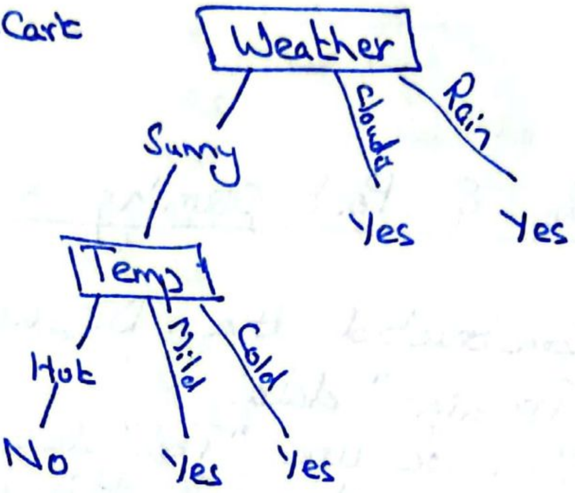
Lecture -6

→ Decision Tree

1) Can be used for both classification & regression, however mainly used for classification.
2) Some important points to consider:
   - a) Tree Structure } — ID3 ~
   - b) Decision Nodes } — Cart
   - c) Leaf Nodes
   - d) Split (pure, impure)
   - e) Entropy
   - f) Information gain
   - g) Pruning (Pre, post)



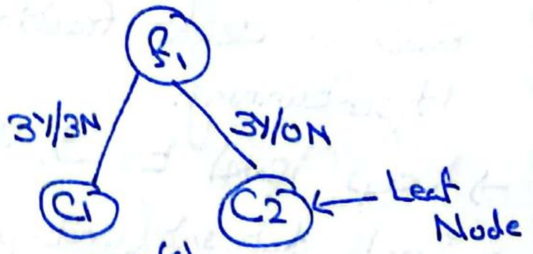
Questions: To check the purity of split we use
↳ Entropy
↳ Gini Index (for large datsets. Have simple math so computationaly efficient)

: Which feature to chose for splitting.
↳ Information Gain.

Entropy

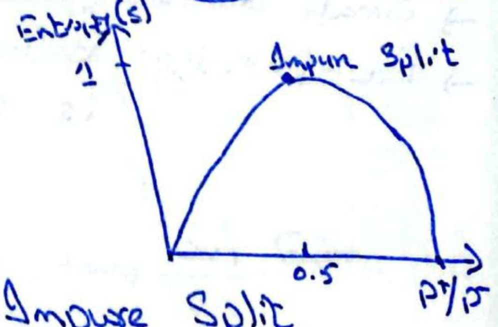
Assuming Binary (Yes/No)

$$\text{Entropy}(s) = p^+ \log_2 p^+ - p^- \log_2 p^-$$



$$S_{C_2} = \frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$$

$$= -1\log_2 1 = \boxed{0} \quad \text{Pure Split}$$



$$S_{C_1} = \frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$

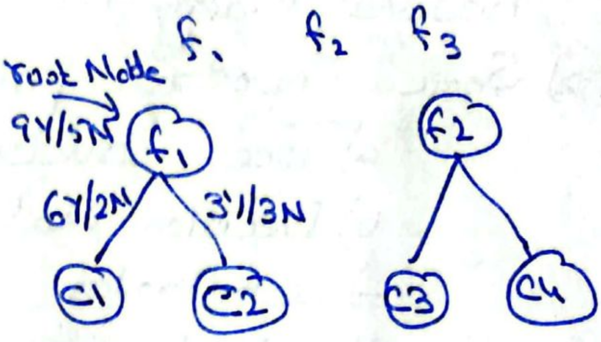
$$= \frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = \boxed{1} \quad \text{Impure Split}$$

Entropy will always be between 0 ~ 1.

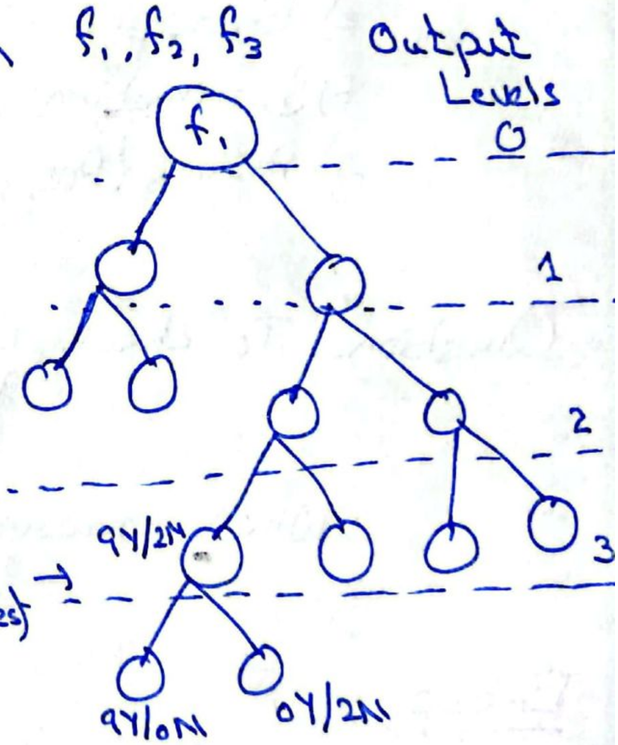
-> How to split, which feature to be the node.

Information Gain

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



-> Pre & Post pruning DT



-> Constructed the DT using "Training" data.

-> When we use "Test" data, the accuracy is lower & this phenomena is known as overfitting.

-> To resolve overfitting
- ↳ Post Prunning
- ↳ Pre Prunning

-> Hyperparameters (Max depth, minimum samples split/leaf, minimum weight fraction). & H Hypertunning.

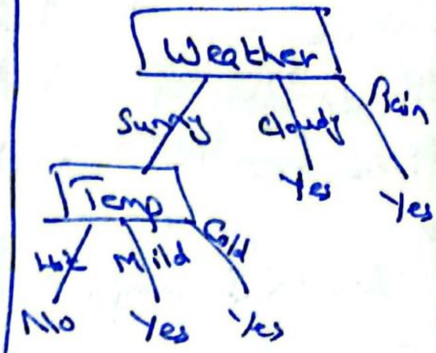
-> Keep level to 3. So reduced the overfitting issue

-> Small dataset (post-prunning), Big datasets (pre-prunning)

-> Pre-prunning: Before constructing DT, at what depth is accuracy highest.

Numerical Example:

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weaks	Yes
D14	Rain	Mild	High	Strong	No



Values(Outlook) = Sunny, Overcast, Rain

$S[9+, 5-]$ $\quad \text{Entropy}(s) = \frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$

$= \boxed{0.94}$

$S_{sunny}[2+, 3-]$ $\quad \text{Entropy}(S_{sunny}) = \frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = \boxed{0.971}$

$S_{overcast}[4+, 0-]$ $\quad \text{Entropy}(S_{overcast}) = \frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = \boxed{0}$

$S_{Rain}[3+, 2-]$ $\quad \text{Entropy}(S_{Rain}) = \frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = \boxed{0.97}$

$$\text{Gain}(S, \text{Outlook}) = \text{Entropy}(S) - \sum_{S \in (\text{overcasting}, \text{Sunny}, \text{Rain})} \frac{|S_v|}{|S|}\text{Entropy}(S_v)$$

$$= 0.94 - \frac{5}{14}(0.971) - \frac{4}{14}(0) - \frac{5}{14}(0.971)$$

$$= 0.2464$$

Attribute: Temp

Values(Temp) = Hot, Mild, Cool

$S = [9+, 5-]$ Entropy(s) = 0.94

S_{Hot} $[2+, 2-]$ $\text{Entropy}(S_{Hot}) = \frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1$

S_{Mild} $[4+, 2-]$ $\text{Entropy}(S_{Mild}) = \frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} = 0.9183$

S_{Cool} $[3+, 1-]$ $\text{Entropy}(S_{Cool}) = \frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.8113$

$$\text{Gain}(S, \text{Temp}) = \text{Entropy}(S) - \sum_{v \in (\text{Hot, Mild, Cool})} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$= 0.94 - \frac{4}{14}(1) - \frac{6}{14}(0.9183) - \frac{4}{14}(0.8113)$$

$$= 0.0289$$

Attribute: Humidity

Value(Humidity) = High, Normal

$S = [9+, 5-]$ Entropy(s) = 0.94

S_{High} $[3+, 4-]$ $\text{Entropy}(S_{High}) = 0.9852$

S_{Normal} $[6+, 1-]$ $\text{Entropy}(S_{Normal}) = 0.5916$

$$\text{Gain}(S, \text{Humidity}) = \text{Entropy}(S) - \sum_{v \in (\text{High, Normal})} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$= 0.94 - \frac{7}{14}(0.9852) - \frac{7}{14}(0.5916)$$

$$= 0.1516$$

Attribute: Wind

Values(Wind) = Strong, Weak

$S = [9+, 5-]$ Entropy(S) = 0.94

S_{Strong} $[3+, 3-]$ $\text{Entropy}(S_{Strong}) = 1$

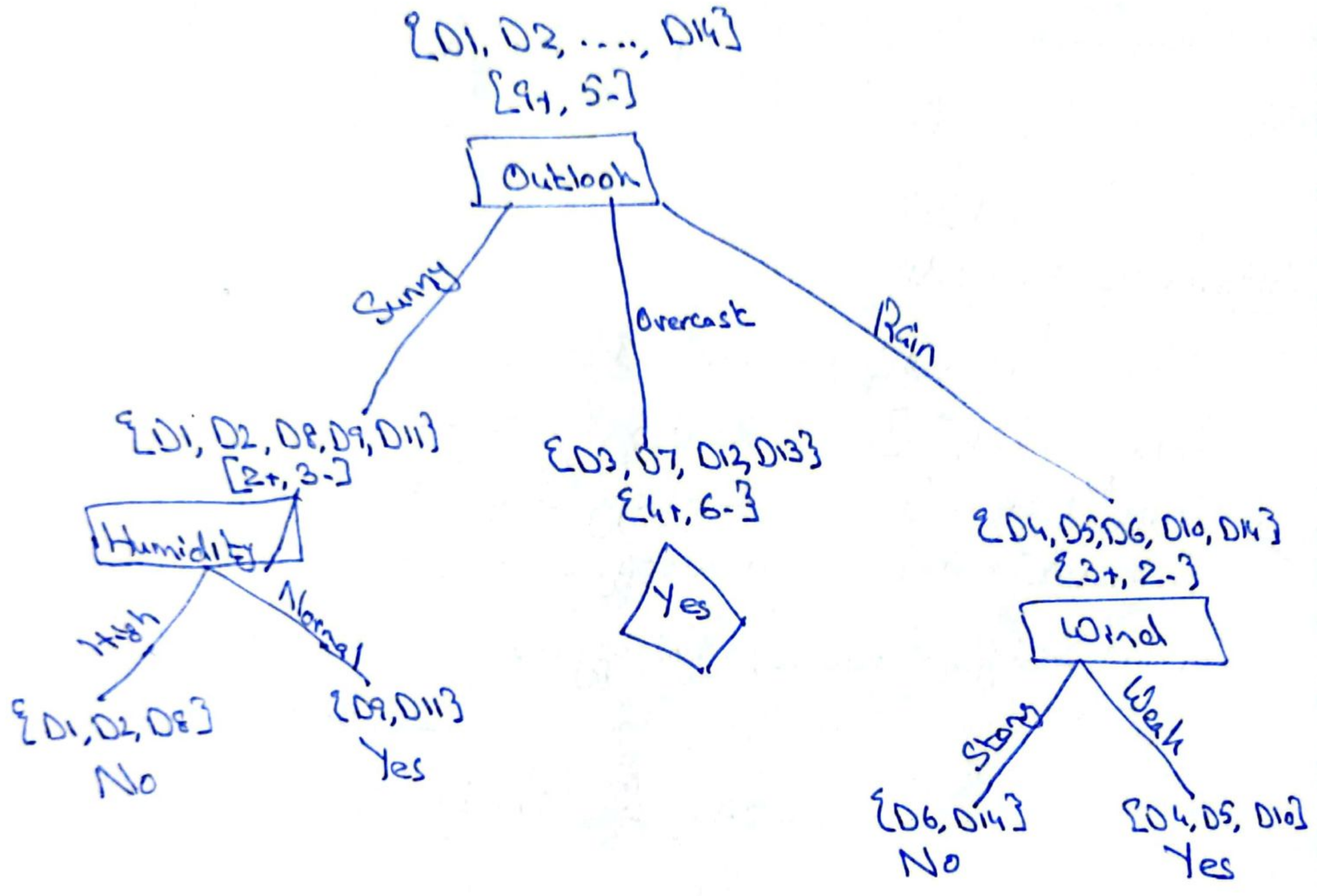
S_{Weak} $[6+, 2-]$ $\text{Entropy}(S_{Weak}) = 0.8113$

$$\text{Gain}(S, \text{Wind}) = 0.94 - \frac{6}{14}(1) - \frac{8}{14}(0.8113)$$

$$= 0.0478$$

So

Gain(S, Outlook) = 0.2464 | Gain(S, Temp) = 0.0289 | Gain(S, Humidity) = 0.1516 | Gain(S, Wind) = 0.0478



Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

Attribute: Temp

Values(Temp) = Hot, Mild, Cool

$S_{sunny} = [2+, 3-]$ Entropy$(S_{sunny}) = 0.97$

S_{Hot} [0+, 2-] Entropy(S_{Hot}) = 0

S_{Mild} [1+, 1-] Entropy(S_{Mild}) : 1

S_{Cool} [1+, 0-] Entropy(S_{Cool}) : 0

$$Gain(S_{sunny}, Temp) = Entropy(S) - \sum_{v \in (H,M,C)} \frac{|S_v|}{|S|} Entropy(S)$$

$$= 0.97 - \frac{2}{5}(0) - \frac{2}{5}(1) - \frac{1}{5}(0)$$

$$= \boxed{0.570}$$

Attribute: Humidity

Values(Humidity) = High, Normal

S_{sunny} = [2+, 3-] Entropy(S) = 0.97

S_{High} [0+, 3-] Entropy(S_{High}) = 0

S_{Normal} [2+, 0-] Entropy(S_{Normal}) = 0

$$Gain(S_{sunny}, S_{Humidity}) = Entropy(S) - \frac{3}{5} Entropy(S_{sunny}) - \frac{2}{5} Entropy(S_{Normal})$$

$$= 0.97 - \frac{3}{5}(0) - \frac{2}{5}(0)$$

$$= 0.97$$

Attribute: Wind

Value(Wind) = Strong, Wind

S_{sunny} = 0.97

S_{strong} [1+, 1-] Entropy(S_{strong}) = 1

S_{weak} [1+, 2-] $Entropy(S_{weak}) = \frac{1}{3}\log_2 \frac{1}{3} - \frac{2}{3}\log_2 \frac{2}{3}$

$= 0.9183$

$$Gain(S_{sunny}, Wind) = 0.0192$$

For Rain

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

Attribute: Temp

Value(Temp) = Hot, Mild, Cool

$S_{Rain} = [3+, 2-]$ Entropy$(S_{sunny}) = 0.97$

$S_{Hot} \leftarrow [0+, 0-]$ Entropy$(S_{Hot}) = 0$

$S_{Mild} \leftarrow [2+, 1-]$ Entropy$(S_{Mild}) = 0.9183$

S_{Cool} $[1+, 1-]$ Entropy$(S_{Cool}) = 1$

Gain$(S_{Rain}, Temp) = 0.0192$

Attribute: Humidity

S_{Rain} $[3+, 2-]$ Entropy$(S_{sunny}) = 0.97$

S_{High} $[1+, 1-]$ Entropy$(S_{High}) = 1$

S_{Normal} $[2+, 1-]$ Entropy$(S_{Normal}) = 0.9183$

Gain$(S_{Rain}, Humidity) = 0.0192$

Attribute: Wind

S_{Rain} $[3+, 2-]$ Entropy$(S_{sunny}) = 0.97$

S_{Strong} $[0+, 2-]$ Entropy$(S_{Strong}) = 0$

S_{Weak} $[3+, 0-]$ Entropy$(S_{Weak}) = 0$

Gain$(S_{Rain}, Wind) = 0.97$