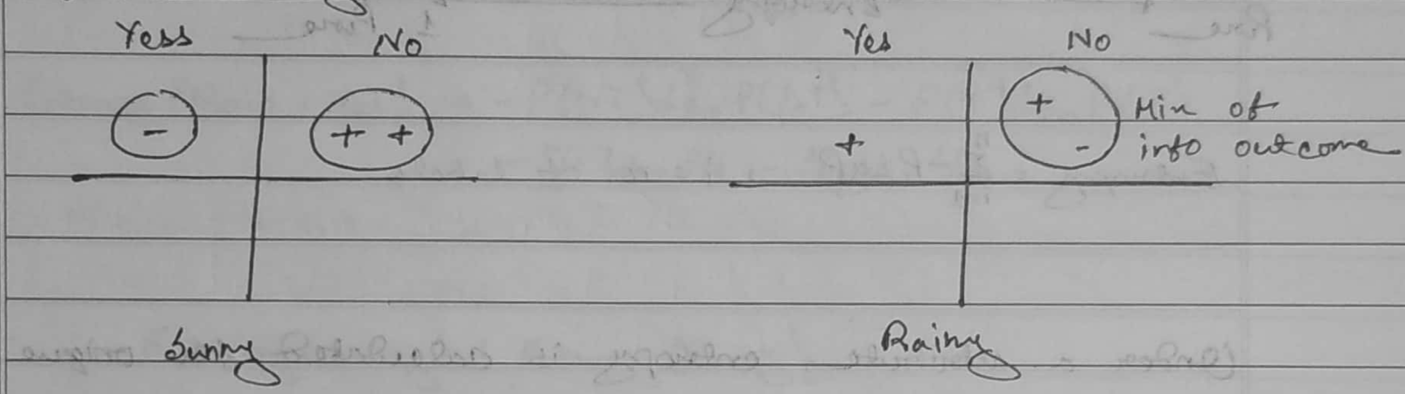
Subject : Decision Tree

Date :

We try to form the decision tree as concise as possible. A concise decision tree is likely to be efficient and given a test data it ~~is~~ can reach the decision easily.

Information Purity:



Take Umbrella → +ve

Don't take Umbrella → −ve

More Pure = More Information

Information wise, sunny is more pure than rain. Because under Yes and No there is no mixture of + and −. Under the class Yes and No they (+, −) are both getting conclusively separated.

Since, sunny is more pure so it has more information. So, if our target is to make a tree as concise as possible then sunny should have more priority as the root node.

It is impossible to intuitively calculate information purity visually when the data set is very big like thousands of samples with 30/40 attributes. In this case, the amount of decision with respect to attribute will also be greater.

The mathematical term to calculate information purity is called entropy.

Subject :

Date :

The higher the value of entropy is, the attribute is less pure.

More ~~Less~~ Pure 0 ——— Entropy ———→ 1 Less ~~More~~ Pure

$$\text{Entropy} = \sum_{i=1}^{n} -P \log P \quad ; \quad i = \text{no. of labels}$$

Under a attribute, entropy is calculated for unique ~~vales~~ values.

Sunny	Rainy	Decision
Yes	No	Don't Take Umbrella
No	Yes	Take Umbrella
No	No	Take Umbrella

Less Entropy = More Pure = More Information

$$\text{Entropy (Sunny = Yes)} = -P(DT)\log_2 P(DT) - P(T)\log_2 P(T)$$
$$= -\tfrac{1}{1}\log_2 \tfrac{1}{1} - \tfrac{0}{1}\log_2 \tfrac{0}{1}$$
$$= 0$$

= So this is very pure

$$\text{Entropy (Sunny = No)} = -P(DT)\log_2 P(DT) - P(T)\log_2 P(T)$$
$$= -0/2 \log_2 P(0/2) - 2/2 \log_2 2/2$$
$$= \log_2 1$$

= 0 So the information is very pure.

Subject : | Date :

Entropy (Rainy = ~~Yes~~ No) $= -P(\Delta T)\log_2 P(\Delta T) - P(T)\log_2 P(T)$

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

$= -0.5\log_2 0.5 - 0.5\log_2 0.5$

$= 1$ → Max possible value. So the information is very impure

Entropy (Rain = Yes) $= -P(\Delta T)\log_2 P(\Delta T) - P(T)\log_2 P(T)$

$= -0/1 \log_2 0/1 - 1/1 \log_2 1/1$

$= 0$

∴ Entropy (Sunny) < Entropy (Rainy)

∴ Sunny is more pure.

Feature selection is based on purity. The more impure the feature is, the less it is priotized.

→ So, for root node Sunny will be picked.

Based on the amount of information gain we select feature for the root node. The decision tree creation algorithm based on the information gain is called ID3. The more the information gain, the more priority it gets to be the root node.

CS CamScanner

Subject :

Date :

E.g. $E(\text{Decision}) = -P(Y)\log_2 P(Y) - P(N)\log_2 P(N)$

$= -9/14 \log_2 9/14 - 5/14 \log_2 5/14$

$= 0.940$

$E(\text{Outlook} = \text{Sunny}) = -P(Y|S)\log_2 P(Y|S) - P(N|S)\log_2 P(N|S)$

$= -2/5 \log_2 2/5 - 3/5 \log_2 3/5$

$= 0.971$

$E(\text{Outlook} = \text{Rain}) = -P(Y|R)\log_2 P(Y|R) - P(N|R)\log_2 P(N|R)$

$= -3/5 \log_2 3/5 - 2/5 \log_2 2/5$

$= 0.971$

$E(\text{Outlook} = \text{Overcast}) = -P(Y|O)\log_2 P(Y|O) - P(N|O)\log_2 P(N|O)$

$= -4/4 \log_2 4/4 - 0/4 \log_2 0/4$

$= 0$

$IG(\text{Outlook}) = E(\text{Decision}) - E(S) - E(R) - E(O)$

$\cancel{= 0.940 - \frac{5}{14}}$

$= E(\text{Decision}) - P(S)E(S) - P(R)E(R) - P(O)E(O)$

$= 0.940 - 5/14(0.971) - 5/14(0.971) - 4/14(0)$

$= 0.246$

Subject :

Date :

For Humidity

$E(D) = -P(Y)\log_2 P(Y) - P(N)\log_2 P(N)$

$= -9/14 \log_2 9/14 - 5/14 \log_2 5/14$

$= 0.940$

~~E(Humidity = High) = -P(H|H) Lg2 P(H|H) - P(~~

$E(\text{Humidity} = \text{High}) = -P(Y|H)\log_2 P(Y|H) - P(N|H)\log_2 P(N|H)$

$= -3/7 \log_2 3/7 - 4/7 \log_2 4/7$

$= 0.985$

$E(\text{Humidity} = \text{Normal}) = -P(Y|N)\log_2 P(Y|N) - P(N|N)\log_2 P(N|N)$

$= -6/7 \log_2 6/7 - 1/7 \log_2 1/7$

$= 0.592$

$IG(H) = E(D) - P(H) * E(H) - P(N) * E(N)$

$= 0.940 - (7/14 * 0.985) - (7/14 * 0.592)$

$= 0.151$

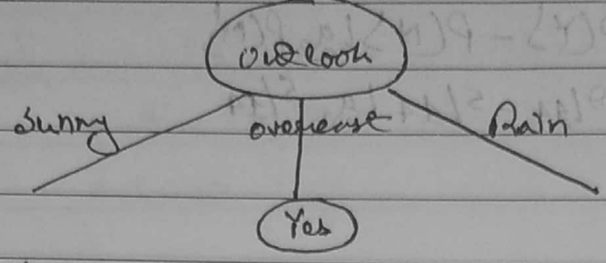
$IG(W) = 0.048$ More IG = More Priority.

$IG(T) = 0.029$

The less the entropy, the greater the information gain.

Here, outlook has the most information gain. So, it will be the root node.

Subject : Date :



The rows of overcast have only the decision value of YES. So there is no conflict of decision. We are getting pure information.

For Temperature under Sunny.

$E(D) = -2/5 \log_2 2/5 - 3/5 \log_2 3/5$

$= 0.970$

$E(T = H) = -P(Y|H) \log_2 P(Y|H) - P(N|H) \log_2 P(N|H)$

$= -2/2 \log_2 2/2 - 0/2 \log_2 0/2$

$= 0$

$E(M) = -1/2 \log_2 1/2 - 1/2 \log_2 1/2 = 1$

$E(C) = 0$

$IG(T|O) = 0.970 - 2/5(0) - 2/5(1) - 1/5(0)$

$= 0.570$

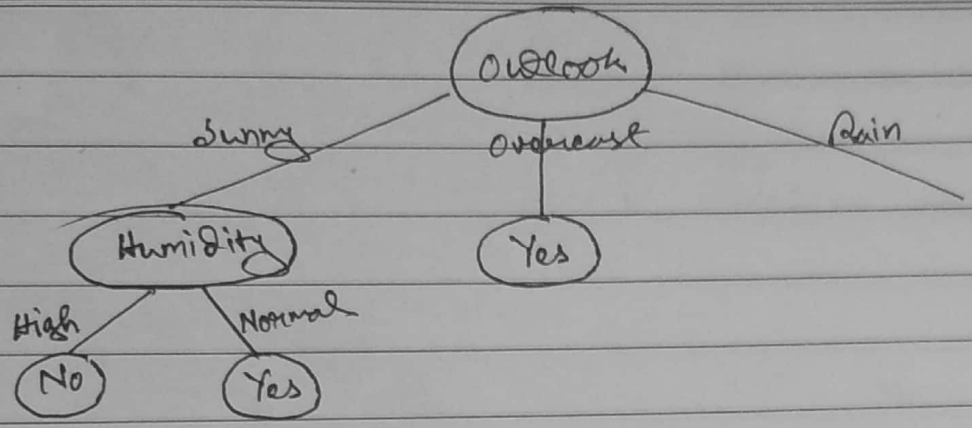
$IG(H|O) = 0.970$

$IG(W|S) = 0.019$

$\therefore$ Under Sunny, Humidity will sit.

Subject :

Date :



20+ } Highest Info needed
20- } High uncertainty

20+ } no info needed
0 - } no uncertainty

~~20~~ 14+ } Needs less info
2- } Less uncertainty.

The higher the information gain, the higher the seperating ability.

The higher the info gain, the higher the no. of branches. i.e
info. gain favours attributes ~~hig~~ having large no. of branches.

$$\text{Gain Ratio} = \frac{\text{Info. Gain}}{\text{Split Information}}$$