\right\}$$

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

$$= 1 - \frac{68}{100}$$

$$= \frac{32}{100}$$

$$= 0.32$$

(3) Information Gain

→ Either we consider Entropy or G.I to find Information Gain.

→ In ID3 algorithm we use entropy to calculate information gain.

→ In CART algorithm we use G.I to calculate information gain.

feature1, feature2, feature3

How to select one feature for root node?

→ find Information gain using entropy or G.I.
(entropy → ID3, G.I → CART)

→ For a big dataset we use G.I.

→ Generally in CART, binary split happens.

→ Generally in ~~IB~~ ID3, more than 2 split happens.

→ Calculations are fast in this Impurities.

Information Gain — Root feature Entropy

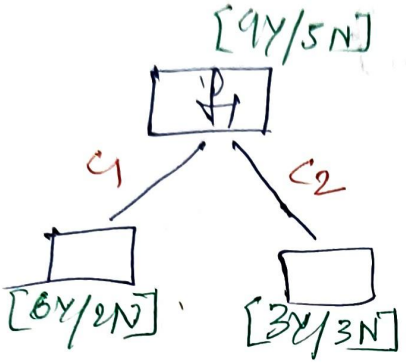
$$Gain(S, F_i) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$$ — v = value

↓

Either we can entropy or this Impurities

Sv → for each each split, entropy

Example



Q Calculate Information Gain?

Ans ~~For this split entropy have been calculated in entropy section. for 6Y/2N and 3Y/3N split.~~

$H(S)$ = Root feature Entropy $= -P_Y \log_2(P_Y) - P_N \log_2(P_N)$

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

$= 0.94$

for 6Y/2N, $H(S) = -\frac{6}{8}\log\left(\frac{6}{8}\right) - \frac{2}{8}\log\left(\frac{2}{8}\right)$

$H(S) = 0.81$

for 3Y/3N, $H(S) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$

$H(S) = 1$

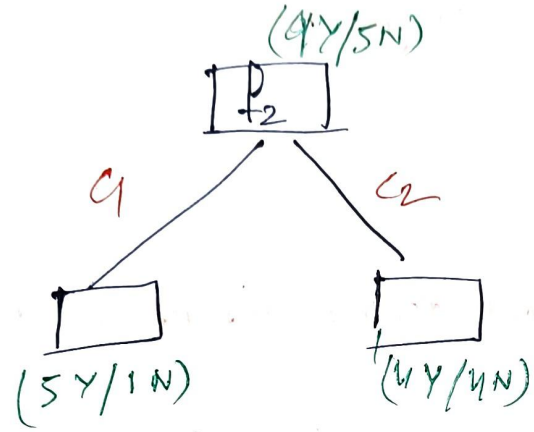
Information gain w.r.t f_1

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

$$\boxed{\text{Gain}(S, f_1) = 0.049}$$

Information gain w.r.t f_2

→ Our ultimate goal is to calculate information gain using E.I or entropy.



$$\text{Gain}(S, f_2) = 0.94 - \left(\frac{6}{14} \times 0.65 + \frac{8}{14} \times 1\right)$$

$$\boxed{\text{Gain}(S, f_2) = 0.09}$$

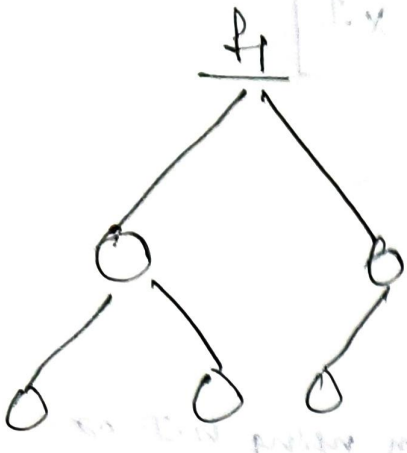
NOTE

$\text{Gain}(S, f_2) > \text{Gain}(S, f_1)$

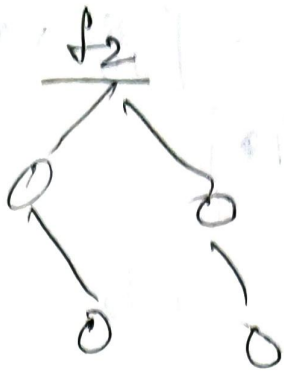
→ We will select f_2 for root node because it is providing more information.

→ f_2 will be better split.

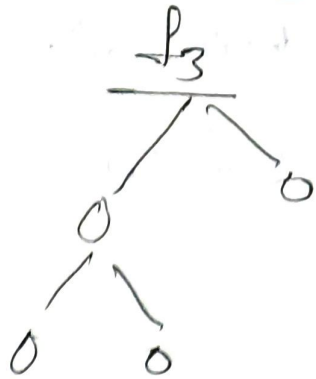
$I.G = 0.98$



$I.G = 0.97$



$I.G = 0.45$



→ Information Gain of f_1 is ~~goo~~ better than others, so f_1 will be root node.

~~Example~~

Example

→ outlook, temp, humidity, wind, decision dataset on 48 page of today's class.

Let's assume outlook has more information Gain.