DECISION TREE: CONNECTING NODES

A Decision Tree is a diagram or chart that is used to define~~d~~/determine a course of action or show a statistically probability.

→ Each branch of the decision tree represents a possible decision, outcome or reaction.

WHY DO WE USE DECISION TREE?

Decision tree provides an effective method of decision making because they...,

↳ Provide a framework to quantify the values of outcomes and the probabilities of achieving them.

↳ Allow us to analyze fully the possible consequences of a decision.

NOTE:

*↳ A primary advantage for using a decision tree is that it is easy to follow and understand.

FACTORS OF DECISION TREE:

The Decision Trees are composed of three parts...

→ DECISION NODES - Denoting choice.

→ CHANCE NODES - Denoting probability

→ END NODES - Denoting outcomes.

↳ Decision Trees can be used to deal with complex datasets, and can be pruned if necessary to avoid overfitting.

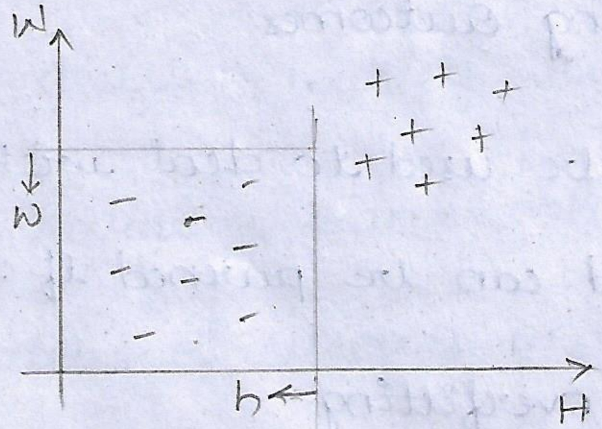
→ As we know that, Decision trees is that where the data is continuously split according to a certain parameter.

→ So, the tree can be explained by two entities, i.e., Decision Nodes & Leaves.

*→ Decision Tree helps in both classification and Regression.

*
↳ Decision Tree creates if-else statements on its own.

CLASSIFICATION :



→ 2 classes
↳ Male
↳ Female.

h & w are less than H & W which belongs to female class.

→ In Programatic level.

I/p – if (height < H and weight < W):

Print ("FEMALE")

else:

Print ('MALE')

But Mathematically, we can't write so,

Geometrically – IT GIVES AXIS PARALLEL LINES

NOTE:

As the depth of the decision tree grows the statements will be complex.

DEPTH SIZE:

The depth of a decision tree is the length of the longest path from a root to a leaf.

↳ The size of a decision tree is the number of nodes in the tree.

NOTE:

→ If each node of the decision tree makes a binary decision, the size can be as large as 2d+1-1, where 'd' is the depth.

ENTROPY:

It is a measure of the randomness in the information being processed.

↳ In other words, it refers to disorder or uncertainity.

NOTE:

* The higher the entropy, the harder it is to draw any conclusions from that data/information.

↳ Entropy in Decision Tree stands for homogeneity.

The Mathematical formulae for Entropy is

$$H(Y) = -\sum_{i=1}^{K} P_i \cdot \log_2 (P_i)$$

where

P_i - Frequentist probability of an element/class 'i' in the data.

↳ If there are 2 classes, i.e., +ve & -ve.

∴ 'i' here could be '+ve' (or) '-ve'.

INFORMATION GAIN:

It is the reduction in Entropy (or) surprise by transforming a dataset and is often used in decision trees.

↳ It is calculated by comparing the entropy of the dataset before and after a transformation.

Mathematically, it is written as

$$\boxed{IG(Y,X) = H(Y) - H(Y|X)}$$

→ We simply subtract the entropy of Y given X from the entropy of just Y to calculate the reduction of entropy.

IG(Y, X) - for variable 'Y' with respect to all features in the data.

$$\Rightarrow \boxed{IG(Y,X) = H(Y) - \Sigma \left\{ \frac{|D_i|}{|D|} * H_{D_i}(Y) \right\}}$$

* Information retrieval is a field of Information gain.

* Information Gain is the increase or decrease in Entropy value when the node is split.

* ↳ IG is the reduction of the entropy.

ENTROPY EXAMPLE:

CASE-1:

X				
Y				

For the above example, let us consider that there are around 100 data points.

CLASS < +ve / -ve. ⇒ DECISION $\begin{cases} Y(+ve) = 0 \\ Y(-ve) = 100. \end{cases}$

→ $H(Y) = 0$ since every thing is constant.

⇒ $\boxed{H(Y) = -\sum_{i=1}^{K} p_i * \log_2(p_i)}$

⇒ $H(Y) = -\{p(Y_{+ve}) \cdot \log_2 p(Y_{+ve}) + p(Y_{(-ve)}) \cdot \log_2 p(Y_{-ve})\}$

⇒ $H(Y) = -\{0 \cdot \log_2 p(0) + 100 \cdot \log_2 p(100)\}$

⇒ $H(Y) = -\{0 + 100 \cdot \log_2 p(100)\}$

⇒ $H(Y) = 0.$

↳ As there is no randomness, we get the output as '0'.

CASE - II:

$$\Rightarrow \text{Decision} \begin{cases} Y(+ve) = 50 \\ Y(-ve) = 50 \end{cases} \rightarrow H(Y) = \text{Max}$$

$$\Rightarrow H(Y) = -\sum_{i=1}^{K} p(Y_i) \cdot \log p(Y_i)$$

$$\Rightarrow H(Y) = -\left\{\frac{1}{2} * \log_2(0.5) + \frac{1}{2} * \log_2(0.5)\right\}$$

$$\Rightarrow H(Y) = -1 \times \log_2(0.5)$$

$$\Rightarrow H(Y) = -1 \times (-1)$$

$$\Rightarrow H(Y) = 1.$$

CASE - III:

$$\Rightarrow \text{Decision} \begin{cases} Y(+ve) = 100 \\ Y(-ve) = 0 \end{cases} \rightarrow H(Y) = 0$$

$$\Rightarrow H(Y) = -\sum_{i=1}^{K} p(Y_i) \cdot \log p(Y_i)$$

$$\Rightarrow H(Y) = -\{100 \cdot \log_2 p(100) + 0 \cdot \log_2 p(0)\}$$

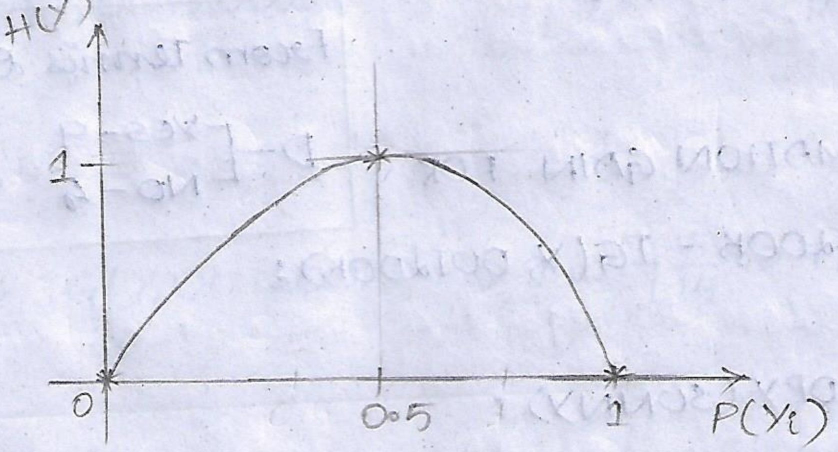
$$\Rightarrow H(Y) = -\{100 \cdot \log_2 p(100) + 0.\}$$

$$\Rightarrow H(Y) = 0.$$

So, from the above 3 cases, we can say that,

$$H(Y) \begin{cases} MAX - 1 \\ MIN - 0 \end{cases}$$

Graphically,



From the Tennis Dataset, lets find the Entropy and also the Information Gain.

FOR OUTLOOK :

OUTLOOK	YES	NO	TOTAL
SUNNY	2	3	5
OVERCAST	4	0	4
RAINY	3	2	5

(i) ENTROPY : H(Y) FOR TOTAL TENNIS GAME :

$$\Rightarrow H(Y) = -\sum_{i=1}^{K} p(Y_i) \cdot \log p(Y_i)$$

$$\Rightarrow H(Y) = -\left\{ \frac{5}{14} \log_2\left(\frac{5}{14}\right) + \frac{9}{14} \log_2\left(\frac{9}{14}\right) \right\}$$

$$\Rightarrow H(Y) = -\{(-0.530) + (-0.409)\}$$

$$\Rightarrow H(Y) = 0.940$$

From Tennis Data

$$D - \begin{cases} YES - 9 \\ NO - 5 \end{cases}$$

(ii) INFORMATION GAIN FOR OUTLOOK - IG(Y, OUTLOOK) :

(a) ENTROPY (SUNNY) :

$$= -\left\{ \frac{2}{5} \log_2\left(\frac{2}{5}\right) + \frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right) \right\}$$

$$= -\{(-0.528) + (-0.442)\}$$

$$= 0.970$$

(b) ENTROPY (OVERCAST) :

$$= -\left\{ \frac{4}{4} \cdot \log_2\left(\frac{4}{4}\right) + \cancel{\frac{0}{4} \cdot \log_2\left(\frac{0}{4}\right)}^{\,0} \right\}$$

$$= -\{(0) + (0)\}$$

$$= 0$$

(c) ENTROPY (RAINY):

$$= -\left\{\frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right) + \frac{2}{5} \cdot \log\left(\frac{2}{5}\right)\right\}$$

$$= -\{(-0.442) + (-0.528)\}$$

$$= 0.970$$

$$\Rightarrow IG(Y, OUTLOOK) = (i) - \left[\frac{5}{14} \times (a) + \frac{4}{14} \times (b) + \frac{5}{14} \times (c)\right]$$

$$\Rightarrow IG(Y, OUTLOOK) = 0.940 - \left[\frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97\right]$$

$$= 0.940 - [0.346 + 0.346]$$

$$= 0.940 - [0.692]$$

$$= 0.246$$

$$\Rightarrow IG(Y, OUTLOOK) \simeq 0.25$$

Illarly

$$\Rightarrow IG(Y, WIND) = 0.0482$$

$$\Rightarrow IG(Y, HUMIDITY) = 0.1518$$

$$\Rightarrow IG(Y, TEMPERATURE) = 0.0292$$