Decision Tree (DT)

- DTs are a type of supervised machine learning algorithms where the data is continously split according to a certain paramater/feature.
- As it is a type of supervised machine learning algorithm so we can use it for classification as well as for regression problems.
- DT can be explained by two entities, ① Decision nodes. ② Leaves nodes.
- Decision nodes represent conditions on the basis of these conditions data is splitted
- Leaves nodes are the decisions or the final outcome.
- Edges represent possible values of node.
  → In case of classification leaves represent classes and in case of regression leaves represent continous values.
- Most important attribute select as root node.

→ ID3 Algorithm

- ID stands for interactive dichotomizer (3rd version) of ID3
- The first step is to find the root node.
- It uses a special function GAIN, to evaluate the gain information of each attribute.
- For example if there are 3 attributes it will calculates the gain information for each.
- Whichever attribute has the maximum gain information becomes the root node.
- The rest of the attributes then fight for the next slots.
→ ID3 based on information gain.

Entropy

- In order to define information gain precisely, we use entropy, which characterizes the purity and impurity of an arbitrary collection of examples.
- The range of entropy is 0 to 1.
- 0 (minimum impurity) means collection of examples belong to only 1 class.
- 1 (maximum impurity) means collection of examples have equal ratio of classes
- Given a collection (S), containing +ve and -ve examples of some target concept, the entropy of S relative to this Boolean classification is:

$$\text{Entropy of } S = E(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

where P_+ is the proportion of positive class/example in S and P_- is the proportion of negative examples in S

- In all calculations involving entropy we define $0 \log 0$ to be 0.
- For example S is a collection of 14 examples of some Boolean concept/problem, including 9 positive and 5 negative examples, then the entropy of S relative to this Boolean classification is:

$$E(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.940$$

So if the collection contains unequal numbers of +ve and negative examples, the entropy is between 0 and 1.

→ Formula of Entropy for multiple classes.

$$E(S) = -\sum_{i=1}^{n} P_i \log_2 P_i$$

where n represents the number of classes

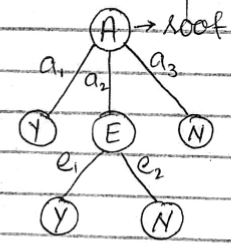
→ Information Gain

- Given entropy as a measure of the impurity in a collection of training examples, we can now define a measure of the effectiveness of an attribute in classifying the training data.
- The measure called information gain, is simply the expected reduction in entropy caused by partitioning the examples according to this attribute.
- That is, if we use the attribute with the maximum information gain as the node, then it will classify some of the instances as positive or negative with 100% accuracy, and this will reduce the entropy for the remaining instance.

→ Information gain (Example)

A	B	E	class
a_1	b_1	e_2	Yes
a_2	b_2	e_1	Yes
a_3	b_2	e_1	No
a_2	b_2	e_2	No
a_3	b_1	e_2	No

Data



Decision Tree.

Steps.

① calculate the entropy of entire data set (S)

$$E(S) = -\sum_{i=1}^{n} P_i \log_2 P_i$$

$$= -P_Y \log_2 P_Y - P_N \log_2 P_N$$

$$= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \quad \because P_Y = 2/5 \quad P_N = 3/5$$

$$= 0.97$$

② calculate the information gain for each attribute

- For attribute A

$$G(S, A) = E(S) - \frac{|S_{a_1}|}{|S|} E(S_{a_1}) - \frac{|S_{a_2}|}{|S|} E(S_{a_2}) - \frac{|S_{a_3}|}{|S|} E(S_{a_3})$$

As $E(S) = 0.97$, $|S| = 5$, $|S_{a_1}| = 1$, $|S_{a_2}| = 2$, $|S_{a_3}| = 2$

$$E(S_{a_1}) = -P_Y \log_2 P_Y - P_N \log_2 P_N$$
$$= -\frac{1}{1} \log_2 \frac{1}{1} - 0 \log_2 0 = 0$$

Similarly

$$E(S_{a_2}) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

and

$$E(S_{a_3}) = -0 \log_2 0 - \frac{2}{2} \log_2 \frac{2}{2} = 0$$

So

$$G(S, A) = 0.97 - \frac{1}{5} \times 0 - \frac{2}{5} \times 1 - \frac{2}{5} \times 0 = 0.57$$

- For attribute B

$$G(S, B) = E(S) - \frac{|S_{b_1}|}{|S|} E(S_{b_1}) - \frac{|S_{b_2}|}{|S|} E(S_{b_2})$$
$$= 0.97 - \frac{2}{5} \times 1 - \frac{3}{5} \times E(S_{b_2})$$

$$E(S_{b_2}) = -P_Y \log_2 P_Y - P_N \log_2 P_N$$
$$= -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.91$$

So

$$G(S, B) = 0.97 - \frac{2}{5} \times 1 - \frac{3}{5} \times 0.91 = 0.024$$

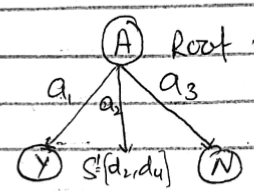
- For attribute E

$$G(S, E) = E(S) - \frac{|S_{e_1}|}{|S|} E(S_{e_1}) - \frac{|S_{e_2}|}{|S|} E(S_{e_2})$$
$$= 0.97 - \frac{2}{5} \times 1 - \frac{3}{5} \times E(S_{e_2})$$

$$E(S_{e_2}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.91$$

$$= 0.97 - \frac{2}{5} \times 1 - \frac{3}{5} \times 0.91 = 0.024$$

③ Select the attribute that has maximum Gain and define splitting criteria



④ Repeat all steps for sub-data set

A	B	E	class
a_2	b_2	e_1	Yes
a_2	b_2	e_2	No.

S'

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

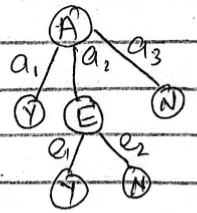
$$G_r(S', B) = E(S') - \frac{|S'b_2|}{|S'|} E(S'b_2)$$

$$= 1 - \frac{2}{2} \times 1 = 0.$$

$$G_r(S', E) = E(S') - \frac{|S'e_1|}{|S'|} E(S'e_1) - \frac{|S'e_2|}{|S'|} E(S'e_2)$$

$$= 1 - \frac{1}{2} \times 0 - \frac{1}{2} \times 0 = 1$$

So



Practice - 1.

Day(D)	Weather(W)	Temperature(T)	Humidity(H)	Wind(W)	Play? (P)
1	Sunny	Hot	High	weak	No
2	cloudy	Hot	High	weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	cool	Normal	Strong	No
7	Rainy	Mild	High	weak	Yes
8	Sunny	Hot	High	Strong	No
9	cloudy	Hot	Normal	weak	Yes
10	Rainy	Mild	High	Strong	No

Dataset (S)

Solution.

① Find entropy of entire dataset(S).

$$E(S) = -P_Y \log_2 P_Y - P_N \log_2 P_N$$
$$= -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1.$$

② Select the attribute that has maximum gain.

For weather(W)

$$G_1(S,W) = E(S) - \frac{|SW_{sunny}|}{|S|}E(SW_{sunny}) - \frac{|SW_{cloudy}|}{|S|}E(SW_{cloudy}) - \frac{|SW_{Rainy}|}{|S|}E(SW_{Rainy})$$

$$= 1 - \frac{3}{10}\times\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{3}{10}\times\left(-\frac{3}{3}\log_2\frac{3}{3} - 0\log_2 0\right) - \frac{4}{10}\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right)$$

$$= 1 - 0.3\times 0.91 - 0.3\times 0 - 0.4\times 0.81 = 0.40.$$

For Temperature(T)

$$G_1(S,T) = E(S) - \frac{|ST_{Hot}|}{|S|}E(ST_{Hot}) - \frac{|ST_{mild}|}{|S|}E(ST_{mild}) - \frac{|ST_{cool}|}{|S|}E(ST_{cool})$$

$$= 1 - \frac{4}{10}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) - \frac{5}{10}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$$

$$- \frac{1}{10}\left(-0\log_2 0 - \frac{1}{1}\log_2\frac{1}{1}\right)$$

$$= 1 - 0.4\times 1 - 0.5\times 0.97 - 0.1\times 0 = 0.11$$

For Humidity(H)

$$G_1(S,H) = E(S) - \frac{|SH_{High}|}{|S|}E(SH_{High}) - \frac{|SH_{normal}|}{|S|}E(SH_{normal})$$

$$= 1 - \frac{7}{10}\left(-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right) - \frac{3}{10}\left(\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)$$

$$= 1 - 0.7\times 0.98 - 0.3\times 0.91 = 0.04$$

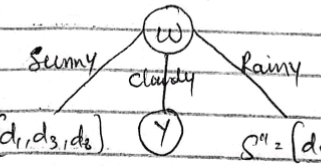
For Wind(Wd)

$$G(S,WD) = E(S) - \frac{|SWD_{weak}|}{|S|}E(SWD_{weak}) - \frac{|SWD_{strong}|}{|S|}E(SWD_{strong})$$

$$= 1 - \frac{4}{10}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) - \frac{6}{10}\left(-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right)$$

$$= 1 - 0.4\times 0.81 - 0.6\times 0.91 = 0.13$$

③ Select Attribute that has maximum gain as root node.



④ Repeat 1 to 3 steps for S' and S''

$$E(S') = -P_Y\log_2 P_Y - P_N\log_2 P_N$$

$$E(S') = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.91$$

$$G(S', T) = E(S') - \frac{|S'T_{Hot}|}{|S'|} E(S'T_{Hot}) - \frac{|S'T_{mild}|}{|S'|} E(S'T_{mild})$$

$$= 0.91 - \frac{2}{3}\left(-0\log_2 0 - \frac{2}{2}\log_2\frac{2}{2}\right) - \frac{1}{3}\left(-\frac{1}{1}\log_2\frac{1}{1} - 0\log_2 0\right)$$

$$= 0.91 - 0.67 \times 0 - \frac{1}{3} \times 0 = 0.91$$

$$G(S', H) = E(S') - \frac{|S'H_{High}|}{|S'|} E(S'H_{High}) - \frac{|S'H_{Normal}|}{|S'|} E(S'H_{Normal})$$

$$= 0.91 - \frac{2}{3}\left(-0\log_2 0 - \frac{2}{2}\log_2\frac{2}{2}\right) - \frac{1}{3}\left(-\frac{1}{1}\log_2\frac{1}{1} - 0\log_2 0\right)$$

$$= 0.91 - 0.67 \times 0 - 0.33 \times 0 = 0.91$$

$$G(S', WD) = E(S') - \frac{|S'WD_{weak}|}{|S'|} E(S'WD_{weak}) - \frac{|S'WD_{strong}|}{|S'|} E(S'WD_{strong})$$

$$= 0.91 - \frac{1}{3}\left(-0\log_2 0 - \frac{1}{1}\log_2\frac{1}{1}\right) - \frac{2}{3}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

$$= 0.91 - 0.33 \times 0 - 0.67 \times 1 = 0.24.$$

FOR S''

$$E(S'') = -P_Y\log_2 P_Y - P_N\log_2 P_N$$

$$= -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$

$$= 0.81$$

$$G(S'', T) = E(S'') - \frac{|S''T_{Mild}|}{|S''|} E(S''T_{Mild}) - \frac{|S''T_{Cool}|}{|S''|} E(S''T_{Cool})$$

$$= 0.81 - \frac{3}{4}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{1}{4}\left(-0\log_2 0 - \frac{1}{1}\log_2\frac{1}{1}\right)$$

$$= 0.81 - 0.75 \times 0.91 - 0.25 \times 0. = 0.12$$

$$G(S'', H) = E(S'') - \frac{|S''H_{High}|}{|S''|} E(S''H_{High}) - \frac{|S''H_{Normal}|}{|S''|} E(S''H_{Normal})$$

$$= 0.81 - \frac{3}{4}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{1}{4}\left(-0\log_2 0 - \frac{1}{1}\log_2\frac{1}{1}\right)$$

$$= 0.81 - 0.75 \times 0.91 - 0.25 \times 0 = 0.12$$

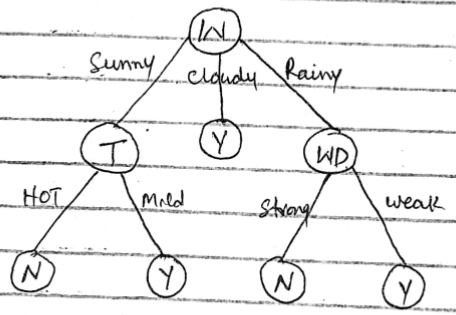
$$G(S'', WD) = E(S'') - \frac{|S''_{WD_{strong}}|}{|S''|} E(S''_{WD_{strong}}) - \frac{|S''_{WD_{weak}}|}{|S''|} E(S''_{WD_{weak}})$$

$$= 0.81 - \frac{3}{4}\left(-0\log_2 0 - \frac{3}{3}\log_2 \frac{3}{3}\right) - \frac{1}{4}\left(-\frac{1}{1}\log_2 \frac{1}{1} - 0\log_2 0\right)$$

$$= 0.81 - 0.75 \times 0 - 0.25 \times 0 = 0.81$$

As we see that

$G_1(S', T)$ and $G(S', H)$ have maximum equal gain so can select randomly.



Note:- If features are continous, internal nodes can test the value of a feature against a threshold