Dataset : Decision Tree Example: ID3

S.NO	color	Attribute outline	dot	shape
1	green	dashed	NO	triangle
2	green	dashed	yes	triangle
3	yellow	dashed	NO	Square
4	red	dashed	NO	Square
5	red	Solid	NO	Square
6	red	Solid	yes	triangle
7	green	Solid	NO	Square
8	green	dashed	NO	triangle
9	yellow	Solid	yes	Square
10	red	solid	NO	Square
11	green	Solid	yes	Square
12	yellow	dashed	yes	Square
13	yellow	Solid	NO	Square
14	red	dashed	yes	triangle

Entropy formula :

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

Information gain formula:

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

1. We need to get the attribute which is giving maximum information among all the attributes. Hence we need to calculate Information gain for every attribute. After this we need to identify which attribute has highest information gain. Then we will take that attribute for tree construction.

We are planning to calculate information gain each attribute.

The attributes in the dataset are:

1. Color
2. outline
3. Dot

Attribute: Color

possible values of attribute Color: {green, yellow, red}

Attribute: outline

possible values of attribute outline: {dashed, solid}

Attribute: Dot

possible values of attribute Dot: {No, Yes}

We need to calculate entropy of entire dataset & entropy of individual attribute of dataset then Info gain.

Entropy of the dataset:

$$\text{Entropy}(S)^{\text{dataset}} = \sum_{i=1}^{c} -p_i \log_2(p_i)$$

In the dataset, we find 5- triangles, & 9 squares

$$\Rightarrow S = [9\square, 5\triangle] \text{ out of 14 records in dataset}$$

$$\text{Entropy}(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= 0.94$$

Similarly we need to compute Entropy of green, yellow, red for Color attribute.

⇒ total records related to green in color attribute
= 5.

Out of them 5 records, 3 triangle & 2 squares:

⇒ $S_{green} \leftarrow [2\square, 3\Delta]$

⇒ $Entropy(S_{green}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$

$= 0.971$

⇒ total records related to yellow in color attribute.
= 4

Out of them 4 records, 4 squares & 0 - triangle.

$S_{yellow} \leftarrow [4\square, 0\Delta]$

$Entropy(S_{yellow}) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)$

we need to calculated. $\log_2$ is zero only.

$= 0$

⇒ total records related to red in color attribute,

$S_{Red} \leftarrow [3\square, 2\Delta]$ ⇒ total ⇒ (5)

$Entropy(S_{Red}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$

$= 0.971$

We need to find:

Now, Information gain of color attribute.

$Gain(S, color) = Entropy(S) - \sum \frac{|S_v|}{|S|}$

$$Gain(S, Color) = Entropy(S) - \sum_{v \in \{green, red, yellow\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$S_v \rightarrow$ The NO. of possible values of 'v"

$$Gain(S, Color) = Entropy(S) - \frac{5}{14} Entropy(S_{green}) - \frac{4}{14} Entropy(S_{yellow}) - \frac{5}{14} Entropy(S_{Red})$$

$$\Rightarrow Gain(S, color) = 0.94 - \frac{5}{14} \times 0.971 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.971$$

$$= 0.2464.$$

Similarly, we need to calculate Information gain for the remaining attributes, also.

Attribute: outline = {dashed, solid}

$S = [9\square, 5\triangle]$ $\quad Entropy(S) = 0.94$

$S_{dashed} \leftarrow [3\square, 4\triangle]$ $\quad$ total = 7 - records.

$$Entropy(S_{dashed}) = -\frac{3}{7} \log_2(3/7) - \frac{4}{7} \log_2(4/7)$$

$$= 0.9852$$

$S_{solid} \leftarrow [6\square, 1\triangle]$ $\quad$ total = 7 - records.

$$Entropy(S_{solid}) = -\frac{6}{7} \log_2(6/7) - \frac{1}{7} \log_2(1/7)$$

$$= 0.5916$$

$$\Rightarrow Gain(S, outline) = Entropy(S) - \sum_{v \in \{dashed, solid\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, outline) = Entropy(S) - \frac{7}{14} Entropy(S_{dashed}) - \frac{7}{14}$$

$$Gain(S, Color) = Entropy(S) - \sum_{v \in \{green, Red, yellow\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$S_v \rightarrow$ The No. of possible values of 'v'

$$Gain(S, Color) = Entropy(S) - \frac{5}{14} Entropy(S_{green}) - \frac{4}{14} Entropy(S_{yellow}) - \frac{5}{14} Entropy(S_{Red})$$

$$\Rightarrow Gain(S, Color) = 0.94 - \frac{5}{14} \times 0.971 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.971$$

$$= 0.2464.$$

Similarly, we need to calculate Information gain for the remaining attributes, also.

Attribute: outline = {dashed, solid}

$S = [9\square, 5\triangle]$ $\quad Entropy(S) = 0.94$

$S_{dashed} \leftarrow [3\square, 4\triangle]$ $\quad$ total = 7 - records.

$$Entropy(S_{dashed}) = -\frac{3}{7} \log_2(3/7) - \frac{4}{7} \log_2(4/7)$$

$$= 0.9852$$

$S_{solid} \leftarrow [6\square, 1\triangle]$ $\quad$ total = 7 - records.

$$Entropy(S_{solid}) = -\frac{6}{7} \log_2(6/7) - \frac{1}{7} \log_2(1/7)$$

$$= 0.5916$$

$$\Rightarrow Gain(S, outline) = Entropy(S) - \sum_{v \in \{dashed, solid\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, outline) = Entropy(S) - \frac{7}{14} Entropy(S_{dashed}) - \frac{7}{14} Entropy(S_{solid})$$

$$Gain(S, outline) = 0.94 - \frac{7}{14} \times 0.9852 - \frac{7}{14} \times 0.5916$$

$$= 0.1516$$

Attribute: ~~Dot~~ = {NO, Yes}

$S = [9\square, 5\triangle]$ Entropy $(S) = 0.94$

$S_{Yes} \leftarrow [3\square; 3\triangle]$ total = 6-records.

$$Entropy(S_{yes}) = \frac{-3}{6} \log_2\left(\frac{3}{6}\right) - \frac{3}{6} \log\left(\frac{3}{6}\right)$$

$$= 1$$

$S_{NO} \leftarrow [6\square, 2\triangle]$ total = 8-records.

$$Entropy\ (S_{NO}) = -\frac{6}{8} \log_2\left(\frac{6}{8}\right) - \frac{2}{8} \log_2\left(\frac{2}{8}\right)$$

$$= 0.8113$$

$$Gain\ (S, Dot) = Entropy(S) - \sum_{v \in \{Yes, No\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= 0.94 - \frac{6}{14} \times 1.0 - \frac{8}{14} \times 0.8113$$

$$= 0.0478$$

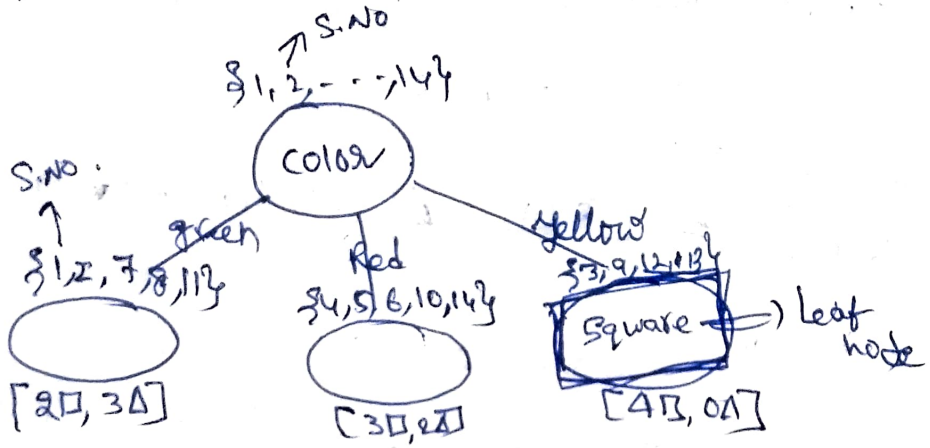
$Gain\ (S, color) = 0.2464$

$Gain\ (S, outline) = 0.1516$

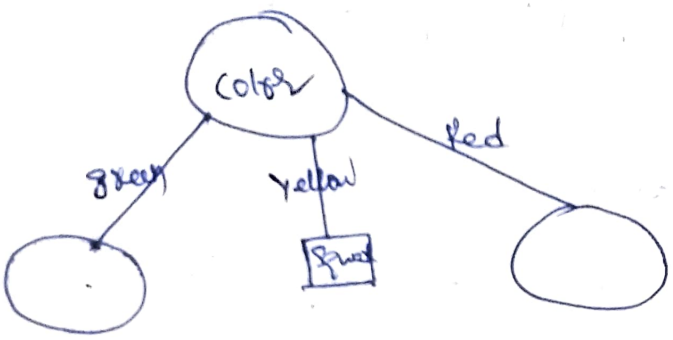
$Gain\ (S, Dot) = 0.0478.$

Here we can conclude that color attribute has highest gain. So, we consider color as root node.

The possibilities of color are green, Red, and yellow.



(ii)



We need to compute the feature. The tree of green & Red. Whue as Yellow become the leaf node.

Now we will start @ green branch side.

Now we need to take the examples of records whichare ~~records~~ related to green only. i.e S.NO {1, 2, 7, 8, 11}.

S.NO	outline	Dot	shape
1	dashed	NO	triangle
2	dashed	yes	triangle
7	solid	NO	Square
8	dashed	NO	triangle
11	solid	yes	Square

$S_{1(\text{green})} = [2□, 3△]$ total 5-records.

Choose attribute outline

$$\text{Entropy}(S_1) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= 0.97$$

Attribute : outline

S_1(dashed) $\leftarrow$ [0□, 3△] $\rightarrow$ total — 3 — records.

$$\text{Entropy}(S_1(\text{dashed})) = 0.0$$

S_1(solid) $\leftarrow$ [2□, 0△] $\rightarrow$ total — 2 — records.

$$\text{Entropy}[S_1(\text{solid})] = 0.$$

~~Gain (S_1, green)~~

$$\text{Gain}(S_1, \text{outline}) = \text{Entropy}(S_1) - \sum_{v \in \{\text{dashed, solid}\}} \frac{|S_v|}{|S_1|}\text{Entropy}(S_v)$$

$$\text{Gain}(S_1, \text{outline}) = 0.97 - \frac{3}{5}\times 0 - \frac{2}{5}\times 0$$

$$= 0.97$$

Attribute : dot $\{$NO, Yes$\}$

S_1(NO) $\leftarrow$ [1□, 2△] $\rightarrow$ total — 3 — records

$$\text{Entropy}(S_1(\text{NO})) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$$

$$= 0.9183$$

S_1(Yes) $\leftarrow$ [1□, 1△] $\rightarrow$ total 2 — records.

$$\text{Entropy}(S_1(\text{Yes})) = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right).$$

$$= 1.$$

$$gain(S_1, dot) = Entropy(S_1) - \sum_{v \in \{yes, no\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= 0.97 - \frac{2}{5} \times 1.0 - \frac{3}{5} \times 0.918$$

$$= 0.0192$$

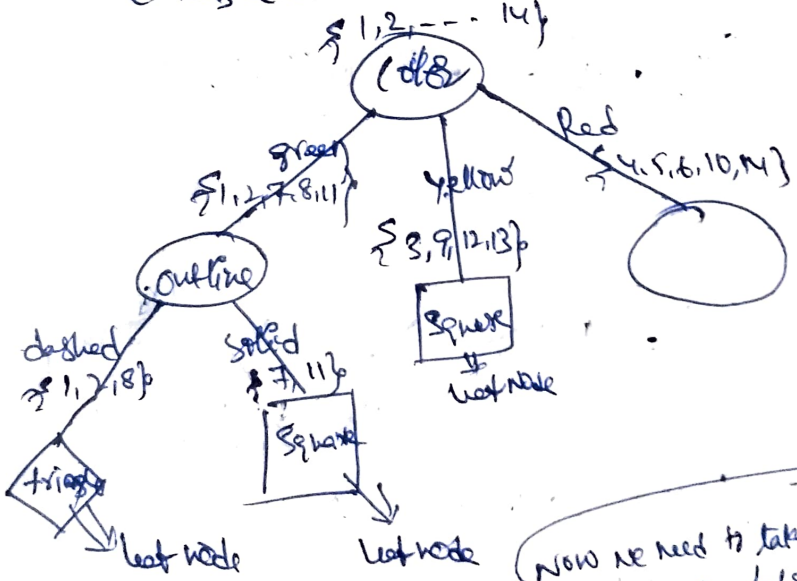
finally

$$gain(S_1, outline) = 0.97$$

$$gain(S_1, dot) = 0.0192$$

outline has highest Information gain, hence it can be taken as node at this level.

@ this level tree will be



Now we need to take examples which are scaled to Red one

S.NO	outline	DOT	shape
4	dashed	NO	Square
5	Solid	NO	square
6	Solid	Yes	Triangle
10	Solid	NO	square
14	dashed	Yes	Triangle

treat this as S_2

$S_2 = [3\square, 2\triangle]$ ⇒ total 5 records.

$$\text{Entropy}(S_2) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$

$$= 0.97$$

Attribute: outline ⇒ {dashed, solid}

$S_2(\text{dashed}) \leftarrow [1\square, 1\triangle]$ ⇒ 2 records

$$\text{Entropy}(S_2(\text{dashed})) = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right)$$

$$= 1$$

$S_2(\text{solid}) \leftarrow [2\square, 1\triangle]$ ⇒ 3 records

$$\text{Entropy}(S_2(\text{solid})) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$$

$$= 0.9183$$

$$\text{Gain}(S_2, \text{outline}) = \text{Entropy}(S_1) - \sum_{v \in \{\text{dashed, solid}\}} \frac{|S_v|}{|S_1|} \text{Entropy}(S_v)$$

$$= 0.97 - \frac{2}{5} \times 1.0 - \frac{3}{5} \times 0.918$$

$$= 0.0192$$

Attribute: dot ⇒ {NO, yes}.

$S_2(\text{NO}) \leftarrow [3\square, 0\triangle]$ ⇒ 3 records.

$$\text{Entropy}(S_2(\text{NO})) = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)$$

$$= 0$$

$S_2(\text{yes}) \leftarrow [0\square, 2\triangle]$

~~$S_2(yes) \leftarrow [3+, 3-]$~~

$S_2(yes) \leftarrow [0+, 2-]$ total − 2 records

$Entropy(S_2(yes)) = 0.$

$$gain(S_2, dot) = Entropy(S_2) - \sum_{v \in \{yes, no\}} \frac{|S_v|}{|S_2|} Entropy(S_v)$$

$= 0.97$

$= 0.97 - \frac{2}{5} * 0.0 - \frac{3}{5} \times 0. = 0.97$

finally

$gain(S_2, outline) = 0.0192$

$gain(S_2, dot) = 0.97$

As "dot" attribute has highest information gain, it can be taken as node at this stage.

final tree