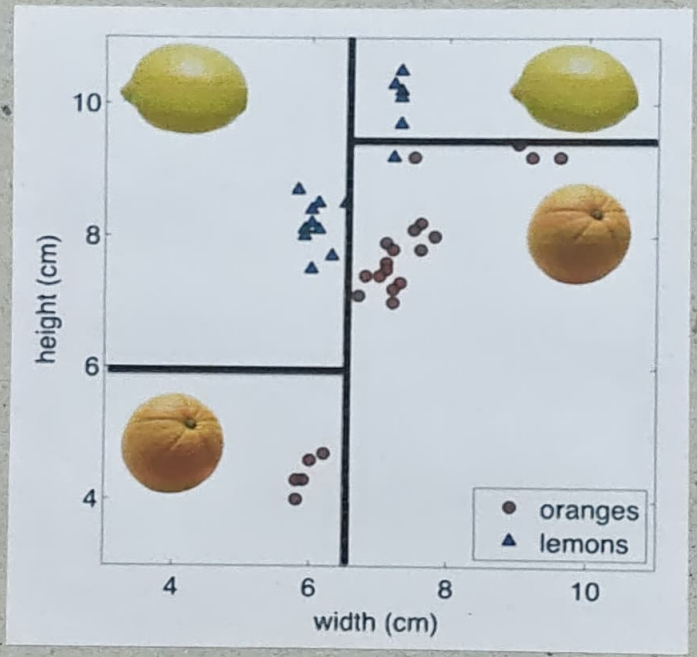
Week 6 : Session 3.

10. Decision Tree classifier (DTC)

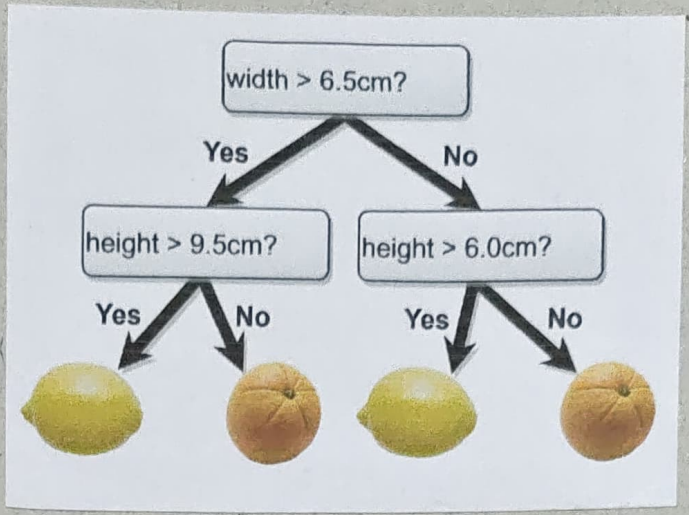
To understand DTC, let us consider the simple dataset, which comprises of Heights and Weights of Lemons and Oranges:



Now, Given a new object, based on these attributes/features, can we characterize / build Machine Learning algorithm which detects whether the object is Lemon or Orange?

Yes!!

One can build Decision Tree Classifier (DTC) as below using the given data set.



Advantages of DTC.

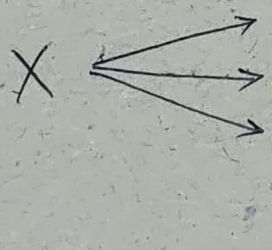
- Interpretable, Intuitive (in contrast to Neural Nets)
- Popular in Medical diagnosis applications.

DTC's are well suited to model discrete outcomes. (ii) Discrete decisions at every stage. To learn / build a DTC, one can use principles of Information theory to choose the best attribute.

(i) Which attribute to split?
Which attribute gives more information?

ENTROPY.

Entropy is the intrinsic information present in a source. Consider a source X with symbols x_i and probabilities $p(x_i)$.



The entropy $H(X)$ of this source is defined as

$$H(X) = p(x_1) \log_2\left(\frac{1}{p(x_1)}\right) + p(x_2) \log_2\left(\frac{1}{p(x_2)}\right) + \ldots$$

$$= \sum_{i=1}^{n} p(x_i) \log_2\left(\frac{1}{p(x_i)}\right)$$

$$\boxed{H(X) = -\sum_{i=1}^{n} p(x_i) \log_2\left(p(x_i)\right)}$$

This is the Entropy (or) Information content.

⊙ Entropy is always positive.

(i) Entropy ≥ 0.

⊙ Entropy is maximum when all symbols are equiprobable.

(i) $p(x_i) = \frac{1}{n}$.

⊙ Entropy is Zero, only if one of the symbols have probability 1 and rest of the symbols have probability 0.

(i) Entropy = 0 if $p(x_i) = 1$

and $p(x_j) = 0, \; j \neq i$

Remember, Sum of all the probabilities has to be 1.

(i) $\sum_{i=1}^{n} p(x_i) = 1$.

"Shannon", who is the Father of Information Theory, first came up with the concept of Entropy.

Example:

Consider the binary event of people who like or dislike Ice Cream $\overline{Ic}$, $\overline{\overline{Ic}}$.

Random variable, $X = \{Ic, \overline{Ic}\}$

Probability that people like Icecream, $P(Ic) = \frac{3}{4}$

Probability that people dont like Icecream, $P(\overline{Ic}) = 1 - P(Ic)$

$= 1 - \frac{3}{4}$

$= \frac{1}{4}$

Now, the Information / Entropy of the Random variable X is given as

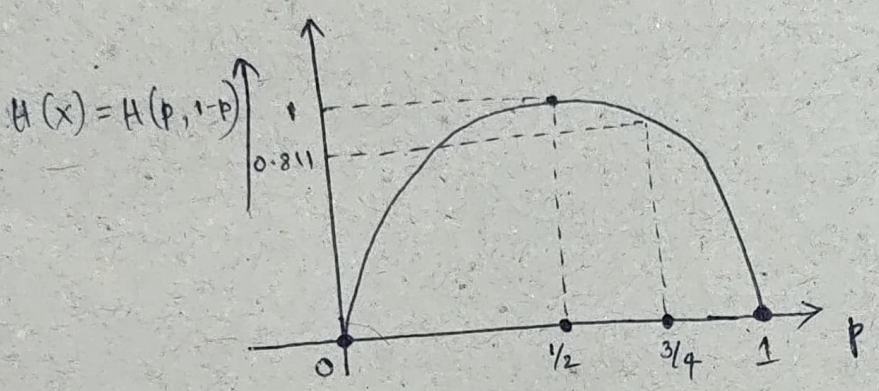
$$H(X) = H\left(P(Ic), P(\overline{Ic})\right)$$

$$= H\left(\frac{3}{4}, \frac{1}{4}\right)$$

$$= \frac{3}{4}\log_2\left(\frac{4}{3}\right) + \frac{1}{4}\log_2(4)$$

$$\boxed{H(x) = 0.811}$$

Plot the Entropy as a function of the probability of the source.



$p = 0 \Rightarrow H(0,1) = 0$

$p = 1 \Rightarrow H(1,0) = 0$

$p = \frac{1}{2} \Rightarrow H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$

$p = \frac{3}{4} \Rightarrow H\left(\frac{3}{4}, \frac{1}{4}\right) = 0.811$

The Entropy $H(x)$ reaches the peak when all the symbols have equal probability. (ii) when $p = \frac{1}{2}$, $H(x) = H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$.

CONDITIONAL ENTROPY

Consider two sources :

- ⊙ X with symbols x_i
- ⊙ Y with symbols y_i

The conditional Entropy, $H(X|Y)$

(i) The information in one Random Variable, given the other Random variable is defined as

$$\boxed{H(X|Y) = \sum_{j=1}^{m} p(y_j)\ H(X|Y = y_j).}$$

(ii) First calculate entropy of X for each value of $Y = y_i$, then take the weighted average of all these entropies, by weighing with respective probabilities $p(y_j)$ and take the sum.

For every possible value that the variable Y can take $Y \in \{y_1, y_2, \ldots, y_m\}$, we calculate the entropy of X and then we take a weighted average of all these entropies by weighing with the respective probabilities of each value of the Random variable Y.

Week 6 : Session 4

Example :

Consider the table below showing Joint Probabilities of $X = \{IC, \overline{IC}\}$, $Y = \{CHOC, \overline{CHOC}\}$.

	IC	$\overline{IC}$
$CHOC$	$1/2$	$1/8$
$\overline{CHOC}$	$1/4$	$1/8$

Eg. Probability that people likes Ice cream and don't like Chocolate.

What is the Conditional Entropy $H(X|Y)$?

$$H(X|Y) = \sum_{j=1}^{m} P(y_j)\, H(X|Y=y_j)$$

$$= P(CHOC) \times H(X|Y=CHOC) + P(\overline{CHOC}) \times H(X|Y=\overline{CHOC})$$

Entropy in X, given person likes chocolate

Entropy in X, given person don't like chocolate

- From the table, we get

$$P(CHOC) = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}$$

$$P(\overline{CHOC}) = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}$$

- Now, $H(X|Y=CHOC) = H\left(P(IC|CHOC), P(\overline{IC}|CHOC)\right)$.

$$P(IC|CHOC) = \frac{P(IC \cap CHOC)}{P(CHOC)} = \frac{1/2}{\frac{1}{2}+\frac{1}{8}} = \frac{4}{5}$$

$$P(\overline{IC}|CHOC) = \frac{P(\overline{IC} \cap CHOC)}{P(CHOC)} = \frac{1/8}{\frac{1}{2}+\frac{1}{8}} = \frac{1}{5}$$

Therefore, $H(X \mid Y = CHOC) = H\left(\frac{4}{5}, \frac{1}{5}\right)$

$= \frac{4}{5}\log_2\left(\frac{5}{4}\right) + \frac{1}{5}\log_2(5)$

$= 0.722.$

⊙ Now, $H(X \mid Y = \overline{CHOC}) = H\left(P(IC \mid \overline{CHOC}), P(\overline{IC} \mid \overline{CHOC})\right)$

$P(IC \mid \overline{CHOC}) = \frac{P(IC \cap \overline{CHOC})}{P(\overline{CHOC})} = \frac{1/4}{\frac{1}{4} + \frac{1}{8}} = \frac{2}{3}$

$P(\overline{IC} \mid \overline{CHOC}) = \frac{P(\overline{IC} \cap \overline{CHOC})}{P(\overline{CHOC})} = \frac{1/8}{\frac{1}{4} + \frac{1}{8}} = \frac{1}{3}$

Therefore, $H(X \mid Y = \overline{CHOC}) = H\left(\frac{2}{3}, \frac{1}{3}\right)$

$= \frac{2}{3}\log_2\left(\frac{3}{2}\right) + \frac{1}{3}\log_2(3)$

$= 0.918.$

Finally, the conditional Entropy is calculated as

$H(X \mid Y) = \frac{5}{8} \times 0.722 + \frac{3}{8} \times 0.918$

$$\boxed{H(X \mid Y) = 0.7955}$$

Information Gain (IG)

The Information Gain (IG) is used to choose the best attribute. The Information Gain is defined as

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

$\Rightarrow \underbrace{IG(X \mid Y)}_{\text{Information gained by observing } Y} = 0.811 - 0.7955 = 0.0155.$

This is also known as the Mutual Information (MI).

Larger the Mutual Information,

More the information Conveyed ....

Week 7 : Session 1.

DTC Feature Selection

- Choose the feature / attribute, that maximizes the Information Gain!

$$IG(X|Y) = H(X) - H(X|Y).$$

- For each attribute Y, we compute Information Gain, and choose the attribute that maximizes the Information Gain.

Example

Consider the table shown below, which says the decisions of each and every customer, to WAIT / NOT WAIT at the restaurants.

Example	Input Attributes										Goal
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
x_1	Yes	No	No	Yes	Some	$$$	No	Yes	French	0–10	y_1 = Yes
x_2	Yes	No	No	Yes	Full	$	No	No	Thai	30–60	y_2 = No
x_3	No	Yes	No	No	Some	$	No	No	Burger	0–10	y_3 = Yes
x_4	Yes	No	Yes	Yes	Full	$	Yes	No	Thai	10–30	y_4 = Yes
x_5	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	y_5 = No
x_6	No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0–10	y_6 = Yes
x_7	No	Yes	No	No	None	$	Yes	No	Burger	0–10	y_7 = No
x_8	No	No	No	Yes	Some	$$	Yes	Yes	Thai	0–10	y_8 = Yes
x_9	No	Yes	Yes	No	Full	$	Yes	No	Burger	>60	y_9 = No
x_{10}	Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10–30	y_{10} = No
x_{11}	No	No	No	No	None	$	No	No	Thai	0–10	y_{11} = No
x_{12}	Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30–60	y_{12} = Yes

The customer's decisions are based on the attributes as defined below.

1.	Alternate: whether there is a suitable alternative restaurant nearby.
2.	Bar: whether the restaurant has a comfortable bar area to wait in.
3.	Fri/Sat: true on Fridays and Saturdays.
4.	Hungry: whether we are hungry.
5.	Patrons: how many people are in the restaurant (values are None, Some, and Full).
6.	Price: the restaurant's price range ($, $$, $$$).
7.	Raining: whether it is raining outside.
8.	Reservation: whether we made a reservation.
9.	Type: the kind of restaurant (French, Italian, Thai or Burger).
10.	WaitEstimate: the wait estimated by the host (0-10 minutes, 10-30, 30-60, >60).

Out of these 10 attributes / features, which one to choose to construct the DTC ?

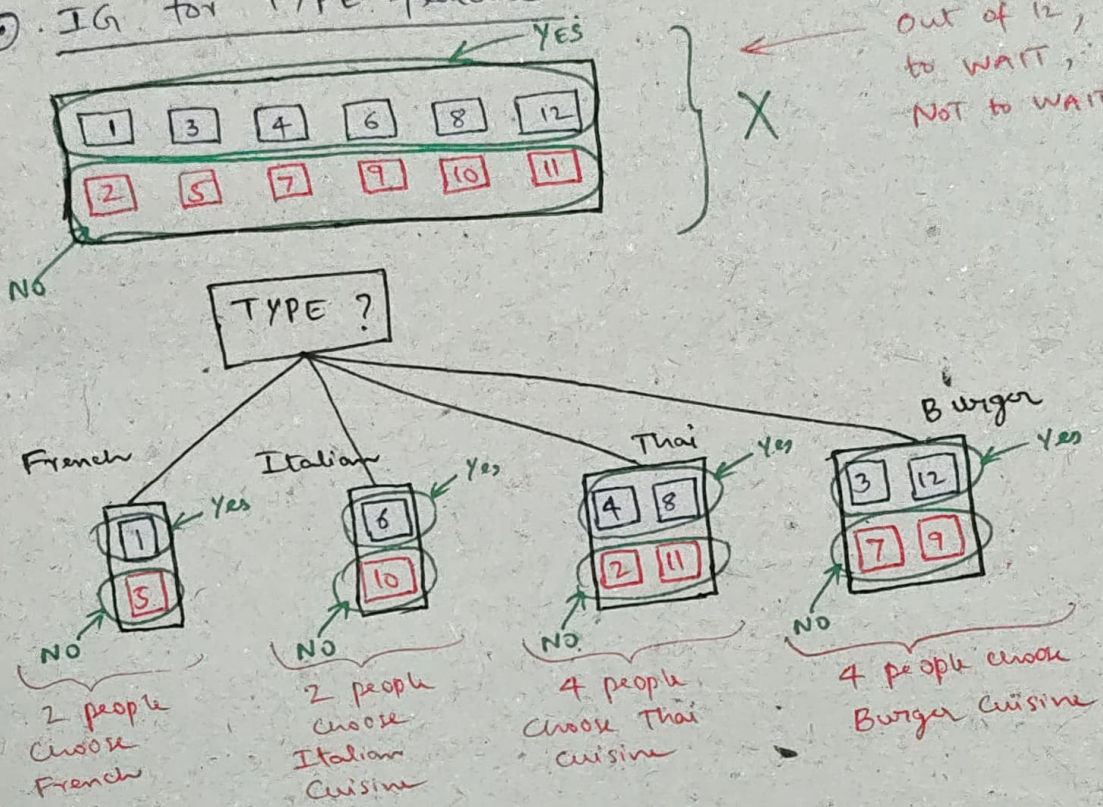
For this example, we consider the two attributes

- TYPE
- PATRONS

Out of these two, which is better for DTC ?

Naturally, we choose the one which has the maximum IG.

⊙. IG for TYPE feature



In any of the cuisines, half people choose to wait, and half people choose NOT to wait.

The IG for TYPE feature is given as

$$IG(TYPE) = H(X) - H(X \mid TYPE).$$

$$\Rightarrow H(X) = H\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{2}\log_2(2) + \frac{1}{2}\log_2(2) = 1.$$

$$\Rightarrow H(X \mid TYPE) = Pr(Fr)\, H(X \mid Fr) + Pr(It)\, H(X \mid It) + Pr(Th)\, H(X \mid Th) + Pr(Bu)\, H(X \mid Bu)$$

$$= \frac{2}{12} H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{2}{12} H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{4}{12} H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{4}{12} H\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$= \frac{2}{12} + \frac{2}{12} + \frac{4}{12} + \frac{4}{12} = 1.$$

Therefore, the IG for the TYPE feature

$$IG(TYPE) = 1 - 1 = 0$$

Inference:

What is the information that we gained, by knowing the Cuisine? NOTHING.

Coz, ⊙ Probability of persons waiting $= \frac{1}{2}$

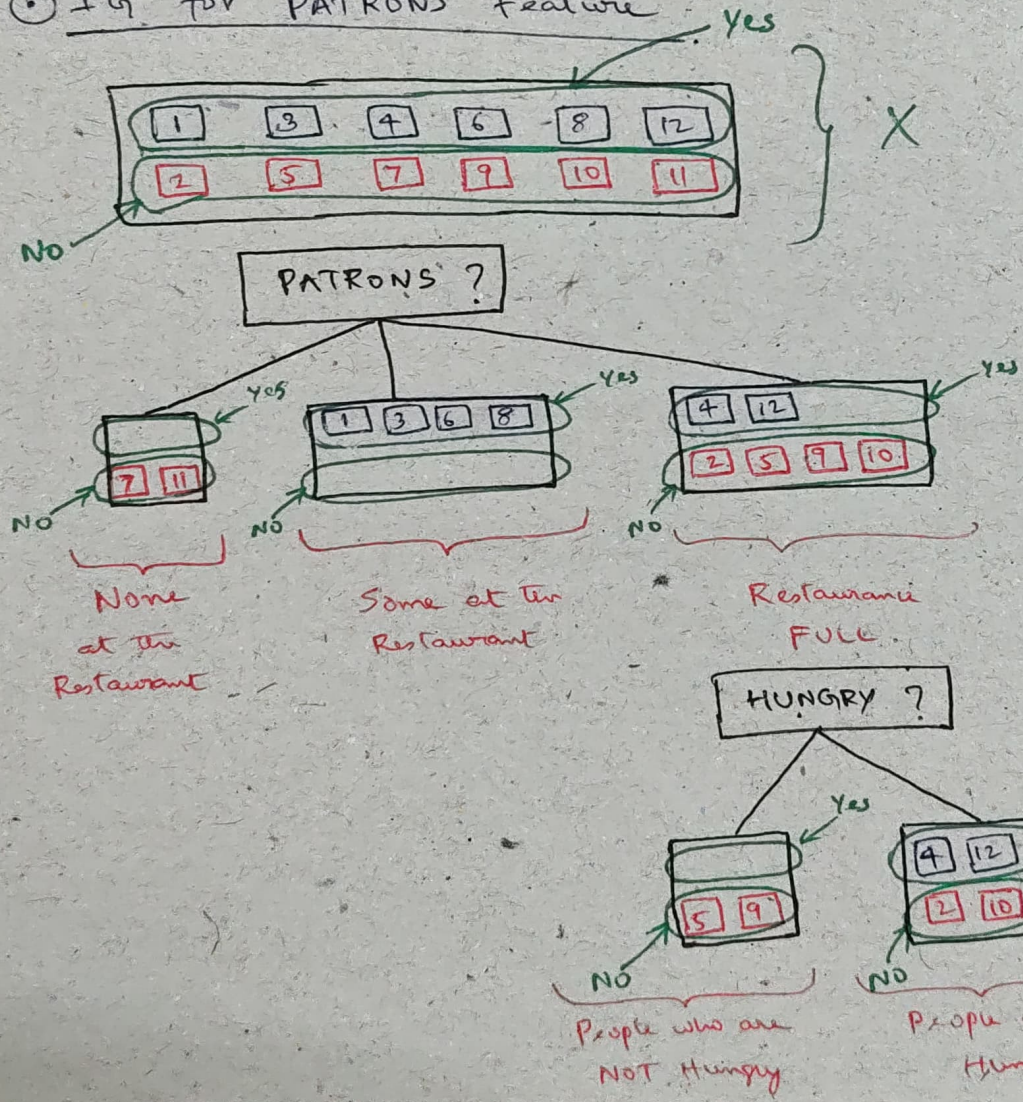
⊙ After knowing the Cuisine, Probability of persons waiting $= \frac{1}{2}$

So, knowing the Cuisine is not helping us. (ie) There is no Information Gain even if we get to know what is the Cuisine (French / Italian / Thai / Burger).

There is no additional information that the Cuisine is giving us about the decision of the customer to wait / Not.

Therefore, the TYPE feature is <u>not suitable</u> for the DTC.

⊙ IG for PATRONS feature



The IG for PATRONS feature is given as

$$IG(\text{PATRONS}) = H(x) - H(X \mid \text{PATRONS})$$

WKT, $H(x) = 1$.

Now, $H(X \mid \text{PATRONS}) = Pr(\text{None})\, H(X \mid \text{None}) + Pr(\text{Some})\, H(X \mid \text{Some}) + Pr(\text{Full})\, H(X \mid \text{Full})$

$$= \frac{2}{12} H(0,1) + \frac{4}{12} H(1,0) + \frac{6}{12} H\left(\frac{1}{3}, \frac{2}{3}\right)$$

$$= \frac{1}{2}\left(\frac{1}{3}\log_2(3) + \frac{2}{3}\log_2\left(\frac{3}{2}\right)\right)$$

$$= 0.46.$$

Therefore, the IG for the PATRONS feature

$$IG(\text{FEATURE}) = 1 - 0.46 = 0.54$$

⊙ Since $IG(\text{PATRONS}) > IG(\text{TYPE})$, we choose PATRONS as the feature to split.

⊙ Final DTC.