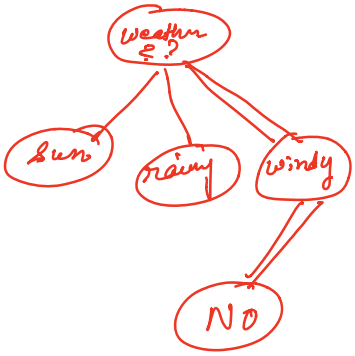
Dec-22 02/03/21 **Decision Trees**

→ Introduction
→ Statistical measures for attribute selm.

Suppose the question is:
"What is the weather outlook today?"

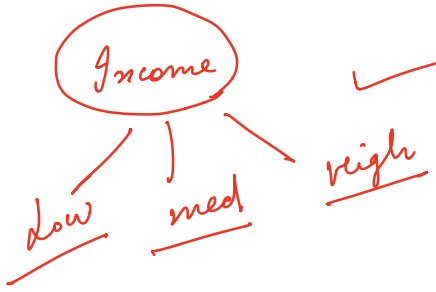
↓
"Is it windy"
↓
"Should we play golf today"
↓
"NO"



< Age, Income, credit rating >

< Entropy Info. gain split into Gini index >

Gini (Age)
Gini (Income) min ✓
Gini (CR)



→ DT are otherwise known as Classifications Trees or hierarchical classifiers. These are rule-based learning - top down induction method.

→ Quinlan [1986] introduced ID3 (Interactive Dichotomizer third series) for DT

ID3: Entropy is used to measure the information content of an attribute.

→ C4.5 is the advanced version of ID3

Information gain is used in C4.5.

① Entropy / Information Gain :

→ Entropy is a measure of how much uncertainty is present in the information.

→ Information gain is a measure of how much inform^n the answer to a specific question provides.

Prob Entropy : $$H(X) = -\sum_{i=1}^{n} P_i \log_2 P_i$$

where P_i is the probability of occurrence of i

Q Let us consider 3 possible events A, B and C having equal prob. of occurence. find the entropy.

Ans 3 possible outcomes with equal prob = 1/3

$$(E) = -\left(P_A \log_2 P_A + P_B \log_2 P_B + P_C \log_2 P_C\right)$$

$$= -\left(\tfrac{1}{3}\log_2 \tfrac{1}{3} + \tfrac{1}{3}\log_2 \tfrac{1}{3} + \tfrac{1}{3}\log_2 \tfrac{1}{3}\right)$$

$$= -\tfrac{1}{3}\left(\log_2 \tfrac{1}{3} + \log_2 \tfrac{1}{3} + \log_2 \tfrac{1}{3}\right)$$

$$= -(-1.59) = 1.59$$

* In information theory, the information content (gain) is maximized, and entropy is minimized.

prob 2 Data for DT is given below :

ID	Age	Income	student	CR	Buy car
1	Y	high	NO	fair	NO
2	Y	H	NO	good	NO
3	M	H	NO	f	Y
4	old	m	NO	f	Y
5	O	l	Y	f	Y
6	O	l	Y	g	N
7	m	l	Y	g	Y
8	Y	m	NO	f	N
9	Y	l	Y	f	Y
10	O	m	Y	f	Y
11	Y	m	Y	g	Y
12	M	m	NO	g	Y
13	M	h	Y	f	Y
14	O	m	NO	g	N

3 attribute selⁿ. measures are

1. Information Gain
2. Gain Ratio (split info)
3. Gini Index

1) Information Gain :

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2 P_i \quad -(1)$$ (also known as Entropy)

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{D} \times Info(D_j) \quad -(2)$$

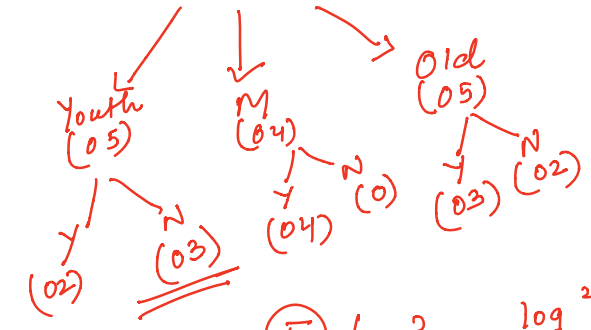
$$Gain(A) = Info(D) - Info_A(D) \quad -(3)$$

Soln : 14 data set → 9 (pos- buy car - yes)
→ 5 (neg - buy car - NO)

$$Info(D) = -\frac{9}{14}\log_2^{9/14} - 5/14 \log_2^{5/14}$$

$$= 0.94 \text{ bits}$$

Info (Age)



$$Info(Age) = \frac{5}{14}\left(-\frac{2}{5}\log_2^{2/5} - 3/5 \log_2^{3/5}\right) + \frac{4}{14}\left(-\frac{4}{4}\log_2^{4/4}\right) + \frac{5}{14}\left(-\frac{3}{5}\log_2^{3/5} - \frac{2}{5}\log_2^{2/5}\right)$$

$$= 0.694$$

$$Gain(Age) = Info(D) - Info(Age)$$

$$= 0.94 - 0.694$$

$$= \boxed{0.246} \checkmark \text{ (maximum)}$$

Gain (Income) = 0.029 bits

Gain (student) : 0.151 bits

Gain (CR) : 0.048 bits

→ Check for maximum information gain or min entropy value for every attribute, so as to select the splitting attribute