DMIA.

Problems:

- Decision Tree
- KNN [K nearest Neighbor]
- Linear Regression
- Bayes

PART-A

5 out of 4

10m

Part B

1 -> 10m comp.

Theory:

- Multiple logistic regression
- logistic regression
- Generalized linear model (GLM)
- Simple problems with algo
- Derivation sigmoid fnn.

KNN

① Decision Tree:

$$Info(D) = -\sum_{i=1}^{m} pi \log_2(pi) \quad \text{[Exact info]}$$

$$Gain(A) = Info(A) - Info(D) \quad \text{[Gain (A)]}$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times info(D_j) \quad \text{[New info].}$$

Outlook	Temp	Humidity	Windy	Play tennis
R	H	H	F	NO
R	H	H	T	No
R	H	H	F	yes.
O	H	H	F	yes
vS	M	H	F	yes.
vS	C	N	T	No
vS	C	N	T	yes
O	C	N	F	No
R	M	H	F	yes
R	C	N	F	yes
vS	M	N	F	yes
R	M	N	T	yes
O	M	H	T	yes
O	H	N	F	yes
vS	M	H	T	NO.

total = 14

yes = 9

NO = 5

Calculate $Info(D) = -\sum_{i=1}^{m} Pi \log_2(Pi)$

$Gain(A) = Info(D) - Info_{(A)}(D)$.

14 for yes; NO;

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

$$\boxed{\text{Info}(D) = 0.940}$$ Total $= 0.637$.

1st Attribute outlook →

	Y	N
S	2	3 ✓
O	3	2
R	4	0

$$\text{Info}_{\text{outlook}}(D) = \frac{5}{14}\left(\frac{-2}{5}\log_2\left[\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)\right) +$$

$$\frac{5}{14}\left(\frac{-3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) +$$

$$\frac{4}{14}\left(\frac{-4}{4}\log_2\frac{4}{4} - \frac{0}{0}\right)$$

$$\text{Info}_{\text{outlook}}(D) = 0.693$$

2nd Attribute Temperature

	Y	N
H	2	2
M	4	2
C	3	1

$$\text{Info}_{\text{outlook}}(D) = \frac{4}{14}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) +$$

$$\frac{6}{14}\left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) +$$

$$\frac{4}{14}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)$$

$$\text{Info}_{\text{outlook}}(D) = 0.911$$

Attribute humidity;

	Y	N
H	3	4
N	6	1

$$\text{Info}_{\text{outlook}}(D) = \frac{7}{14}\left(-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right) + \frac{7}{14}\left(\frac{-6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right)$$

$$\text{Info}_{\text{outlook}}(D) = 0.788$$

Attribute windy;

	Y	N
W	6	2
S	3	3

$$\text{Info}_{\text{outlook}}(D) = \frac{8}{14}\left(-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right) + \frac{6}{14}\left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right)$$

$$\text{Info}_{\text{outlook}}(D) = 0.892.$$

$$\boxed{\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)}$$

$$\text{Gain(outlook)} = 0.940 - 0.693$$

$$= 0.247$$

$$\text{Gain(Temp)} = 0.940 - 0.911$$

$$= 0.029$$

$$\text{Gain(Hum)} = 0.940 - 0.788$$

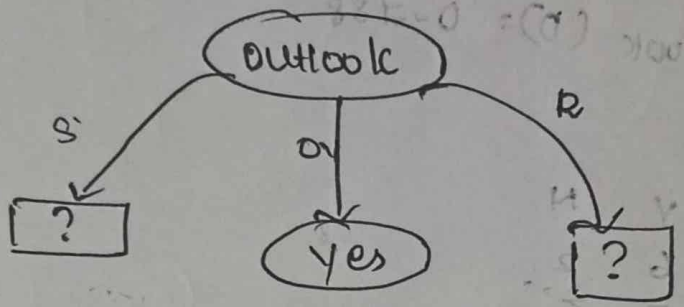
$$= 0.152$$

$$\text{Gain(windy)} = 0.940 - 0.892 = 0.048$$

Attribute	Gain
Outlook	0.247 ✓
Temp	0.029
Hum	0.152
windy	0.048

Highest

Root Node: Outlook



Consider the Outlook table.

Outlook	Temp	Humi	Windy	Tennis
	H	High	W	N
Sunny	H	High	S	N
	M	High	W	N
	C	Nor	W	Y
	M	Nor.	S	Y

Outlook	Temp	Humi	Windy	Tennis
	M	H	W	Y
	C	N	W	Y
	C	N	S	N
Rainy	M	N	W	Y
	M	H	S	N

sunny, Yes = 2 , NO = 3

$$Info(sunny) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}$$

$$= 0.971$$ Total

Humidity,

	Y	N
H	0	3
N	2	0

$$\Rightarrow \frac{3}{5}\left(-\frac{0}{0} - \frac{3}{3}\log\frac{3}{3}\right) + \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2}\right)$$

$$= 0$$

$$Info_{Star\ Hum}(sunny) = 0$$

Tem,

	Y	N
H	0	2
M	1	1
C	1	0

Windy,

	Y	N
W	1	2
S	1	1

$$Info_{win}(sunny) = \frac{3}{5}\left(-\frac{1}{3}\log\frac{1}{3}\right) + \frac{2}{5}(\ldots$$

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

$$Info_{win}(sun) = 0.951$$

~~$Info_{tem}(sun)$~~ =

Temp,

	Y	N
H	0	2
C	1	0
M	1	1

$$\Rightarrow \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2}\right) + \frac{1}{5}\left(\frac{-1}{1}\log_2\frac{1}{1}\right) + \frac{2}{5}\left(\frac{-1}{2}\log_2\frac{1}{2}\right)$$

$$\frac{2}{5}\left(\frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

Info temp (sun) $\Rightarrow 0.4$

Gain (Temp) $= 0.971 - 0.4$

$= 0.571$

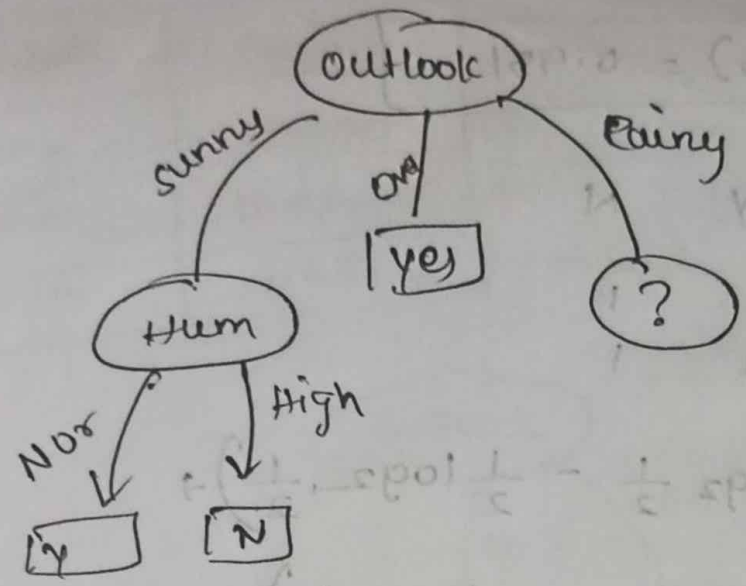
Gain (Hum) $= 0.971 - 0$

$= 0.971$

Gain (Win) $= 0.971 - 0.951$

$= 0.02.$

Attribute	Gain	
Tem	0.571	
Hum	0.971 ✓	Highest
Win	0.02	



Rainy;

Consider the Outlook table;

Outlook	Temp	Humi	windy	Tennis
	M	H	W	Y
	C	N	W	Y
Rainy	C	N	S	N
	M	N	W	Y
	M	H	S	N.

Rainy; Yes = 3, NO = 2.

$$\text{Info(Rainy)} = -\frac{3}{5}\left[\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right]$$

	Y	N
M	1	1
C	1	1

$\Rightarrow 0.971$ Total

$$\text{Info}_{\text{Temp}}(\text{Rain}) = \frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) +$$

$$\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

$\boxed{Info_{Temp}(Rain) = 0.951}$

Hum,

	Y	N
H	1	1
N	2	1

$$\Rightarrow \frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)$$

$\boxed{Hum \Rightarrow 0.951}$

Windy,

	Y	N
W	3	0
S	0	2

$$Info_{win}(Rainy) = \frac{3}{5}\left(\frac{-3}{3}\log_2\frac{3}{3} - 0\right) + \frac{2}{5}\left(-0 - \frac{2}{2}\log_2\frac{2}{2}\right)$$

$\boxed{Info_{win}(Rain) = 0}$

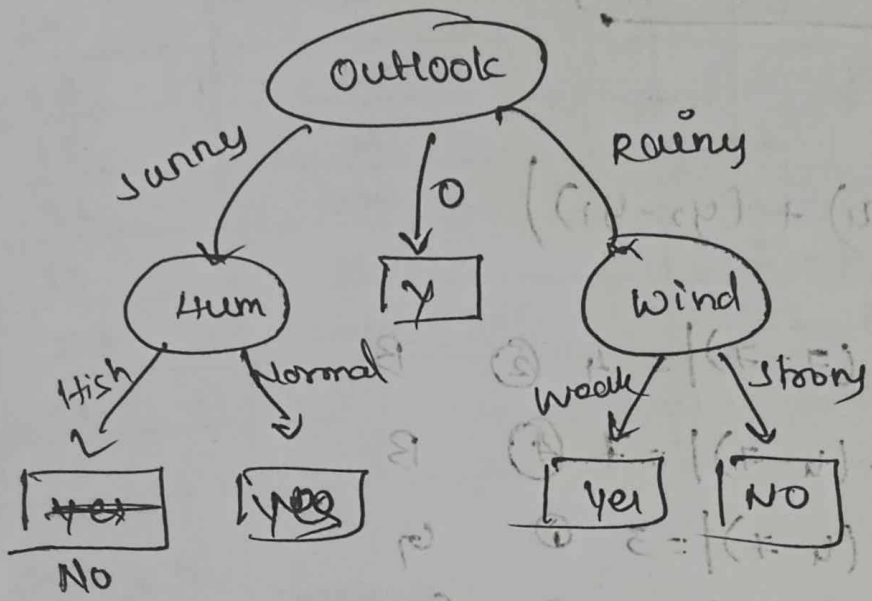
$$Gain(Temp) = 0.971 - 0.951$$
$$= 0.020$$

$$Gain(Hum) = 0.971 - 0.951$$
$$= 0.020$$

$$Gain(wind) = 0.971 - 0$$
$$= 0.971$$

Attribute	Gain
Tem	0.020
Hum	0.020
Win	0.971 ✓

Highest



(2) K Nearest Neighbors :

↳ Powerful classification algo. used in pattern recognition.

Distance formulas;

Euclidean → $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

Manhattan → $|(x_2-x_1)+|y_2-y_1|$ (or) $|x_1-y_1|$

Adv :

↳ Simple
↳ Applied from any distribution
↳ Good classification

Dis :

↳ Choosing K may b tough
↳ Large no. of samples
↳ More time to classify a new example.

x	y	class
7	7	B
7	4	B
3	4	G
1	4	G

$k = 3$

New instance (3, 7)

Predict the class.

Distance:

$$|(x_2 - x_1) + (y_2 - y_1)|$$

B $|(3-7)+(7-7)| = 4$ ② B

B $|(7-3)+(4-7)| = 7$ ④ B

G $|(3-3)+(4-7)| = 3$ ① G

G $|(1-3)+(4-7)| = 5$ ③ G

If, $k=1$; Good

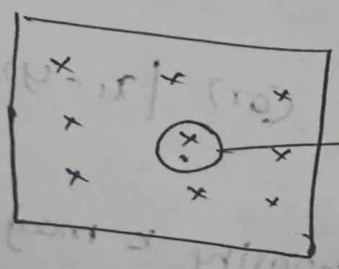
$k=2$; Bad

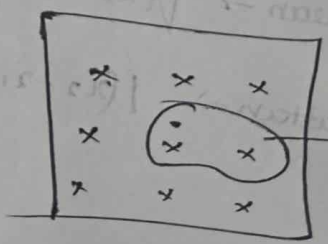
$k=3$; Good

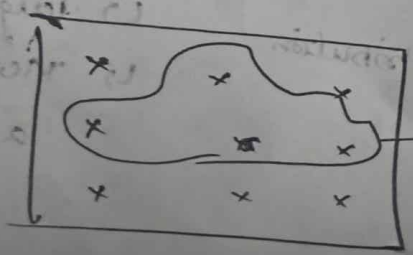
~~$k=4$; Bad~~

∴ If $k=3$ given then (3, 7) is good.

Majority → Good.







$K = 5$ gn.

Height	weight	class underw/Normal	Distance
167	51	UW	7 - (5)
182	62	N	19
176	69	N	20
173	64	N	12 - (3)
172	65	N	12
169	56	UW	2 - (2)
173	58	N	6 - (4)
170	57	N	2 - (1)
174	56	N	5 - (3)
170	55	?	?

(170, 55) -> ? Instance

$|(170 - 167) + (55 - 51)| = 3 + 4 = 7$

If $K = 1 \rightarrow N$

$K = 2 \rightarrow UW$ ✓

$K = 3 \rightarrow N$

$K = 4 \rightarrow$ ~~UW~~ N

$K = 5 \rightarrow UW$ ✓

$K = 1 \rightarrow N$

$K = 2 \rightarrow UW$

$K = 3 \rightarrow N$

$K = 4 \rightarrow N$

$K = 5 \rightarrow UW$

Majority Normal

(170, 55) -> Normal

③ Baye's theorem:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

B -> evidence

A -> hypothesis.

P(playtennis = "yes") = $\frac{9}{14}$ = 0.643

P(playtennis = "No") = $\frac{5}{14}$ = 0.357

conditional probability;

Outlook	Y	N
S	2/9	3/5
R	3/9	2/5
O	4/9	0

Temp	Y	N
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Hum	Y	N
High	3/9	4/5
Nor	6/9	1/5

Windy	Y	N
S	3/9	3/5
W	6/9	2/5

compute P(x|ci) for each class;

P[outlook = "sunny" | playtennis = "yes") = 2/9 = 0.222

④ Linear Regression :-

least square method:-

$$y = b_0 + b_1 x$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\text{Slope } (m) = \frac{S_{xy}}{S_{xx}}$$

Simple linear regression

$$\hat{y} = b_0 + b_1 x + e$$

No. of hr	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$
1	14	-1	-6	1
3	24	1	4	1
2	18	0	-2	0
1	17	-1	-3	1
3	27	1	7	1
10/2=5	20			4

$\bar{x}=2$ $\bar{y}=20$

Slope $b_1 = \dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$

$\Rightarrow \dfrac{(-1\times-6)+(+1\times4)+(0\times-2)+(-1\times-3)+(1\times7)}{4}$

$b_1 \Rightarrow \dfrac{6+4+0+3+7}{4} = \dfrac{20}{4} = 5$

y-Intercept

$b_0 = \bar{y} - b_1\bar{x}$

$= 20-5(2)$

$= 20-5(2)$

$= 10$

$\hat{y} = 10+5x$

Least square method:

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	2	-2	-2	4	4
2	4	-1	0	0	1
3	5	0	1	0	0
4	4	1	0	0	1
5	5	2	1	2	4
3	4			6	10

$$\text{m.slope} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2} = \frac{6}{10} = \frac{3}{5}$$

$y = mx + c \rightarrow$ intercept

$\bar{y} = 4 \quad \bar{x} = 3$

$$4 = \frac{6/3}{10/5} \times 3 + c$$

$$4 = \frac{3}{5} \times 3 + c$$

$$4 = \frac{9}{5} + c$$

$$\frac{9}{5} + c = 4$$

$$\frac{9}{5} - 4 = c$$

$$\frac{9-20}{5} = c$$

$$\boxed{c = 2.2}$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
2	50	-3.2	-25	80	10.24
3	60	-2.2	-15	33	4.84
5	80	-0.2	5	-1	0.04
7	90	2.2	15	33	4.84
9	95	3.8	20	76	14.44
				221	34.4
5.2	75				

$$m = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$m = \frac{3.8 \times 20}{34.4} = \frac{221}{34.4} = 6.424$$

$y = mx + c$

$75 = 6.424 + 5.2 + c$

$\boxed{c = 41.6}$

Predict for $x = 4$;

$y = 6.424 \times 4 + 41.6 = 67.3$

$\boxed{y = 67.3}$

$y = mx + c$

$y = 6.424 \times 4 + 41.6$

R^2 coefficient of determination:

$R^2 = \frac{SSR}{SST}$ → sum of square reg.

SST → Total sum of reg.

$\boxed{SST = SSR + SSE}$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	14	-1	-6	6	1
3	24	1	4	4	1
2	18	0	-2	0	0
1	17	-1	-3	3	1
3	27	1	7	7	1
2	20			20	4

$$m = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2} = \frac{20}{4} = 5$$

$$y = mx + c$$

$$20 = 5 \times 2 + c$$

$$\boxed{c = 10}$$

$$\hat{y} = mx + c \Rightarrow \hat{y} = 5x + 10$$

	$y-\bar{y}^2$	y	$\hat{y}-\bar{y}$	$(\hat{y}-\bar{y})^2$
1	36	15	5	25
3	16	25	5	25
2	4	20	0	0
1	9	15	5	25
3	49	25	5	25
	114	X	X	100

$5 \times 2 + 10$

$$R^2 = \frac{\Sigma (\hat{y}-\bar{y})^2}{\Sigma (y-\bar{y})^2}$$

$$R^2 = \frac{100}{114}$$

$$= 0.877$$

$$R^2 \Rightarrow 88.1\%$$

(4) Derivation for sigmoid fn

(5) Multiple logistic reg:

given:-

$\beta_0 = 0.67$ $\beta_2 = 0.47$

$\beta_1 = 0.58$ $\beta_3 = 0.36$

Age x_1	BMI x_2	BP x_3	Dia y
45	28	140	1
50	25	135	0
40	30	150	1
35	24	125	0

Predict age 25

BMI 20

BP = 120

Dial = ?

$x_1 = 25$

$x_2 = 20$

$x_3 = 120$

$Y = ?$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

$$z = 0.67 + 0.58 \times 25 + 0.58 \times 20 + 0.36 \times 120$$

$$z = 67.77$$

$$\hat{\beta} = \frac{1}{1+e^{-z}} = \frac{1}{1+0} = 1$$

5 (1) -> 10×4 = 40
1 (10) -> 1×10 = 10
50

DMP.

Problem:

- ✓ * Decision Tree
- ✓ * KNN
- ✓ * Linear Regression
- * Bayes

Theory:

same
- * Multiple logistic Regression
- * logistic regression
- * Generalized linear model
- * Derivation sigmoic fun.

① Decision Tree:

$$Info(D) = -\sum P_i \log_2(P_i)$$

$$Gain_{(A)} = Info(D) - Info_A(D)$$

SNO	Outlook	temperature	Humidity	windy	Playtennis
1	S	H	High	weak	No
2	S	H	H	Strong	NO
3	O	H	H	W	Yes
4	R	M	H	W	Yes
5	R	C	Normal	W	Yes
6	R	C	N	S	NO
7	O	C	N	S	Yes
8	S	M	H	W	NO
9	S	C	N	W	Yes
10	R	M	N	W	yes
11	S	M	N	S	yes
12	O	M	H	S	yes
13	O	H	N	W	yes
14	R	M	H	S	No

Total = 14. Yes = 9, NO = 5 0.41

$$\text{Info}(D) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

Info (D) = 0.940

Attributes

Outlook

	Y	N	
S	2	3	0.44
O	4	0	0.53
R	3	2	0.5

$$\text{Info}_{\text{outlook}}(D) = \frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4} - 0\right) + \frac{5}{14}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$$ 0.34 0.31

$$= \frac{5}{14}(0.97) + \frac{4}{14}(1) + \frac{5}{14}(0.97)$$

$$= 0.35 + 0.29 + 0.35$$

$$= 0.70$$

Info temp (D) =

temp	Y	N
H	2	2
M	4	2
C	3	1

$$\text{Info}_{\text{temp}}(D) = \frac{4}{14}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right] + \frac{6}{14}\left[-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right] +$$

$$\frac{4}{14}\left[\frac{-3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right]$$

$$= \frac{4}{14}(\cdot 1) + \frac{6}{14}(0.92) + \frac{4}{14}(0.81)$$

$$= 0.29 + 0.39 + 0.23$$

$$= 0.91$$

Humidity:

	Y	N
H	3	4
N	6	1

0.52 +
0.12

$$\text{Info}_{\text{Hum}}(D) = \frac{7}{14}\left[\frac{-3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right] +$$

$$\frac{7}{14}\left[\frac{-6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right]$$

$$= \frac{7}{14}(0.98) + \frac{7}{14}(0.59)$$

$$= 0.49 + 0.30$$

$$= 0.788.$$

0.1029
0.4284

$\text{Info}_{\text{windy}}(D) =$ 0.3113

		Y	N
windy	W	6	2
	S	3	3

$$\text{Info}_{\text{windy}}(D) = \frac{8}{14}\left[\frac{-6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right] +$$

$$\frac{6}{14}\left[\frac{-3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right]$$

$$= \frac{8}{14}(0.8113) + \frac{6}{14}(1) \qquad 0.4636$$

$$\simeq 0.8922$$

Gain (A) = Info(D) - Info$_A$(D)

Gain(Outlook) = 0.940 - 0.710

= 0.24 ✓

Gain(Temp) = 0.940 - 0.911

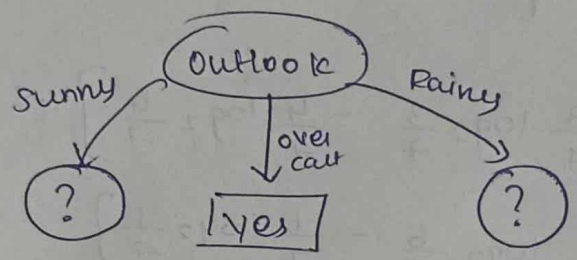
= 0.029

Gain(Hum) = 0.940 - 0.788

= 0.152

Gain(wind) = 0.940 - 0.892

= 0.048.



Sunny, Rainy table.

Outlook	Temp	Hum	Windy	PT
Sunny	Hot	High	Weak	NO
	Hot	High	Strong	NO
	Mild	High	Weak	NO
	Cold	Nor	Weak	yes
	Mild	Nor	Strong	yes

Yes = 2, NO = 3

$$Info(sunny) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.5288 + 0.4422$$

Sunny = 0.971

Attributes, Temp

	Y	N
H	0	2
M	1	1
C	1	0

$$Info_{(Temp)}(D) = \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2}\right) + \frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1}\right)$$

$$= \frac{2}{5}(0) + \frac{2}{5}(1.0) + \frac{1}{5}(0)$$

$$Info_{Temp}(D) = 0.4$$

Hum,

	Y	N
H	0	3
N	2	0

$$Info_{Hum}(D) = \frac{3}{5}\left(-\frac{3}{3}\log_2\frac{3}{3}\right) + \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2}\right)$$

$$= 0$$

Windy

	Y	N
W	1	2
S	1	1

0.5283

$$Info_{wind}(D) = \frac{3}{5}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

$$= \frac{3}{5}(0.9183) + \frac{2}{5}(1)$$

$$= 0.5510 + 0.4$$

$$= 0.9510$$

Gain (Temp) = 0.971 - 0.4

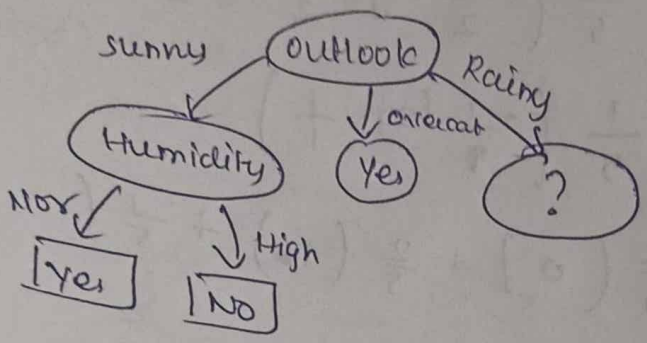
= 0.5710

Gain(Hum) = 0.971 - 0

= 0.971 ✓

Gain(wind) = 0.971 - 0.9510

= 0.02



Rainy table :-

Outlook	Temp	Hum	Windy	PT
Rainy	Mild	High	W	Y
	Cool	Nor	W	Y
	Cool	Nor	S	N
	Mild	Nor	W	Y
	Mild	High	S	N

0.3900

0.4422

Rainy = Yes = 3, NO = 2

$$Info(Rainy) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

= 0.9710

Attributes; Temp

	Y	N
M	2	1
C	1	1

$$\text{Info}_{\text{temp}}(D) = \frac{3}{5}\left[-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right]$$

$$= \frac{3}{5}(0.9183) + \frac{2}{5}(1)$$

$$= 0.9510$$

Hum,

	Y	N
H	1	1
N	2	1

$$\text{Info}_{\text{Hum}}(D) = \frac{2}{5}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] + \frac{3}{5}\left[-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right]$$

$$= \frac{2}{5}(1) + \frac{3}{5}(0.9183)$$

$$= 0.9510$$

windy,

	Y	N
W	3	0
S	2	2

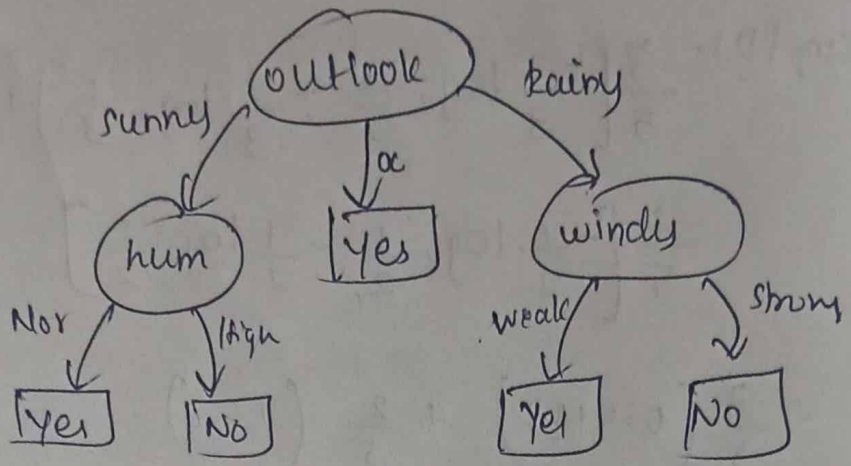
$$\text{Info}_{\text{windy}}(D) = \frac{3}{5}\left(-\frac{3}{3}\log_2\frac{3}{3} - 0\right) + \frac{2}{5}\left(-\frac{2}{5}\log_2\frac{2}{5}\right)$$

$$= 0.$$

$$\text{Gain}(\text{Temp}) = 0.9710 - 0.9510 = 0.026$$

$$\text{Gain}(\text{~~Wind~~ Hum}) = 0.9710 - 0.9510 = 0.020$$

$$\text{Gain}(\text{~~Hum~~ wind}) = 0.9710 - 0 = 0.9710 \checkmark$$



(2.) KNN:

- Non-parameteric
- classification & Regression
- lazy learner algo.

x	y	class
7	7	
7	4	Bad
3	4	Bad
1	4	Good
	4	Good

x_1 y_1

New instance (3, 7)

Give $K = 3$.

$$= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$= \sqrt{(x_2 - y_1)}$$ ↓ Euclidean

, $|(x_2 - x_1) + (y_2 - y_1)|$ ↓ Manhattan

B $|(7-3) + (7-7)| = |4 + 0| = 4$ (2)

B $|(7-3) + (4-7)| = |4 + (-3)| = 7$ (4)

G $|(3-3) + (4-7)| = |0 + (-3)| = 3$ (1)

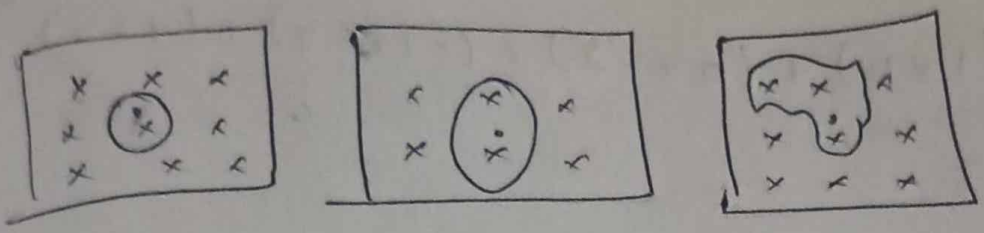
G $|(1-3) + (4-7)| = |-2 + (-3)| = 5$ (3)

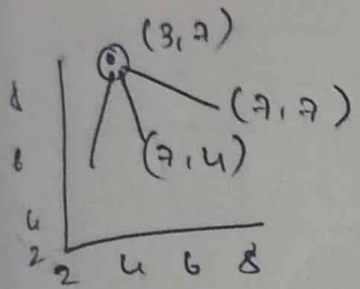
$K = 1$: Good

$K = 2$ Bad

$K = 3$ Good

$K = 3$ ∵ (3, 7) the good.





③ Linear regression:

a) least square method:

$$\boxed{y = b_0 + b_1 x}$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$) slope m

x	y	$x - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$
1	14	-1	1	-6
3	24	-1	1	4
2	18	0	0	-2
1	17	-1	1	-3
3	27	+1	1	7
10/5	100/5	-2	4	4

$\bar{x} = 2$ $\bar{y} = 20$

$b_1 = (-1 \times$

$= (-1 \times -6) + (1 \times 4) + (0 \times -2) + (-1 \times -3) + (1 \times 7)$

$= 6 + 4 + 3 + 7$

$= \frac{20}{4}.5$

$= 5$

$b_1 = 5$ slope.

$b_0 = \bar{y} - b_1 \bar{x}$

$y = b_0 + b_1 x$

$b_0 = \bar{y} - b_1 \bar{x}$

$b_0 = 20 - 5 \times 2$

$= 20 - 10$

$\hat{y}$.

$b_0 = 10$

$\hat{y} = b_0 + b_1 x + \epsilon$

$\hat{y} = 10 + 5x$

$(x - \bar{x})(y - \bar{y})$ x y $x_i - \bar{x}$ $y_i - \bar{y}$ $(x_i - \bar{x})^2$ $(y_i - \bar{y})^2$

b) R^2

$$R^2 = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2}$$

$\hat{y} = mx + c$.

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$\hat{y}$	$\hat{y}-\bar{y}$	$(\hat{y}-\bar{y})^2$
1	14	-1	-6	6	1	36	15	5	25
3	24	-1	4	4	1	16	25	5	25
2	18	0	-2	2	0	4	20	0	0
1	17	-1	-3	3	1	9	15	5	25
3	27	1	7	7	1	49	25	5	25
2	20			20		114			100.

$\frac{20}{4}$

$\hat{y} = 5x + 10$

$5(1) + 10$

$\hat{y} = mx + c$

$\boxed{y = 5x + 10}$

$\bar{y} = m\bar{x} + c$

$20 = 5 \times 2 + c$

$20 = 10 + c$

$10 + c = 20$

$c = 20 - 10$

$\boxed{c = 10}$

$$R^2 = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2}$$

$$R^2 = \frac{100}{114} = 0.877$$

$R^2 = 88\%$

Sigmoic Function

$$\text{odds}(\theta) = \frac{\text{Prob of event happening}}{\text{Prob of event NOT happening}}$$

$$\text{odds}(\theta) = \frac{p}{1-p}$$

Value of odds range from 0 to ∞

$$y = \beta_0 + \beta_1 x$$

Take log on both sides

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

Exponentiating both sides

$$e^{\ln \frac{p(x)}{1-p(x)}} = e^{\beta_0 + \beta_1 x}$$

$$\frac{p(x)}{1-p(x)} = e^{\beta_0 + \beta_1 x} \qquad [\because e^{\ln} = 1]$$

let $y = e^{\beta_0 + \beta_1 x}$

Then, $\frac{p(x)}{1-p(x)} = y$

$$p(x) = y\,[1-p(x)]$$

$$= y - y\,p(x)$$

$$p(x) + y\,p(x) = y$$

$$p(x)\,(1+y) = y$$

$$p(x) = y/1+y$$

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

divide $e^{\beta_0 + \beta_1 x}$

$$= \frac{e^{\beta_0 + \beta_1 x} / e^{\beta_0 + \beta_1 x}}{1 / e^{\beta_0 + \beta_1 x} + \frac{e^{\beta_0 + \beta_1 x}}{e^{\beta_0 + \beta_1 x}}}$$

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$