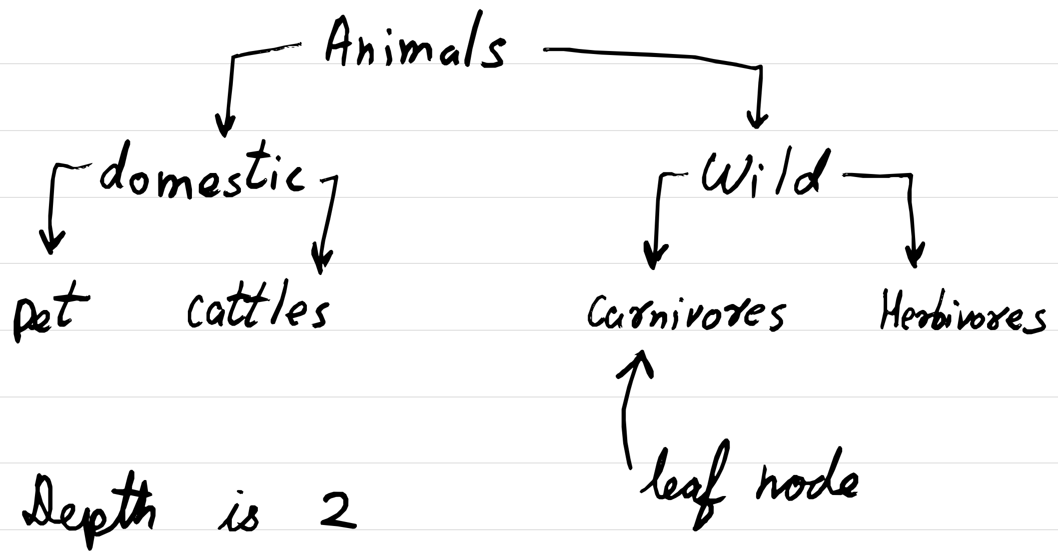
Decision Tree

* Used for classification

A dataset can be classified by various parameters. Decision Trees are used to select such parameter



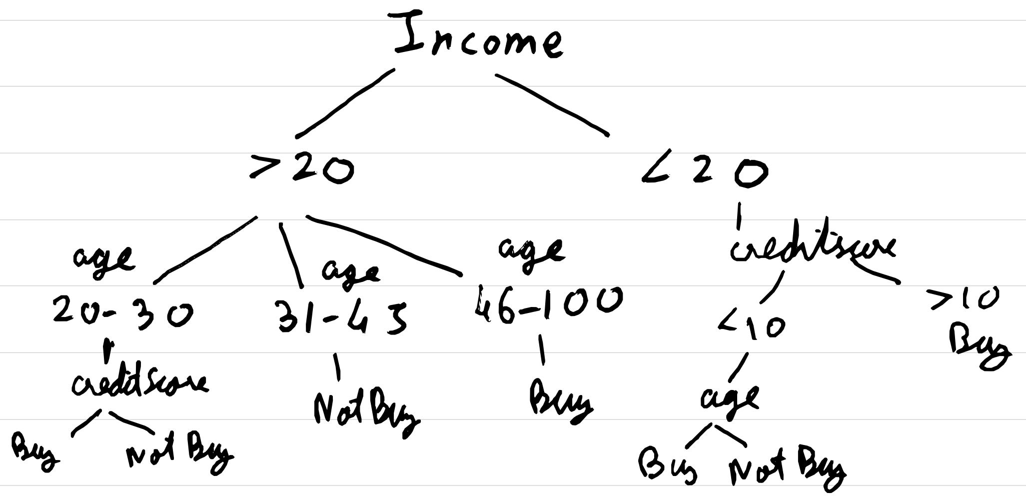
Depth is 2

Example if below data is provided to machine. Machine has to decide if person can buy BMW or not

Income	age	credit Score	buy BMW
20	40	100	Y
10	25	20	N
30	45	90	Y
12	45	60	N
40	46	80	Y
40	46	80	Y

Machine has to decide which creterion to select for classification

eg one such creiteron can be



Machine uses entropy & gain for classification to decide which is the best creterion for such a classification

There are huge number of calculation for this model.

Aim of the model is to decide which tree is the best.

If there are multiple parameters then calculations increase rapidly.

Decision tree & CART both are Top down Non backtracking greedy approaches.

Recursive Partitioning algorithms

Entropy

Entropy is randomness or "impurity" of an attribute

$$H = -\sum_{i=1}^{n} P(x_i) \log_2 (x_i)$$

when $P(x) = 0.5$, $H = 1$

① calculate entropy for Gender

Gender	count
Male	9
Female	5

$\longrightarrow$ $P(\text{male}) = \frac{9}{14}$

$P(\text{Female}) = \frac{5}{14}$

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

$$= -0.94$$

Entropy is 0 if outcome is certain

Entropy is maximum if we have no knowledge of the system, or if any outcome is equally Possible

Information gain

Information is in form of entropy.

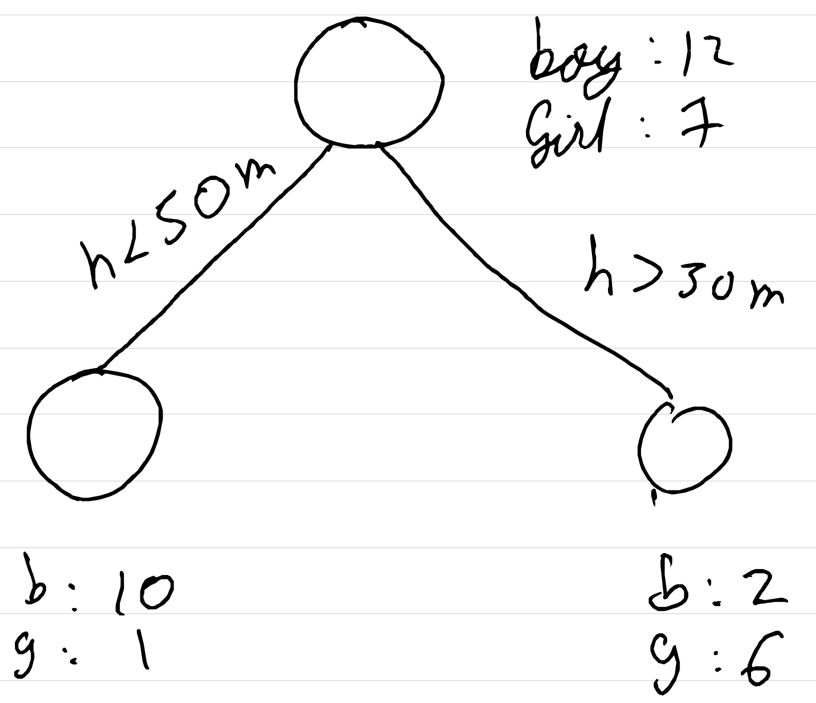
Info gain = Info Parent − Info attribute

Info Parent → H(output)

Info attribute → weighted sum of H of all unique value of that attribute (feature)

Calculate Info gain for split of height

Target → gender



Info parent = H(Parent)

$$= \frac{12}{19}\log\left(\frac{12}{19}\right) + \frac{7}{19}\log\frac{7}{19}$$

$$= -0.949$$

h < 50	h > 50
$H = \frac{10}{11}\log\left(\frac{10}{11}\right) + \frac{1}{11}\log\frac{1}{11}$	$H = \frac{2}{8}\log\left(\frac{2}{8}\right) + \frac{6}{8}\log\frac{6}{8}$
$= +0.439$	$= +0.811$

Info split $= \frac{11}{19}(0.439) + \frac{8}{19}(0.811)$

$$= +0.5956$$

Info gain $= 0.949 - 0.5936$

$= 0.3533$

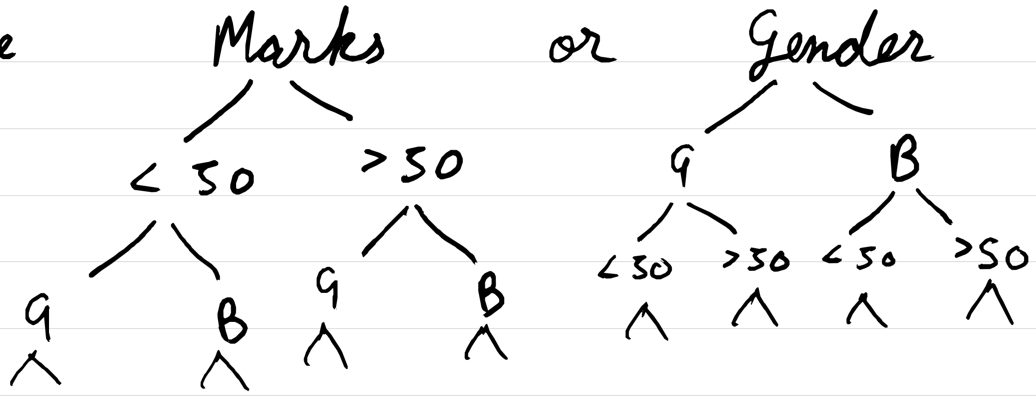
Decision tree

which attribute would you decide to split the dataset into a decision tree

Gender	Marks
G	65
G	46
B	56
B	43
B	53
B	49
G	42
B	84
B	44
G	42
G	40

should data be split on the basis of gender or marks?

ie



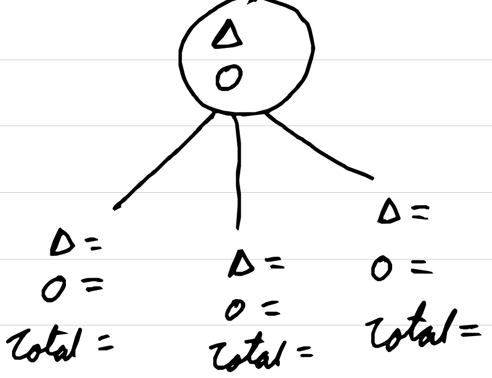
Theese decisions will be taken by the decision tree based on the info gain

Split with the highest info gain is chosen

Decision tree algorithm

$H = \sum P \log_2 P$

step 1. find H of Parent
step 2. Select an attribute &
split data in that attribute
. count the number of Target classes
for the attribute



step 3. Calculate entropy for every split
step 4. Entropy for the attribute is the
weighted sum of the attributes $(\sum P \cdot H)$
step 5. Calculate Information gain
step 6. Repeat steps 2-4 for all attributes
Select lowest value of entropy ie the
highest value of info gain
step 7. Split the dataset based on the
parameter into parts
step 8. Repeat the process for all parts
note → dataset will be smaller for a part
Exclude the attribute for the part

Generally entropy values will be in rang 0.8 - 0.99
or 0 or 1

Dont forget to take $\log_2$

The lower the value of H, that means
more accurate or skewed the split

eg If split is like 6 : 4 H will be towards 1
9 : 1 H will be Lesser

Also $0 \log_2 0 = 0$ consider

Class labelled Training tuples

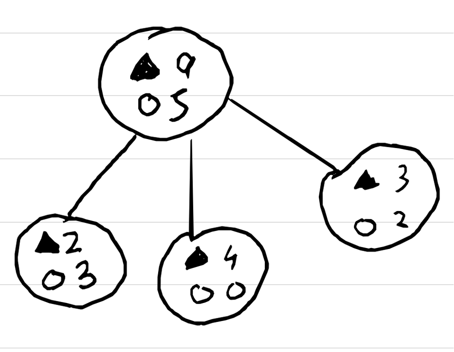
id	age	income	student	crediting	buy Car
1	Youth	high	no	fair	no
2	y	h	no	excellent	no
3	m-a	h	no	f	y
4	s	medium	no	f	y
5	s	low	yes	f	y
6	s	L	yes	e	n
7	m-a	L	yes	e	y
8	y	m	no	f	n
9	y	L	y	f	y
10	s	m	y	f	y
11	y	m	y	e	y
12	ma	m	n	e	y
13	ma	h	y	f	y
14	s	m	n	e	n

info (D) on the target value

$$info(D) = -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right)$$

$= -0.940$

$info(D)_{age}$ → Trying root, age



age

youth 5 — y: 2, n: 3

middle age 4 — y: 4, n: 0

senior 5 — y: 3, n: 2

$$H(\text{youth}) = \frac{5}{14} \times -\left(\cdot\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right)$$

↑ weighted sum

$$H(\text{middle}) = \frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4} - 0\log_2 0\right)$$

$= 0$

$$H(\text{Senior}) = \frac{5}{14}\left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right)$$

$H(\text{age}) = H_1 + H_2 + H_3 = 0.693$

Info gain $= 0.940 - 0.693 = \underline{0.247}$

for Income

Income

high 4 — y: 2, h: 2

medium 6 — y: 4, h: 2

low 4 — y: 3, n: 1

$$H = \frac{4}{14}\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) \rightarrow \text{for high}$$

$$+ \frac{6}{14}\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) \rightarrow \text{for medium}$$

$$+ \frac{4}{14}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) \rightarrow \text{for low}$$

$= 0.911$

Info gain $= 0.940 - 0.911 = \underline{\underline{0.029}}$

for student

student

no (7) — y: 3, n: 4

yes (7) — y: 6, n: 1

$$H = \frac{7}{14}\left(-\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}\right)$$

$$+ \frac{7}{14}\left(-\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7}\right)$$

$= 0.788$

Info gain $= 0.940 - 0.788$
$= 0.15$

for credit score

credit

fair 8 — y: 6, n: 2

excellent 6 — y: 3, n: 3

$$H = \frac{8}{14}\left(\frac{6}{8}\log\frac{6}{8} + \frac{2}{8}\log\frac{2}{8}\right)$$

$$+ \frac{6}{14}\left(\frac{3}{6}\log\frac{3}{6} + \frac{3}{6}\log\frac{3}{6}\right)$$

$= 0.892$

Info gain $= 0.94 - 0.892 = 0.048$

Information Gain

age → 0.247

Income → 0.029

student → 0.15

credit → 0.048

Highest value is by age. Hence age is root node

age

youth: 1 2 3 4 5

middle age: 1 2 3 4

senior: 1 2 3 4 5

resolving problem recursively after removing age parameter & breaking into 3 parts

*

Value was 0 ∴ leaf

when there is 0 distribution for a node, leaf is found max info gain found

eg student

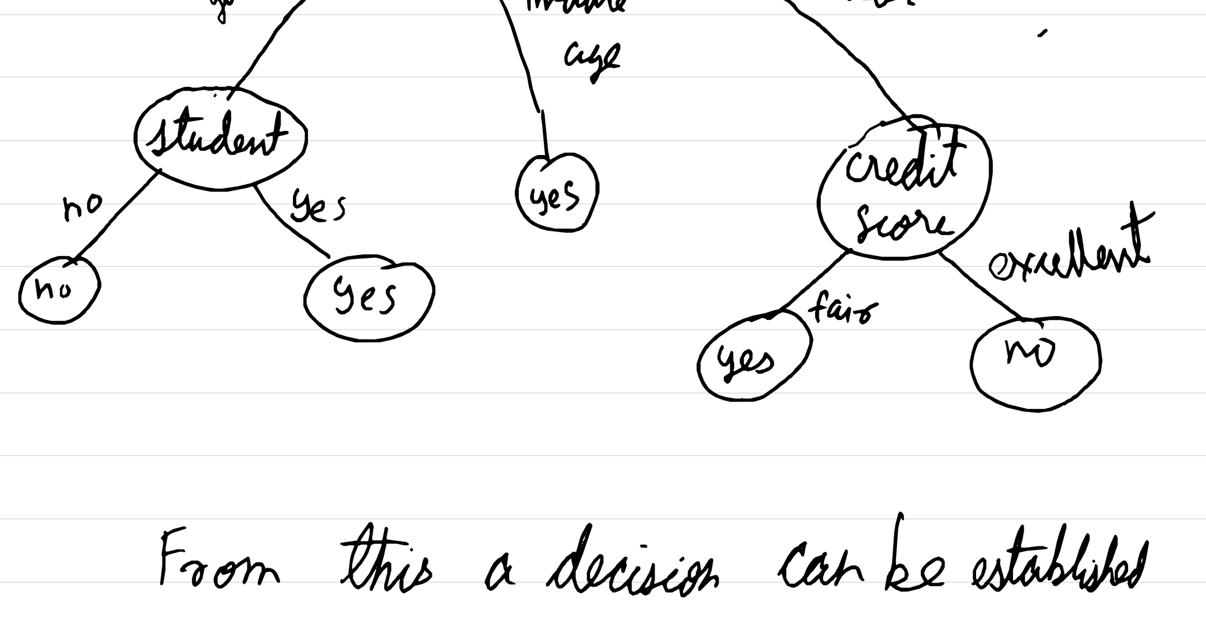
3 — y: 0, h: 3

3 — y: 2, n: 0

(for middle age, all values are yes ∴ terminal node found)

on solving further for senior, we can see clear corelation with it & the ans so it becomes leaf node

Final Answer →



From this a decision can be established

eg if age is youth, it is a student then car can be buyed

Q2

Day	Outlook	Temp	Humidity	Wind	play Tennis?
1	Sunny	Hot	High	Weak	No
2	S	H	H	Strong	N
3	Overcast	H	H	W	Yes
4	Rain	Mild	H	W	y
5	R	Cool	Normal	W	y
6	R	C	N	S	N
7	O	C	N	S	y
8	S	M	H	W	N
9	S	C	N	W	y
10	R	M	N	W	y
11	S	M	N	S	y
12	O	M	H	S	y
13	O	H	N	W	y
14	R	M	H	S	N

Step 1 → Entropy for parent → Tennis

yes : 9
No : 5
$\overline{\overline{14}}$

$$H = -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14}$$

$$= 0.940$$

Step 2 → Entropy for attribute Outlook

outlook

Sunny — y : 2, n : 3, $\overline{\overline{5}}$

overcast — y : 4, n : 0, $\overline{\overline{4}}$

Rain — y : 3, n : 2, $\overline{\overline{5}}$

$$H = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.970$$

$$H = -\frac{4}{4}\log\frac{4}{4} - 0 = 0$$

$$H = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.970$$

$$H = \frac{5}{14} \times (0.970) + \frac{4}{14} \times (0) + \frac{5}{14} \times 0.970$$

$$= 0.692$$

Info gain = 0.940 − 0.692 = 0.248

Attribute : Temp.

Temp

hot — y : 2, N : 2, $\overline{\overline{4}}$

Mild — y : 4, N : 2, $\overline{\overline{6}}$

Cool — y : 3, N : 1, $\overline{\overline{4}}$

$$H = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$$

$$H = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6} = 0.918$$

$$H = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.811$$

$$H = \frac{4}{14} \times (1) + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.811$$

$$= 0.910$$

Info gain = 0.940 − 0.910 = 0.03

Attribute Humidity

Humidity

High — y : 3, n : 4, $\overline{\overline{7}}$

Normal — y : 6, n : 1, $\overline{\overline{7}}$

$$H = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} = 0.985$$

$$H = -\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7} = 0.59$$

$$H = \frac{7}{14}(0.985) + \frac{7}{14}(0.59)$$

$$= 0.7875$$

Info gain = 0.940 − 0.7875 = 0.152

Attribute : Wind

Wind

Weak — y : 6, n : 2, $\overline{\overline{8}}$

strong — y : 3, n : 3, $\overline{\overline{6}}$

$$H = -\frac{2}{8}\log\frac{2}{8} + -\frac{6}{8}\log\frac{6}{8} = 0.811$$

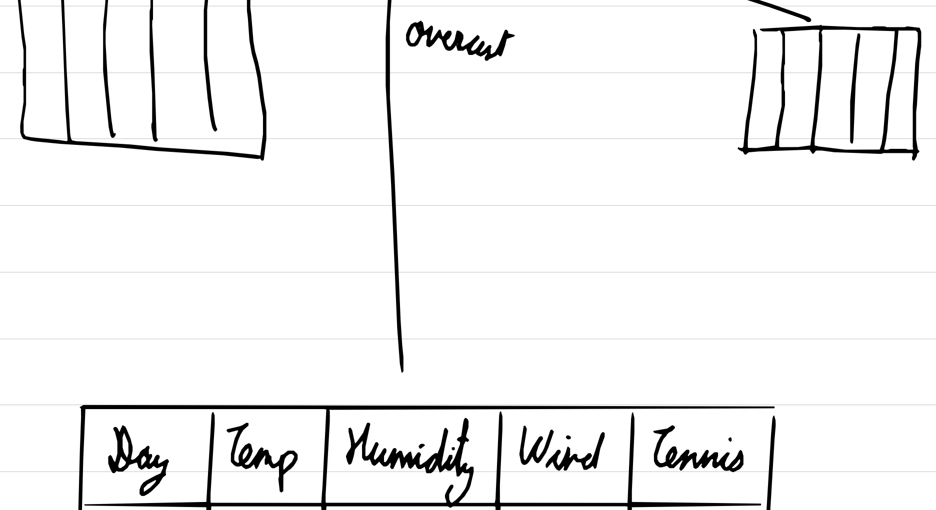
$$H = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$$

$$H = \frac{8}{14} \times (0.811) + \frac{6}{14} \times 1$$

$$= 0.892$$

Info gain = 0.940 − 0.892 = 0.048

Best attribute for the split is with the Most Info gain that is outlook



Day	Temp	Humidity	Wind	Tennis
D_3	H	H	W	y
⋮	⋮	⋮	⋮	⋮

Q3

CGPA	Communication	Skills	Employable
High	Good	Good	Yes
Medium	G	G	y
H	G	Bad	y
H	G	G	y
H	Bad	G	y
M	Good	G	y
L	Bad	B	No
L	Bad	B	No
M	Good	B	y
M	Bad	G	No
M	Good	B	y

→ Parent

Yes: 8
No: 3

$H = -\frac{8}{11}\log\frac{8}{11} - \frac{3}{11}\log\frac{3}{11} = 0.845$

Communication

Good
Yes: 7
No: 0

Bad
Yes: 1
No: 3

$H = -\frac{7}{7}\log\left(\frac{7}{7}\right) - 0$

$= 0$

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

$= +0.8112$

$H = \frac{7}{11} \times 0 + \frac{4}{11} \times 0.8112$

$= 0.295$

Info gain $= 0.55$

C.G.P.A

Yes: 8
No: 3

High

y: 4
N: 0
Total 4

$H = \Sigma P_i \log_2 P_i$

$= \frac{4}{4}\log 1 + 0\log 0$

$= 1$

Medium

y: 4
n: 1
tot: 5

$H = \Sigma P_i \log P_i$

$= \frac{1}{5}\log\frac{1}{5} + \frac{4}{5}\log\frac{4}{5}$

$= -0.217$

low

y: 0
n: 2
tot: 2

$H = \Sigma P_i \log P_i$

$= \frac{2}{2}\log 2 + \frac{0}{0}\log 0$

$= 1$

$P = \frac{4}{11} \times 1 + \frac{5}{11} \times -0.217 + \frac{2}{11} \times 1 = 0.4468$

Info gain $= 0.845 - 0.4468 = 0.398$

CART

Classification & Regression Trees.

CART is a variation of the decision tree algorithm.

Cart builds a tree like structure consisting of nodes and branches. Nodes represent decision points while branches represent the possible outcomes of the decisions

Best split at every point is calculated using Gini impurity.

Gini requires binary split. Hence if the split is more than 2 then parts must be made.

CART

(Classification And Regression tree)

* Based on Gini Index

* $\text{Gini Index } (D) = 1 - \sum_{i=1}^{m} P_i^2$

Partition

P_i is the probability that a tuple belongs to class C_i

* Gini Needs Binary split

eg Income

Low   High   Medium

} wont work with Gini

Data must be divided into only two catagories
In order to do that, consider all possible subsets of the sets

({high, low}, medium) → P_1
({high, medium}, low) → P_2
({low, medium}, high) → P_3

3 combinations $^3C_2$

Example → calculate Gini value of the below split & info gain

up: 4
down: 6

Past Trend

Positive

up: 4
down: 2
6

Negative

up: 0
down: 4
4

→ Gini Parent

$G = 1 - P_+^2 - P_-^2$

$= 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = 0.48$

Gini for +ve

$G = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2$

$= 0.44$

Gini for -ve

$G = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2$

$= 0$

Gini for split $= \frac{6}{10} \times (0.44) + 0 \times \frac{4}{10}$

$= 0.264$

Info gain $= 0.48 - 0.264 = 0.216$

Q3) Make decision tree for gini Indexing on basis of Income

id	age	income	student	crediting	buy Car
1	Youth	high	no	fair	no
2	y	h	no	excellent	no
3	m-a	h	no	f	y
4	S	medium	no	f	y
5	S	Low	yes	f	y
6	S	L	yes	e	n
7	m-a	L	yes	e	y
8	y	m	no	f	n
9	y	L	y	f	y
10	S	m	y	f	y
11	y	m	y	e	y
12	ma	m	n	e	y
13	ma	h	y	f	y
14	S	m	n	e	n

→

Parent yes = 9
no = 5

$$Gini(Parent) = 1 - \sum P_i^2 = 1 - (P_{yes}^2 + P_{No}^2)$$

$$= 1 - \left[\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right]$$

$$= 0.459$$

Possible splits → i) {L, m}, h
ii) {L, h}, m
iii) {m, h}, L

Consider the Binary split as {L, m}, h

$D_1 = \{L, m\}$
$D_2 = h$ } These are the partitions

$$Gini_A(D) = \left(\frac{D_1}{D}\right) Gini(D_1) + \left(\frac{D_2}{D}\right) Gini(D_2)$$

Income

Low, Medium — high

yes = 7
no = 3
10

yes = 2
No = 2
4

$$Gini(D) = \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{2}{4}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$

$$= 0.443$$

Consider split as {L, h}, m

Income

Low, high — Medium

yes = 5
no = 3
8

yes = 4
No = 2
6

$$Gini = 1 - \left(\frac{5}{8}\right)^2 - \left(\frac{3}{8}\right)^2$$

$$= 0.468$$

$$Gini = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2$$

$$= 0.44$$

$$Gini(D) = \frac{8}{14} \times (0.468) + \frac{6}{14} \times (0.44)$$

$$= 0.457$$

Consider split as {h, m}, L

Income

high Medium — Low

yes = 6
no = 4
10

yes = 3
No = 1
4

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

$$= 0.48$$

$$Gini = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$$

$$= 0.375$$

$$Gini = \frac{10}{14}(0.48) + \left(\frac{4}{14}\right)(0.375)$$

$$= 0.45$$

Among all the three, the smallest giniindex is of ({L m}, h) hence consider that partition.

Q2

Past trend	Open interest	Trading volume	Return
+ve	Low	High	Up
-ve	High	Low	Down
+ve	L	H	U
+ve	H	H	U
-ve	L	H	D
+ve	L	L	D
-ve	H	H	D
-ve	L	H	D
+ve	L	L	D
+ve	H	H	U

Parent
Total up: 4
down: 6

$$\text{Gini Parent} = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2$$

$= 0.48$

For Past Trend

Past Trend

Positive

up: 4
down: 2
6

$$\text{Gini} = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2$$

0.44

Negative

up: 0
down: 4
4

$$\text{Gini} = 1 - 0 - \left(\frac{4}{4}\right)^2$$

$= 0$

$$\text{Gini} = \frac{6}{10} \times 0.44 = 0.266$$

Gini Open Interest

Open Interest

Low

up: 2
down: 4
6

Gini

$$= 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2$$

$= 0.44$

High

up: 2
down: 2
4

Gini

$$1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$$

$= 0.5$

$$\text{Gini} = \frac{6}{10} \times 0.44 + \frac{4}{10} \times 0.5$$

$= 0.464$

Gini for Trading volume

High

up: 4
down: 3
7

Gini

$$1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2$$

$= 0.489$

Low

up: 0
down: 3
3

$$1 - 0 - \left(\frac{3}{3}\right)^2$$

$= 0$

Gini =

$$\frac{7}{10} \times 0.489 + \frac{3}{10} \times 0$$

$= 0.3423$

Smallest Gini value is of Past Trend

Advantages of decision trees

1) Understandable rules (Interpretable)
2) Easy calculation
3) Handel both continouse & catagorical variables

Disadvantages of decision trees

1) No global optimization
2) Error prone
3) Overfitting (especially deeptrees)

Methods to avoid overfitting

① Prepruning - stop growing when data split is not statistically significant

② Postpruning - Grow full tree & remove some nodes.

Prepruning

Early stopping

If some condition is met, the current node will not be split, even if it is not 100% pure

It will be a leaf node with the label of the majority class in current set

eg

Age
>20

Yes : 9 } yes
No : 1

$\rightarrow$

Age
>20
yes

Common stopping creiterion

1. Entropy / Gini impurity
2. No of samples
3. Depth of tree

Post Pruning

Prone nodes in bottom up manner if it decreases validation error

validation → unseen dataset

XOR problem & linearity

XOR is a non linearly seperable problem

x_1	x_2	output
0	0	0
1	0	1
0	1	1
1	1	0

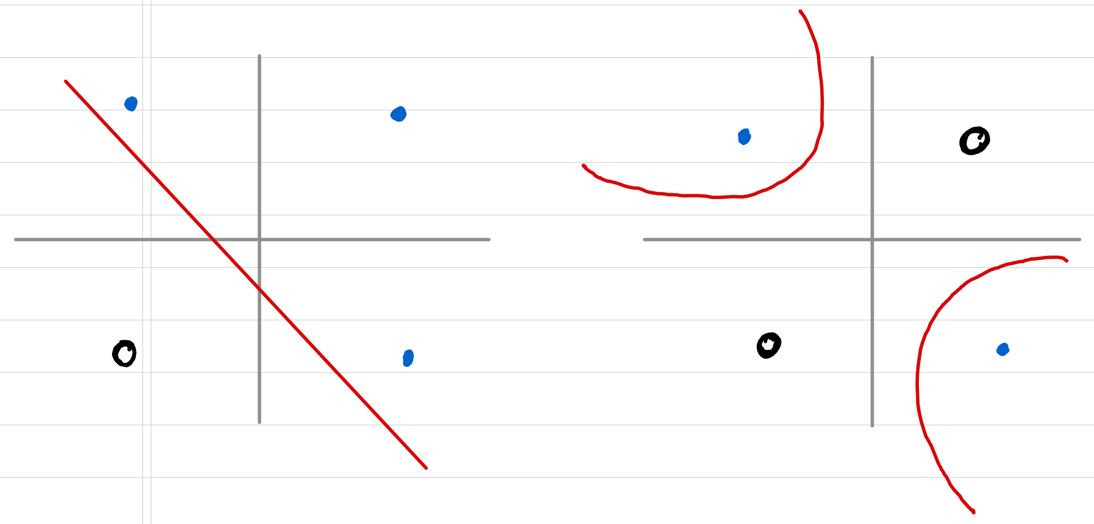
No straight line can seperate 1 & 0

When there is such relationship in the data, it is a non linear data.

1) logistic regression is a linear classifier

meaning it tries to find a single decision boundary in higher dimentions to seperate the classes.

Since XOR requires at least two decision boundaries [for (0,0) and (1,1) from (0,1) and (1,0)], logistic regression cannot model it



• → output 1
o → output 0
— → classification line

2) Naive Bayes → Naive Bayes assumes conditional independance between the features and class labels

However in XOR, the two features are not independant, their interaction determines the output

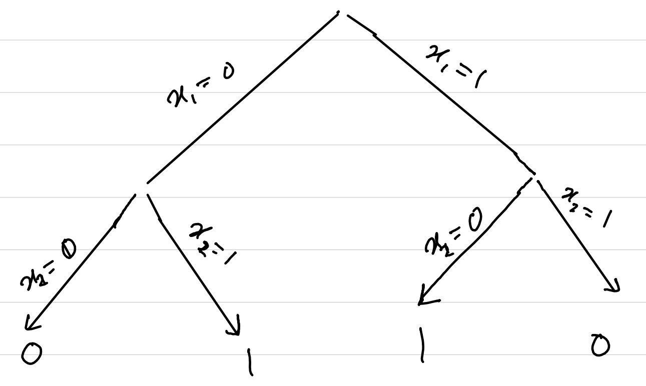
x_1	x_2	y
0	0	0
1	0	1
0	1	1
0	0	0

$P(y=0 \mid x_1=0) = 0.25$
$P(y=0 \mid x_1=1) = 0.25$
$P(y=1 \mid x_1=0) = 0.25$
$P(y=1 \mid x_1=1) = 0.25$
$P(y=0 \mid x_2=0) = 0.25$
$P(y=0 \mid x_2=1) = 0.25$
$P(y=1 \mid x_2=0) = 0.25$
$P(y=1 \mid x_2=1) = 0.25$

No decision can be made in this manner

3) Linear SVM → fails for the same reason as logistic regression

4) Decision trees → split feature space into regions non linearly



Can represent the XOR and such non linear data well.

5) Random forest → since it is ensemble of D.T., it can also represent the data well

6) Kernalized SVM → Works well as the non linear boundaries are supported

7) Neural networks → can do non linear boundaries