P(No | New Instance) = P(No) * P(color = Red | No) * P(TYPE = SUV | No) * P(origin = Domestic | No)

$= 0.5 \times \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} = 0.072$

P(Yes | New Instance) < P(No | New Instances)

The new car is not getting stolen.

Q.2 The following table shows the dataset on hand. Our target is to predict whether to play golf or not. Draw a decision tree using the ID3 algorithm to find a feature best suitable as its root.

Outlook	Temp	Humidity	Wind	Play Tennis ?
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rain	Mild	High	False	Yes
Rain	Cool	Normal	False	Yes
Rain	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	false	No
Sunny	Cool	Normal	False	Yes
Rain	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rain	Mild	High	True	No

DOMS | Page No.
Date / /

ID3 Algorithm :-

The attribute which gives the maximum information. it will be considered as root node.

Attribute 1 :- Outlook

values(outlook) = Sunny, overcast, Rain

$S = [9+, 5-]$ (9+ ↑ Yes, 5- ↑ No)

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

$= 0.94$

this is entropy for entire dataset, we have calculate entropy for sunny, overcast Rain.

$S_{sunny} = [2+, 3-]$ $Entropy[S_{sunny}] = \frac{-2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.971$

$S_{overcast} = [4+, 0-]$ $Entropy[S_{overcast}] = \frac{-4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}$

$= 0$

$S_{Rain} = [3+, 2-]$ $Entropy[S_{Rain}] = \frac{-3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.971$

$Information\ Gain(S, Outlook) = Entropy(S) - \sum_{v \in (sunny, overcast, Rain)} \frac{|S_v|}{S} Entropy(S_v)$

$= 0.94 - 5/14 \times 0.971 - 4/14 \times 0 - 5/14 * 0.971$

$= 0.2466$

DOMS | Page No.
Date / /

- Attribute 2 :- Temperature
Values (Temperature) = Hot, Mild, Cool

S_{Hot} = [2+, 2-] Entropy $[S_{Hot}] = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$

$= 1$

S_{mild} = [4+, 2-] Entropy $[S_{mild}] = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}$

$= 0.9182$

S_{cool} = [3+, 1-] Entropy $[S_{cool}] = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$

$= 0.8112$

Information Gain (S, Temp) = Entropy(S) $- \sum_{v \in (Hot, mild, cool)} \frac{|S_v|}{S}$ Entropy(S_v)

$= 0.94 - \frac{4}{14} \times 1 - \frac{6}{14} \times 0.9182 - \frac{4}{14} \times 0.8112$

$= 0.029$

- Attribute 3 :- Humidity
Values (Humidity) :- High, Normal

S_{High} = [3+, 4-] Entropy $(S_{High}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$

$= 0.9852$

S_{Normal} = [6+, 1-] Entropy $(S_{Normal}) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$

DOMS | Page No.
Date / /

$= 0.5916$

Information Gain (S, Humidity)

$= \text{Entropy}(S) - \sum_{v \in (\text{High}, \text{Normal})} \frac{|S_v|}{S} \text{Entropy}(S_v)$

$= 0.94 - \frac{7}{14} \times 0.9852 - \frac{7}{14} \times 0.5916$

$= 0.1516$

- Attribute 4 = Wind

Values(Wind) = Weak, Strong

$S_{\text{First}} = [6+, 2-]$ $\text{Entropy}(S_{\text{False}}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$

$= 0.8112$

$S_{\text{True}} = [3+, 3-]$ $\text{Entropy}(S_{\text{True}}) = \frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$

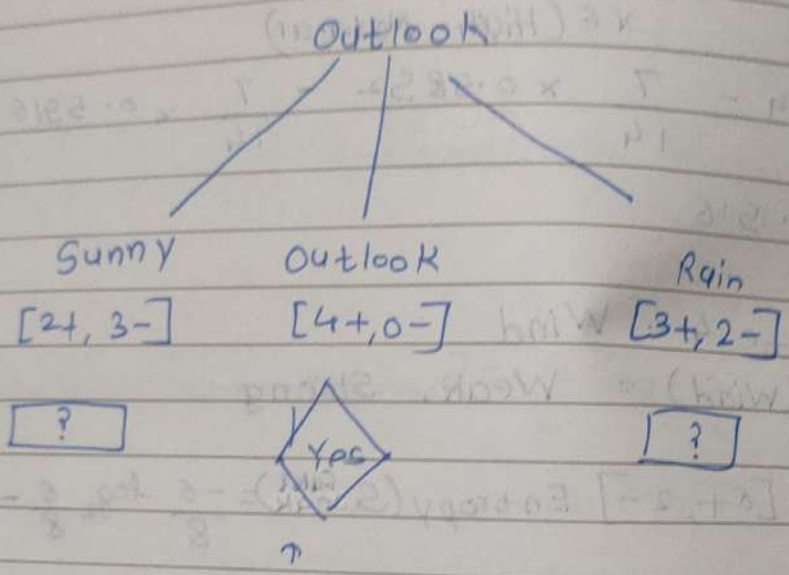
$= 1$

Information Gain (S, Wind) $= \text{Entropy}(S) - \sum_{v \in (\text{False}, \text{True})} \frac{|S_v|}{S} \text{Entropy}(S_v)$

$= 0.94 - \frac{8}{14} \times 0.8112 - \frac{6}{14} \times 1$

$= 0.04788$

IG (S, outlook) = 0.2466 ⇒ maximum
IG (S, Temperature) = 0.029
IG (S, Humidity) = 0.1516
IG (S, wind) = 0.04788



It is yes because there are no values

Only consider Sunny's rows for further calc.

Day	Temp.	Humidity	Wind	Play tennis
D_1	Hot	High	False	No
D_2	Hot	High	True	No
D_8	Mild	High	False	No
D_9	Cool	Normal	False	Yes
D_{11}	Mild	Normal	True	Yes

DOMS | Page No.
Date / /

Attribute : Temperature

$S_{sunny} = [2+, 3-]$ Entropy $[S_{sunny}] = \frac{-2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.97$

$S_{Hot} = [0+, 2-]$ Entropy $[S_{Hot}] = \frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}$

$= 0$

$S_{mild} = [1+, 1-]$ Entropy $[S_{mild}] = \frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

$= 1$

$S_{cool} = [1+, 0-]$ Entropy $[S_{cool}] = \frac{-1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}$

$= 0$

Information Gain $(S_{sunny}, Temp) = Entropy(S) - \sum\frac{(S_v)}{S} Ent(S_v)$

$\forall v \in (Hot, mild, cool)$

$= 0.97 - \frac{2}{5}\times 0 - \frac{2}{5}\times 1 - \frac{1}{5}\times 0$

$= 0.97 - 0.4 = 0.57$

Attribute :- Humidity

$S_{High} = [0+, 3-]$ Entropy $(S_{High}) = -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3}$

$= 0$

$S_{Normal} = [2+, 0-]$ Entropy $(S_{Normal}) = -\frac{2}{2}\log_2\frac{2}{2} - 0$

$= 0$

Information Gain (S_sunny, Humidity) = Entropy(S) − Σ (|S_v| / S) Entropy(S_v), v ∈ (High, normal)

= 0.97 − 0 − 0

= 0.97

Attribute :- Wind

S_{False} = [1+, 2−] Entropy(S_{False}) = $-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

= 0.9182

S_{True} = [1+, 1−] Entropy(S_{True}) = $-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

= 1

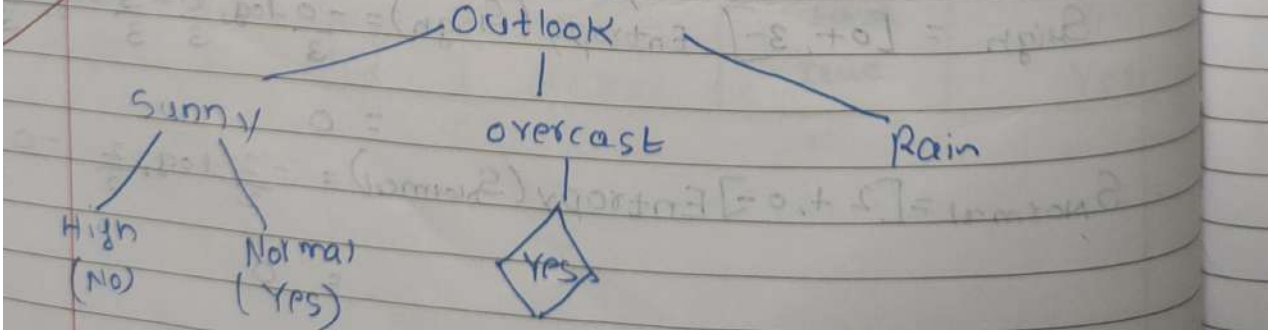
Gain(S_{sunny}, wind) = $0.97 - \frac{3}{5} \times 0.9182 - \frac{2}{5} \times 1$

= 0.0192

IG(S_{sunny}, Temp) = 0.57

IG(S_{sunny}, Humidity) = 0.97 ⇒ maximum

IG(S_{sunny}, Wind) = 0.0192



Now only consider Rain's rows for further calc.

Temperature	Humidity	Wind	Play Tennis
Mild	High	Weak	Yes
Cool	Normal	Weak	Yes
Cool	Normal	Strong	No
Mild	Normal	Weak	Yes
Mild	High	Strong	No

Attribute: Temperature

$S_{mild} = [2+, 1-]$ $Entropy(S_{mild}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$

$= 0.9183$

$S_{cool} = [1+, 1-]$ $Entropy(S_{cool}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

$= 1.0$

$IG_{(mild)} = Entropy(S_{Rain}) - \sum_{v \in (mild, cool)} \frac{|S_v|}{5} Entropy(S_v)$

$= 0.971 - \frac{3}{5} \times 0.918 - \frac{2}{5} \times 1$

$= 0.9912$

Attribute: Humidity

$S_{High} = [1+, 1-]$ Entropy$(S_{High}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

$= 1$

$S_{Normal} = [2+, 1-]$ Entropy$(S_{Normal}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$

$= 0.9182$

$IG(S_{Rain}, Humidity) = 0.97 - \frac{2}{5} \times 1 - \frac{3}{5} \times 0.9182$

$= 0.0912$

Attribute: Wind

$S_{False} = [3+, 0-]$ Entropy$(S_{False}) = -\frac{3}{3}\log_2\frac{3}{3} - 0$

$= 0$

$S_{True} = [0+, 2-]$ Entropy$(S_{True}) = \frac{0}{0} - \frac{2}{2}\log_2\frac{2}{2}$

$= 0$

~~$IG(S_{Rain}, Wind)$ Entropy$(S_{strong}) = \frac{0}{0} - \frac{2}{2}$~~

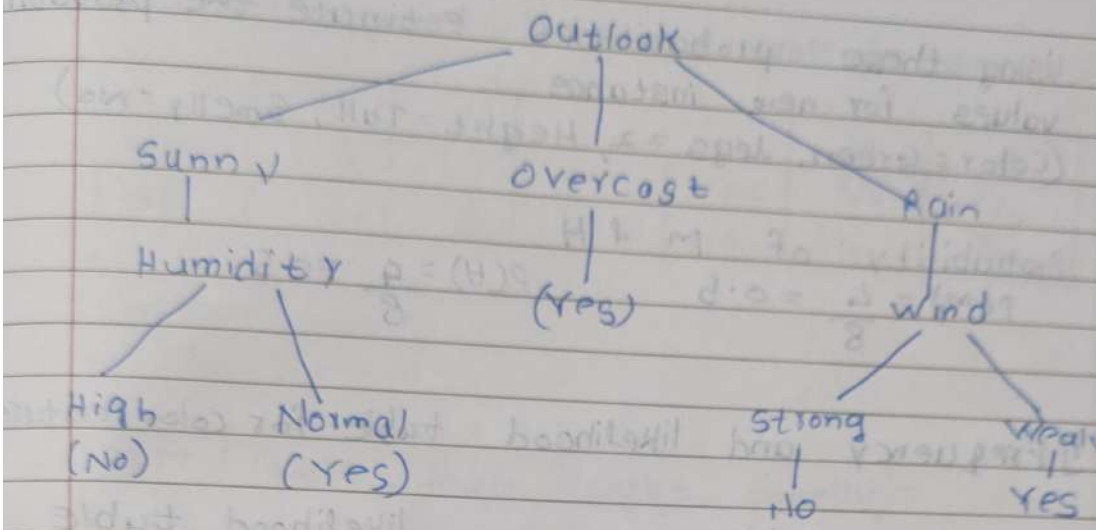
$IG(S_{Rain}, Wind) = 0.97 - \frac{3}{5} \times 0 - \frac{2}{5} \times 0$

$= 0.97$

$IG(Temp) = 0.0912$

$IG(Humidity) = 0.0912$

$IG(Wind) = 0.97 \Rightarrow$ maximum



2.3. Estimate conditional probabilities (Naive Bayes Classifier) of each attribute {color, legs, height, smelly} for species classes: {M, H} using data given in table.

Color	Legs	Height	Smelly	Species
White	3	Short	Yes	M
Green	2	Tall	No	M
Green	3	Short	Yes	M
White	3	Short	Yes	M
Green	2	Short	No	H
White	2	Tall	No	H
White	2	Tall	No	H
White	2	Short	Yes	H

DOMS | Page No.
Date / /

Using these probabilities estimate the probabilities values for new instance
(Color = Green, legs = 2, Height = Tall, Smelly = No)

→ Probability of M & H

$P(M) = \frac{4}{8} = 0.5$ $P(H) = \frac{4}{8} = 0.5$

1] Frequency and likelihood table for color feature

Frequency table

		species	
		M	H
Color	White	2	3
	Green	2	2

likelihood table

	P(M)	P(H)
White	2/4	3/4
Green	2/4	1/4

2] Frequency & likelihood table for legs

	species	
	M	H
3	3	0
2	1	4

	species	
	M	H
3	3/4	0
2	1/2	1

3] Frequency & likelihood table for height

	species	
	M	H
Short	3	2
Tall	1	2

	species	
	M	H
Short	3/4	1/2
Tall	1/4	1/2

DOMS | Page No.
Date / /

4] Frequency & likelihood table for smelly

	M	H
Yes	3	1
No	1	3

	M	H
Yes	3/4	1/4
No	1/4	3/4

Probability:-
$P(M/x) =$

$P(M / \text{Colors}, \text{legs}, \text{Height}, \text{Smelly}) =$

$= P(\text{color}/M) * P(\text{legs}/M) * P(\text{Height}/M) * P(\text{Smelly}/M)$

$= \frac{2}{4} * \frac{1}{4} * \frac{1}{4} * \frac{1}{4} * \frac{4}{8}$

$= \frac{1}{256} = 0.0039$

Probability —
$P(H/x) = P(H / \text{Color}, \text{legs}, \text{Height}, \text{Smelly})$

$= \frac{1}{4} * 1 * \frac{1}{2} * \frac{3}{4} * \frac{4}{8} = \frac{3}{64}$

$= 0.046$

Q.2 Draw a decision tree for a given dataset using the Gini Index Algorithm. You can develop tree till level 1.

Age	Income	Age	Student	Credit rating	Buys Computer
<=30	High		No	Fair	No
<=30	High		No	Excellent	No
31-40	High		No	Fair	Yes
>40	Medium		No	Fair	Yes
>40	Low		Yes	Fair	Yes
>40	Low		Yes	Excellent	No
31-40	Low		Yes	Excellent	Yes
<=30	Medium		No	Fair	No
<=30	Low		Yes	Fair	Yes
>40	Medium		Yes	Fair	Yes
<=30	Medium		Yes	Excellent	Yes
31-40	Medium		No	Excellent	Yes
31-40	High		Yes	Fair	Yes
>40	Medium		No	Excellent	No

→ There are two possible output values - Yes & No. The data has 9 instances of Yes & 5 of No.

$$Gini(S) = 1 - \left[\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right] = 1 - \left[\frac{81+25}{196}\right] = \frac{196}{196}$$

$$= 0.459$$

DOMS | Page No.
Date / /

Gini Index for Age attribute
It has 3 possible attributes = <=30, 31-40, >40

For age <=30
$Gini(S) = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right] = 1 - \left[\frac{4+9}{25}\right] = \frac{12}{25} = 0.48$

For age = 31-40
$Gini(S) = 1 - \left[\left(\frac{4}{4}\right)^2 + (0)^2\right] = 0$

Age = >40
$Gini(S) = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = \frac{12}{25} = 0.48$

weighted avg(Age) $= \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48$

$= 0.3428$

Gini Index for IncomeAge
It has 3 possible attr. = High, Low, Medium

For IncomeAge = ~~Low~~ High
$Gini(S) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 1 - \left[\frac{8}{16}\right] = 0.5$

For IncomeAge = Low
$Gini(S) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 1 - \left[\frac{10}{16}\right] = 0.375$

For IncomeAge = Medium
$Gini(S) = 1 - \left[\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right] = 1 - \left[\frac{20}{36}\right] = 0.444$

weighted avg (IncomeAge)

$= \frac{4}{14} \times 0.5 + \frac{4}{14} \times 0.375 + \frac{6}{14} \times 0.444$

$= 0.4401$

Gini Index for student Attribute

for student = No

$Gini(S) = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = \left[\frac{49 - 25}{49}\right] = 0.4897$

For student = Yes

$Gini(S) = 1 - \left[\left(\frac{6}{7}\right)^2 + \left(\frac{1}{7}\right)^2\right] = 1 - \left[\frac{37}{49}\right] = 0.2448$

weighted age (student)

$= \frac{7}{14} \times 0.4897 + \frac{7}{14} \times 0.2448 = 0.3672$

Gini Index for credit-ranking

for credit-ranking = Fair

$Gini(S) = 1 - \left[\left(\frac{6}{8}\right)^2 + \left(\frac{2}{8}\right)^2\right] = 1 - \frac{40}{64} = 0.375$

for credit-ranking = Excellent

$Gini(S) = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] = 1 - \frac{18}{36} = 0.5$

weighted avg (credit-ranking)

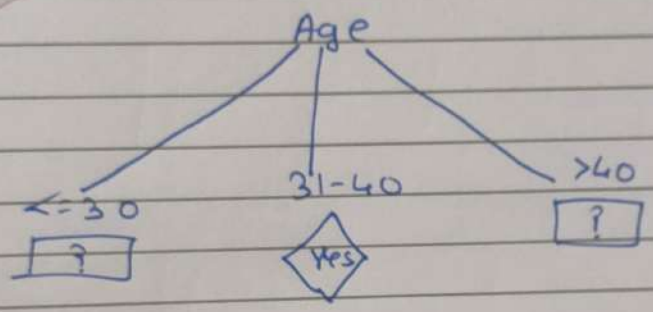
$= \frac{8}{14} \times 0.375 + \frac{6}{14} \times 0.5$

$= 0.2142 + 0.2142 = 0.4284$

DOMS | Page No.
Date / /

Gini(Age) = 0.3428 ⇒ Minimum
Gini(IncomeAge) = 0.4401
Gini(Student) = 0.3672
Gini(credit-ranking) = 0.4284

Age attr. has minimum Gini Index



8/10/11/23