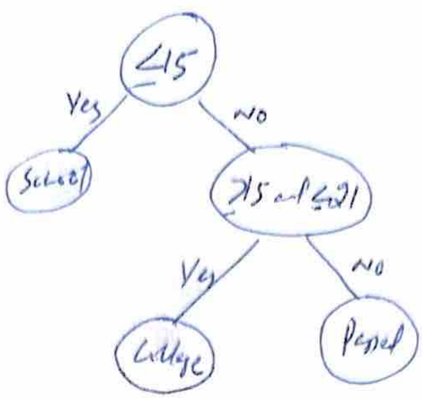
Decision Tree:

(1)

1-data Point

age = 14
if (age ≤ 15):
    Print("School")
elif(age > 15 and age < 21):
    Print ("college")
else
    Print ("Passed college")

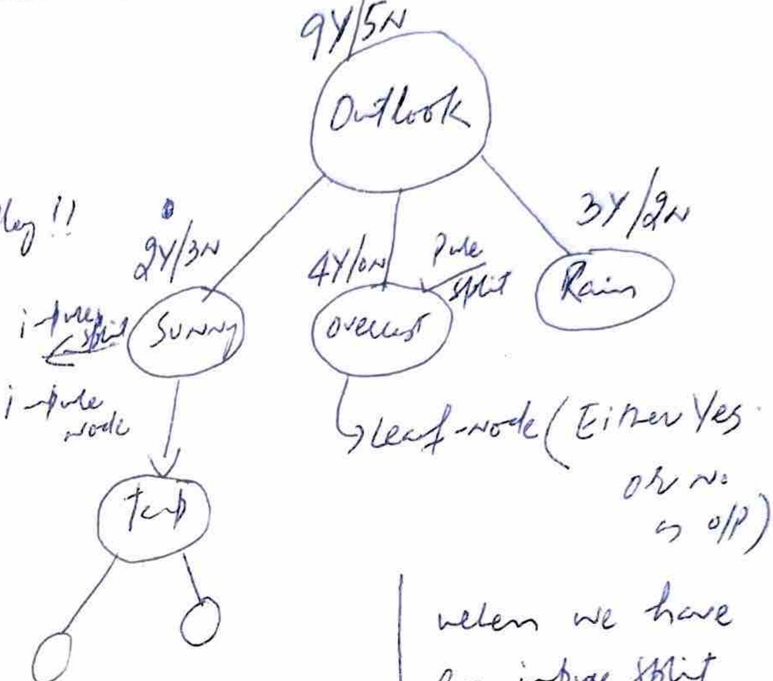


Similarly apply this on the dataset

Play Tennis dataset:

Search Tennis Play !!

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	[illegible]	No
Sunny	Hot	:		No
overcast	Hot	:		Yes
Rain	mild	:		Yes
Rain	cool	Normal		Yes
Rain		:		No
overcast		High		
Sunny		Normal		
Sunny		:		
Rain		:		
Sunny		High		
overcast		:		
overcast				
Rain				



when we have an impure split we take another feature

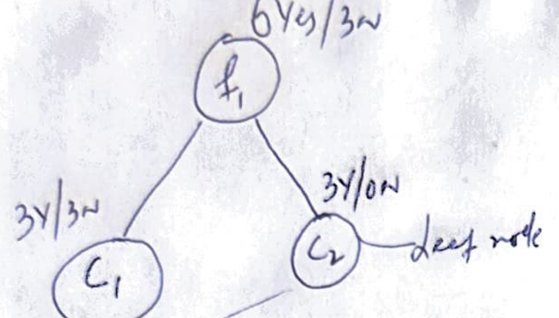
Terminology: (1) Purity: How to check is pure split pure or not (mathematically)?
    ↳ (1) Entropy
    ↳ (2) Gini index / impurity
(2) information Gain → How the features are Selected.

2

① Entropy:

for binary classification

$$H(S) = (-P_+ \log_2 P_+) - (P_- \log_2 P_-)$$



$$H(C_2) = -\frac{3}{3} \log_2 (3/3) - \left(\frac{0}{3} \log 0/3\right)$$

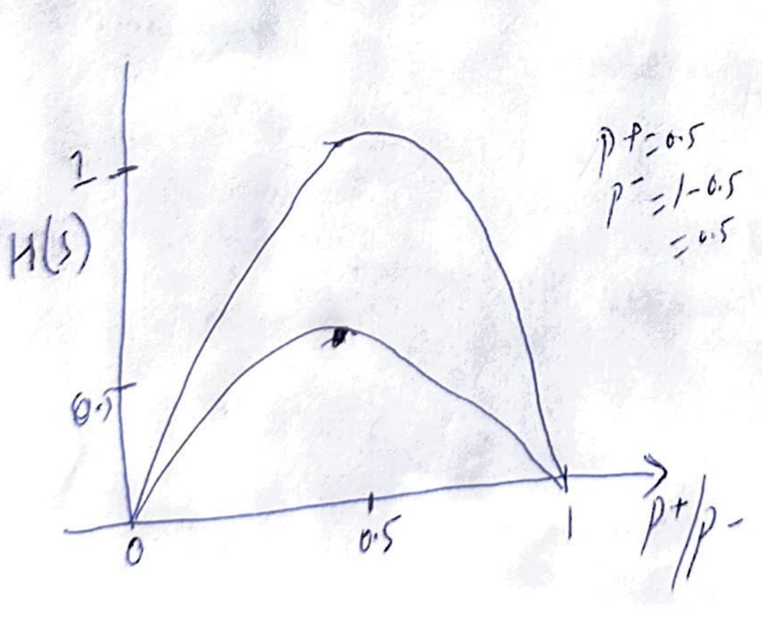
$$= -1 \log_2 1$$

$$= 0 \Longrightarrow \text{Pure Split}$$

$$H(C_1) = -\frac{3}{6} \log_2 (3/6) - \frac{3}{6} \log_2 (3/6)$$

$$= 1 \Rightarrow \text{impure split highly}$$

Entropy range $(0 - 1)$



② Gini impurity

$$G(I) = 1 - \sum_{i=1}^{n} (P)^2$$

for two classes

$$G(I) = 1 - \left((P_+)^2 + (P_-)^2\right)$$

Example ①

3Y/3N

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

$$= 1 - \frac{1}{2}$$

$$= 1/2 = 0.5$$

for impure split $G(I)$ is always 0.5

Example ②

3Y/0N

$$G(I) = 1 - \left((1)^2 - 0\right)$$

$$= 1 - 1$$

$$= 0 \rightarrow \text{Pure Split}$$

G.I range $(0 \rightarrow 0.5)$

---

when to use GI or Entropy

for large dataset use GI

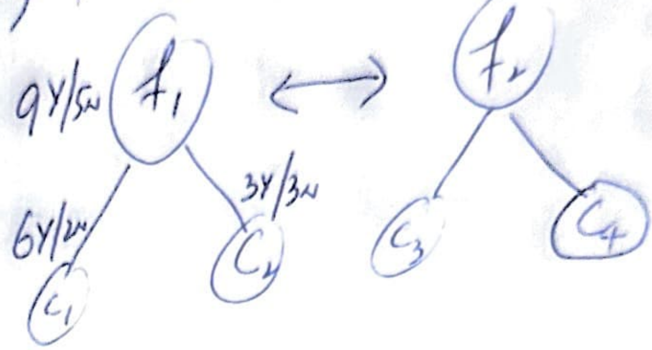
for smaller one use entropy

③

Information Gain (entropy)

$$Gain(S, f_1) = H(S) - \sum_{v \in Val} \frac{|S_v|}{|S|} H(S_v)$$

$f_1, \; f_2, \; f_3$



$$H(S) = -P_+ \log_2(P_+) - P_- \log_2(P_-)$$

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

$$= 0.94$$

S_v = no. of attributes

S = total attributes

$$H(C_1) = -\frac{6}{8}\log_2(6/8) - 2/8 \log_2(2/8)$$

$$= 0.811$$

$$H(C_2) = 1$$

$$Gain(S, f_1) = 0.94 - \left[\frac{8}{14} \times 0.811 + \frac{6}{14} \times 1\right]$$

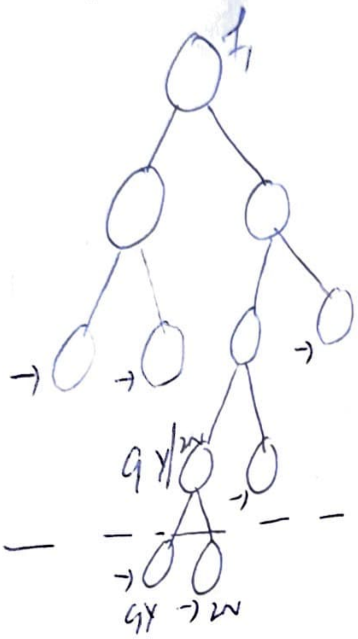
$$= 0.049$$

$$Gain(S, f_2) = 0.052$$

Gain of $f_2 > f_1$ hence we split by starting from f_2.

Pruning: Post Pruning / Pre Pruning

f_1 f_2 f_3 O/P



- leads to overfitting

⊛ Prevention:

(i) Post Pruning → make decision tree then cut

(ii) Pre Pruning

MAX depth

9Y/2N

← No need to split

stop here

⑤

Generating the decision Tree:

Rule: (i) Calculate entropy / Gini impurity of the nodes: — Already discussed

Rule (ii) Calculate information gain of the input features

of the Information Gain $\boxed{(I.G) = \underset{\substack{\uparrow \\ \text{entropy}}}{H(S)} - \frac{|S_1|}{S} H(S_1) - \frac{|S_2|}{|S|} H(S_2)}$

Calculate the IG based criterion to construct a DT for the tennis Example.

→ Let's use IG based criterion to construct a DT for the tennis Example.

Consider feature "<u>wind</u>". Root contains all examples $S = [9+, 5-]$ (Yes, No)

Entropy: $H(S) = -\left(\frac{9}{14}\right)\log_2(9/14) - \left(5/14\right)\log_2(5/14) = 0.94$

$S_{weak} = [6+, 2-] \Rightarrow H(S_{weak}) = -\left(\frac{6}{8}\right)\log_2\left(\frac{6}{8}\right) - \left(\frac{2}{8}\right)\log_2\left(\frac{2}{8}\right) = 0.811$

$S_{strong} = [3+, 3-] \Rightarrow H(S_{strong}) = 1$ {highly impure split entropy is = 1}

$IG(S, wind) = H(S) - \frac{|S_W|}{S} H(S_1) - \frac{|S_S|}{S} H(S_2)$ {$S_W \to S_{weak}$, $S_S \to S_{strong}$}

$= 0.94 - \frac{8}{14} \times 0.811 - \frac{6}{14} \times 1$

$\boxed{= 0.048}$

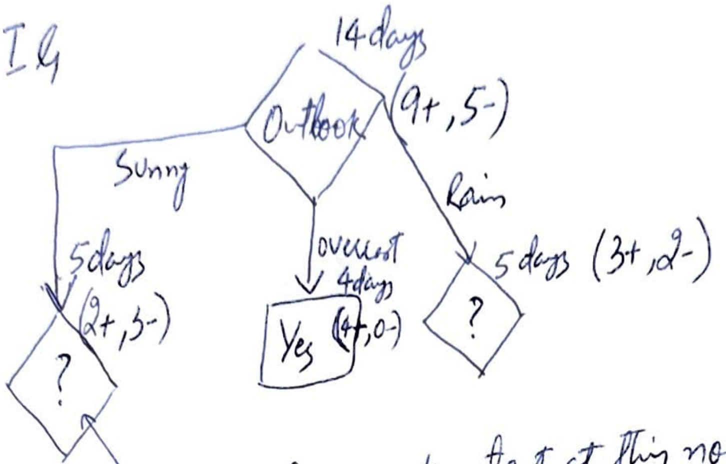
Likewise, at root: $IG(S, \text{outlook}) = 0.246$, $IG(S, \text{humidity}) = 0.151$,

$IG(S, \text{temp}) = 0.029$

Thus we choose "Outlook" feature to be tested as root node

Now, how to grow the DT, i.e what to do at next level? which feature to test next?

__Rule:__ iterate - for each child node, select the feature with the highest IG



which feature to test at this node ?

Proceeding as before, for level 2, left node, we can verify that

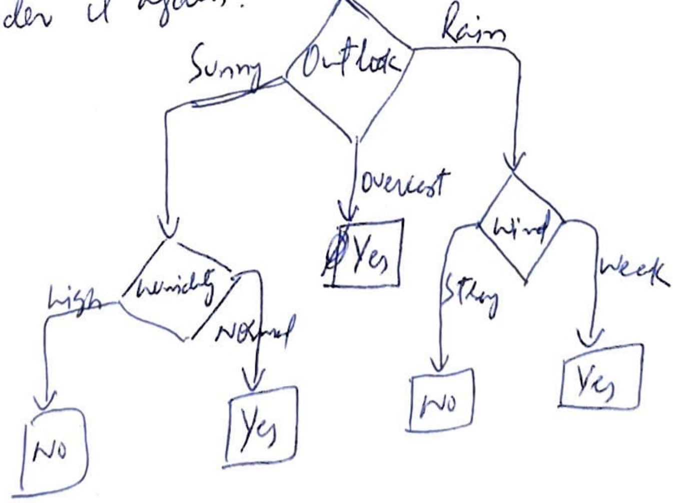
→ $IG(S, temp) = 0.570$, $IG(S, humidity) = 0.970$, $IG(S, wind) = 0.019$

→ Thus humidity chosen as the feature to be tested at level 2, left node

→ No need to expand the middle node (already "Pure" - all "Yes" training Examples)

→ Can also verify that wind has the largest IG for the right node.

__Note:__ If a feature has already been tested along a path earlier, we don't consider it again.



Predict

Q: X (outlook = Sunny, temperature = cool, humidity = high, ~~windy = false~~) (Play ?