> The biggest challenge of decision tree algorithm is to find out which feature to split upon.

> i.e Identifying the feature, i.e the data should be split, and the split should cont examples / datatems belonging to a single cla

> Then only partitions are considered to be pure

So, if we take the same example dataset first we need to understand, how many attributes?.

Are they binary / Multi-valued?

So, among the 4 attributes, which attribute would be my root node? which would be the next level node & so on?

> To answer this, we need to find out information gain of every attribute in the problem.

> Once we calculate information gain, we can conclude, which attribute has the highest importance.

Now, to calculate information gain, you need to first calculate the entropy.

Entropy

- Entropy is a measure of impurity of a collection of examples. (uncertainty in a given dataset)
- Entropy depends on the distribution of random variable p.
  - → S is a collection of training examples
  - → p_+ the proportion of +ve examples
  - → p_- the proportion of -ve examples

$$\text{Entropy}(S) = -P_{\oplus} \log_2 P_{\oplus} - P_{\ominus} \log_2 P_{\ominus}$$

Eg:- There are (3) cases

(1) If Entropy([14+, 0-]) = $-14/14 \log_2(14/14) - 0 \log_2(0) = 0$

If you have all positive examples, like above where we have 14 +ve example and 0 -ve example or vice versa, you don't need to apply formula as Entropy will always be 0.

(Combination of true & -ve examples

(2) Entropy $(9+, 5-) = -9/14 \log_2(9/14) - 5/14 \log_2(5/14)$

$= 0.94$

have to apply formula.

(Equal true & -ve examples

(3) Entropy $(7+, 7-) = -7/14 \log_2(7/14) - 7/14 \log_2(7/14)$

$= 1/2 + 1/2 = 1$

In this case entropy will always be 1.

Information Gain Information gain is calculated on the basis of decrease in entropy(s) caused by the partitioning/datasplit of the examples according to the attribute

> The attribute with 'minimum' impurity is the important node.

> Constructing a decision tree is all about finding an attribute that returns highest information gain

Decision Tree - ID3 Algorithm 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	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No.

Attribute : Overlook

value (outlook) : Sunny, Overcast, Rainy

$S = [9+, 5-]$

$$\text{Entropy}(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.94$$

$S_{sunny} \leftarrow [2+, 3-]$

$$\text{Entropy}(S_{sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.971$$

$S_{overcast} \leftarrow [4+, 0-]$

$$\text{Entropy}(S_{overcast}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

$S_{Rain} \leftarrow [3+, 2-]$

$$\text{Entropy}(S_{Rain}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= 0.971$$

$$\text{Gain}(S_{outlook}) = \text{Entropy}(S) - \sum_{v \in (Sunny, Overcast, Rain)} \frac{|S_v|}{|S|}\text{Entropy}(S_v)$$

$$= \text{Entropy}(S) - \frac{5}{14}\text{Entropy}(S_{sunny}) - \frac{4}{14}\text{Entropy}(S_{overcast}) - \frac{5}{14}\text{Entropy}(S_{Rain})$$

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

$= 0.2464$

Gain (S, outlook) = 0.2464

Similarly we need to calculate Information gain for other attributes.

Attributes: (temp) → Hot, mild, cool

Value (temp)

Entropy (S) = 0.94

S_{Hot} [2+, 2-] Entropy (S_{hot}) = 1.0

S_{mild} [4+, 2-] 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-] Entropy (S_{cool}) = $\frac{-3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$ = 0.8113

Gain (S_{temp}) = $0.94 - \frac{4}{14} * 1.0 - \frac{6}{14} * 0.9183 - \frac{4}{14} * 0.8113 = 0.0289$

= 0.0289

Attribute : Humidity

values (Humidity) = High, Normal

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

$S_{high} \leftarrow [3+ 4-]$ $\quad Entropy(S_{high}) = 0.9852$

$S_{Normal} \leftarrow [6+, 1-]$ $\quad Entropy(S_{normal}) = 0.591$

$Gain(S, Humidity) = Entropy(S) - \frac{7}{14} * 0.9$
52

$- \frac{7}{14} * 0.5916 = 0.1516$

0.1516

Attribute : Wind

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

$S_{strong} \leftarrow [3+, 3-]$ $\quad Entropy(S_{strong}) = 1.0$

$S_{weak} \leftarrow [6+, 2-]$ $\quad Entropy(S_{weak}) = 0.811$

Fig

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

$$= 0.94 - \frac{6}{14} \times 1.0 - \frac{8}{14} * 0.8113$$

$$= 0.0478$$ //

After calculating gain of all attributes we need to check which attribute has the highest information gain.

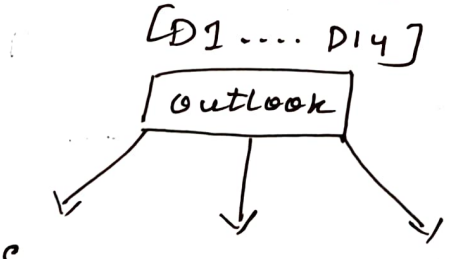
Gain (S, outlook) = 0.2464 ✓

Gain (S, temp) = 0.0289

Gain (S, humidity) = 0.1516

Gain (S, wind) = 0.0478

Outlook is the root node.



Now, lets consider for Sunny

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

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

$= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

$S_{hot} \leftarrow [0+2-]$ $= Entropy(S_{hot}) = 0$

$S_{Mild} \leftarrow [1+, 1-]$ $= Entropy(S_{mild}) = 1.0$

$S_{cool} \leftarrow [1+, 0-]$ $= Entropy(S_{cool}) = 0.0$

$Gain_{temp}(S_{sunny}, temp)$

$= Entropy(S_{sunny}) - \sum_{v \in (Hot, mild, cool)} \frac{|S_v|}{|S|} Entropy(S_i)$

$= 0.97 - \frac{2}{5} * 0 - \frac{2}{5} * 1.0 - \frac{1}{5} * 0$

$= 0.570$ //

Attribute : Humidity

value : High, Normal

$S_{sunny} = 0.97$

$S_{high} \leftarrow [0+, 3-]$ $\quad Entropy(S_{high}) = 0$

$S_{Normal} \leftarrow [2+, 0-]$ $\quad Entropy(S_{Normal}) = 0.0$

$$Gain(S_{sunny}, humidity) = Entropy(S_{sunny}) - \sum_{v \in [High, Normal} \frac{|S_v|}{|S|} Entropy(S_v)$$

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

$$= 0.97$$

Attribute: Wind

value: (Strong, Weak)

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

$S_{strong} [1+, 1-] =$ $\quad Entropy(S_{strong}) = 1.0$

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

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

So, $Gain(S_{sunny, temp}) = 0.570$

$Gain(S_{sunny, Humidity}) = 0.97$

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

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

values (temp)

$S[Rain] = [3+, 2-] = 0.97$

$S_{Hot} \leftarrow [0, 0]$ Entropy $[S_{hot}] = 0.0$

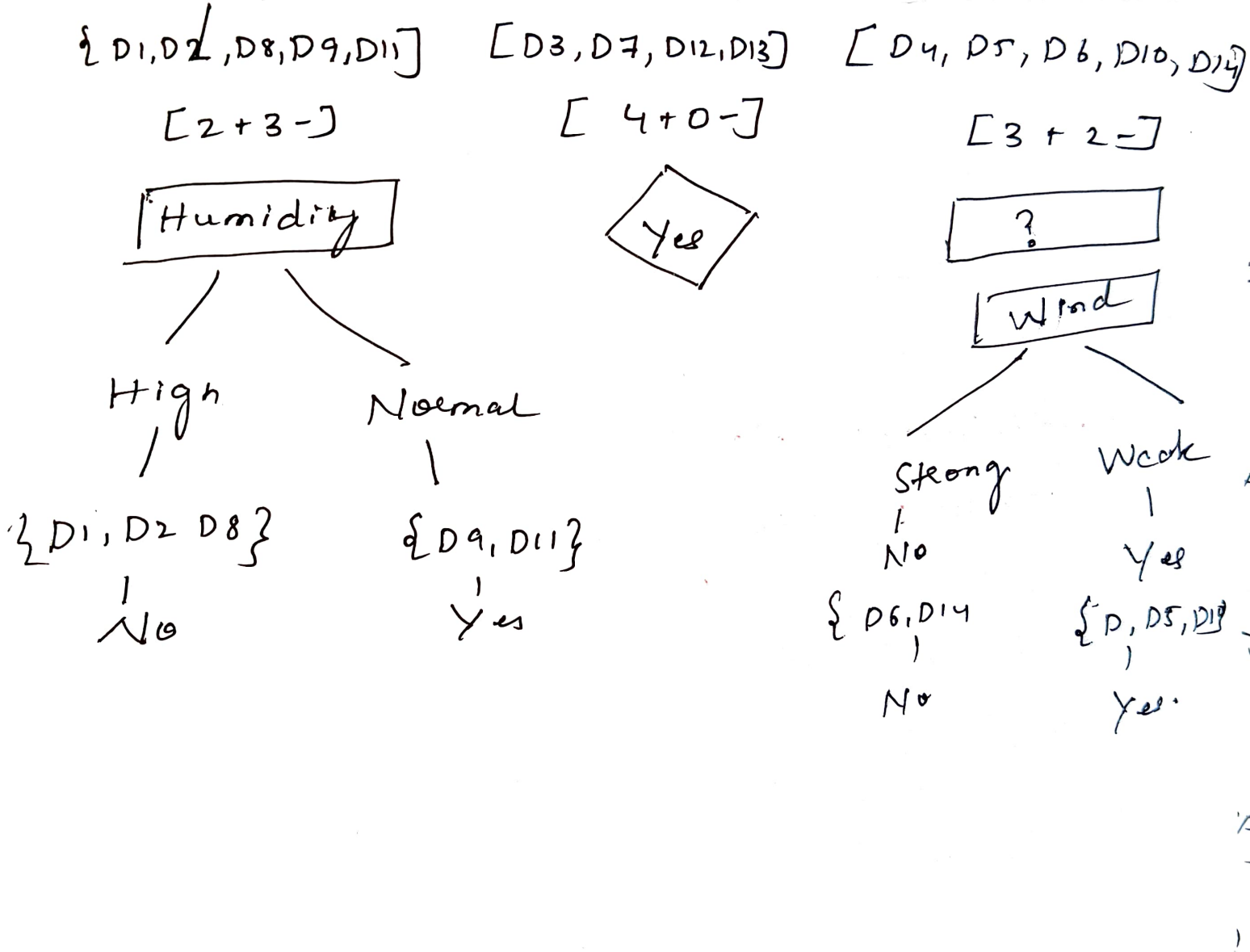
$S_{mild} \leftarrow [2+, 1-]$ Entropy $[S_{mild}] = 0.9183$

$S_{cool} = [1+, 1-]$ Entropy $[S_{cool}] = 1.0$

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

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

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


[D1, D2, D8, D9, D11]
[2+ 3-]
[D3, D7, D12, D13]
[4+ 0-]
[D4, D5, D6, D10, D14]
[3+ 2-]
Humidity
High
Normal
{D1, D2 D8}
No
{D9, D11}
Yes
Yes
?
Wind
Strong
Weak
No
{D6, D14}
No
Yes
{D, D5, D1
Yes.