Bayes Theorem

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

↓ Event A when B has occured

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

Eg. n = 14

Eg. < outlook = sunny, Temp. = cool, Humidity = high, Wind = strong >

$P(\text{yes}) = \frac{5}{14}$, $P(\text{no}) = \frac{9}{14}$

$P(\text{sunny}|\text{yes}) = \frac{2}{5} \times \frac{14}{9} = \frac{28}{45}$

$P(\text{sunny}|\text{no}) = \frac{3}{5} \times \frac{14}{5} = \frac{42}{25}$

$P(\text{sunny}|\text{yes}) = \frac{5}{14} \times \frac{9}{14}$

$P(\text{sunny}|\text{yes}) = \frac{2}{5} \times \frac{14}{9} = \frac{28}{45}$ $\frac{2}{9}$

$P(\text{sunny}|\text{no}) = \frac{3}{5} \times \frac{14}{5} = \frac{3}{5}$

$P(\text{cool}|\text{yes}) = \frac{3}{9}$ $P(\text{high}|\text{yes}) = \frac{3}{9}$

$P(\text{cool}|\text{no}) = \frac{1}{45}$ $P(\text{high}|\text{no}) = \frac{4}{5}$

$P(\text{strong}|\text{yes}) = \frac{3}{9}$ $P(\text{st}|\text{no}) = \frac{3}{5}$

$$P(E) = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = \frac{54}{6561}$$

$$P(yes) = P(yes)\, P(sl|y)\, P(c|y)\, P(h|y)\, P(wt|y)$$
$$= 0.0053$$

$$P(no) = P(no)\, P(sl|n)\, P(c|n)\, P(h|n)\, P(wt|n)$$
$$= 0.0206$$

$$V_{NB}(yes) = \frac{V_{NB}(yes)}{V_{NB}(yes) + V_{NB}(no)} = 0.205$$

$$V_{NB}(no) = 0.795$$

Eg $X = \{Red, SUV, Domestic\}$

$$P(R|y) = \frac{P(R|R)\, P(R)}{P(R)} = \frac{\frac{3}{5} \times \frac{5}{10}}{\frac{5}{10}} = \frac{3}{5}$$

$$P(R|n) = \frac{\frac{2}{5} \times \frac{5}{10}}{\frac{5}{10}} = \frac{2}{5}$$

$$P(SUV|y) = \frac{\frac{1}{5} \times \frac{4}{10}}{\frac{5}{16}} = \frac{4}{25}$$

$$P(SUV|n) = \frac{\frac{3}{5} \times \frac{4}{10}}{\frac{5}{10}} = \frac{3}{5}$$

$$P(D|y) = \frac{\frac{2}{5} \times \frac{5}{10}}{\frac{5}{10}} = \frac{2}{5}$$

$$P(D|n) = \frac{\frac{3}{5} \times \frac{5}{10}}{\frac{5}{10}} = \frac{3}{5}$$

$$V_{NB}(y_i) = P(y_i)\,P(R|y_i)\,P(S|y_i)\,P(D|y_i)$$

$$= \frac{1}{2} \times \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5}$$

$$= \frac{3}{125}$$

$$V_{NB}(n) = \frac{12}{125}$$

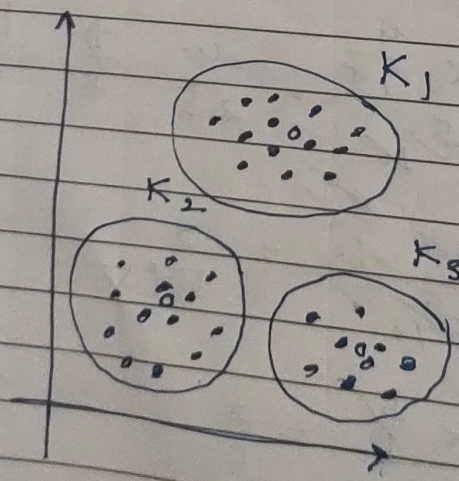
$$P(x|n) > P(x|yes)$$

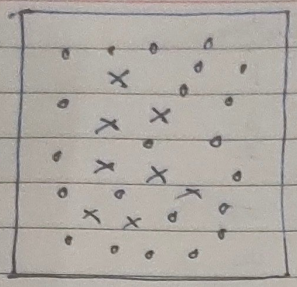
$\Rightarrow$ Not stolen

Clustering

→ no labelled data

* It is dist. based unsupervised ML alg. where D.P. close to each other are grouped in a given no. of clusters / groups



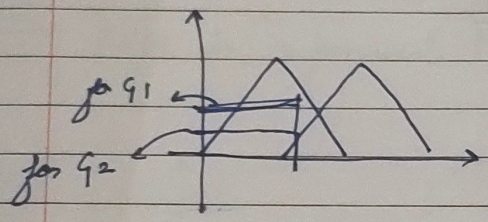


Two centres
C_1 C_2

Clustering

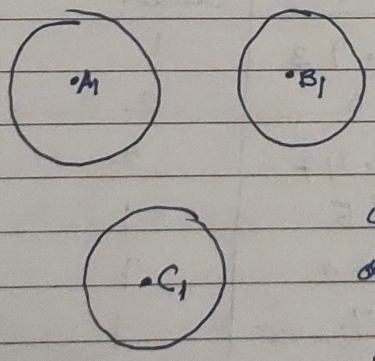
Hard clustering
↓
Each datapoint assigned to only a single cluster
(K-means, K-medoid)
KNN

Soft clustering
↓
Each datapoint belongs to a cluster within a certain prob. value known as Membership values
(Fuzzy C-Mean)
$A = \{a, \mu_a\}$
↓ element ↓ Membership value



Suppose we have following pts -

A1 (2, 10), A2 (2, 5), A3 (8, 4), B1 (5, 8), B2 (7, 5)
B3 (6, 4), C1 (1, 2), C2 (4, 9)



$$d_{euc} = \sqrt{(x_1 - x_i)^2 + (y_1 - y_i)^2}$$

$d_{A11} = 0$

$d_{A1A2} = \sqrt{5}$ ✓

$d_{A13} = 6\sqrt{2}$

$d_{B11} = 0$

$d_{B12} = \sqrt{13}$ ✓

$d_{B13} = \sqrt{17}$

$d_{C11} = 0$ $\quad d_{C12} = \sqrt{58}$ ✓

			(2,10)	(5,8)	(1,2)	Cluster
A_1	2	10	0	3.61	8.06	1
A_2	2	5	5	4.24	3.16	3
A_3	8	4	8.49	5	7.28	2
B_1	5	8	3.61	0	7.21	2
B_2	7	5	7.07	3.61	6.71	2
B_3	6	4	7.21	4.12	5.39	2
C_1	1	2	8.06	7.21	0	3
C_2	4	9	2.24	1.41	7.62	2

New centroid $= \frac{A_3 + B_1 + B_2 + B_3}{5}$

New cen. $A_1 = (2, 10)$

$B_1 = (6, 6)$

$C_1 = (1.5, 3.5)$

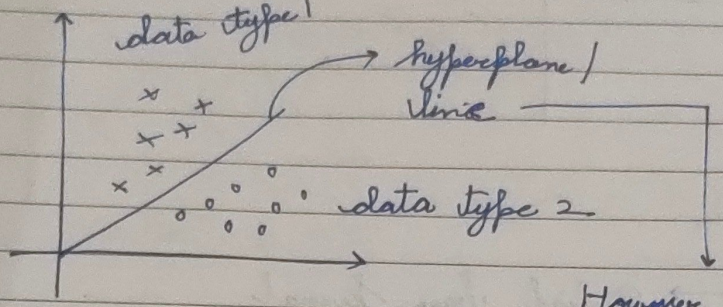
Repeat the process with new centroids until the centroids are stable.

Points		(185, 72)	(170, 56)	Cluster
185	72	0	21.93	1
170	56	24.515	0	2
168	60	18.78	4.47	2
179	68	7.21	15	1
182	72	3	20	1
188	77	5.83	27.65	1

New cen. $= [(183.5, 72.25), (169, 58)]$

SVM (Support Vector Machine)

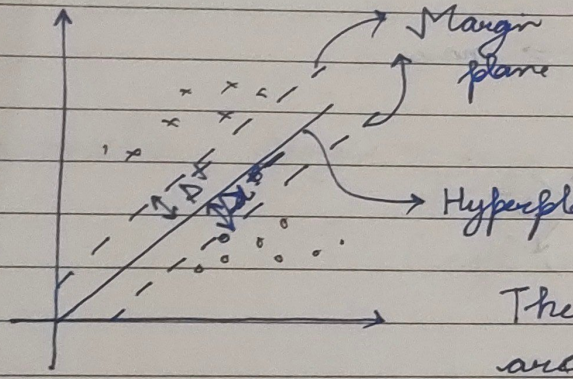
* Classifier
* Regression



However, low margin poses a problem

→ SVM is a supervised MLA

Hence SVM is used

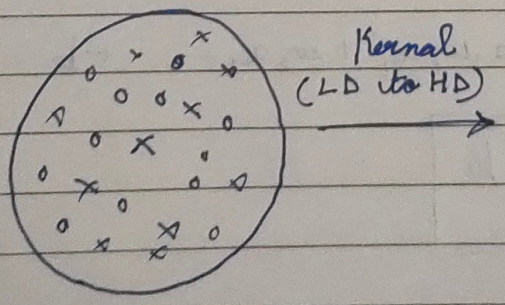


Margin $= D^+ + D^-$

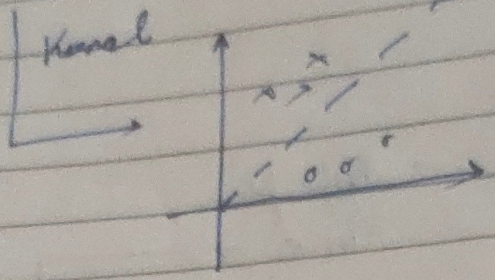
The pts. touching the margin are called support vector

The best classifier has the max. margin.

If data is non-linearly separable



Eg

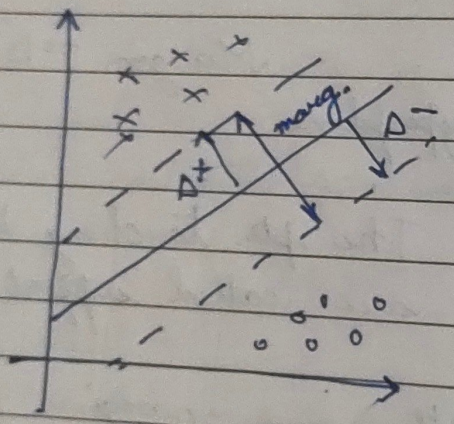


→ Kernal

- Radial basis kernal
- Quadratic kernal

Linear SVM

Since, margin → max.



$y = mx + c$, $y = b_1 x + b_0$

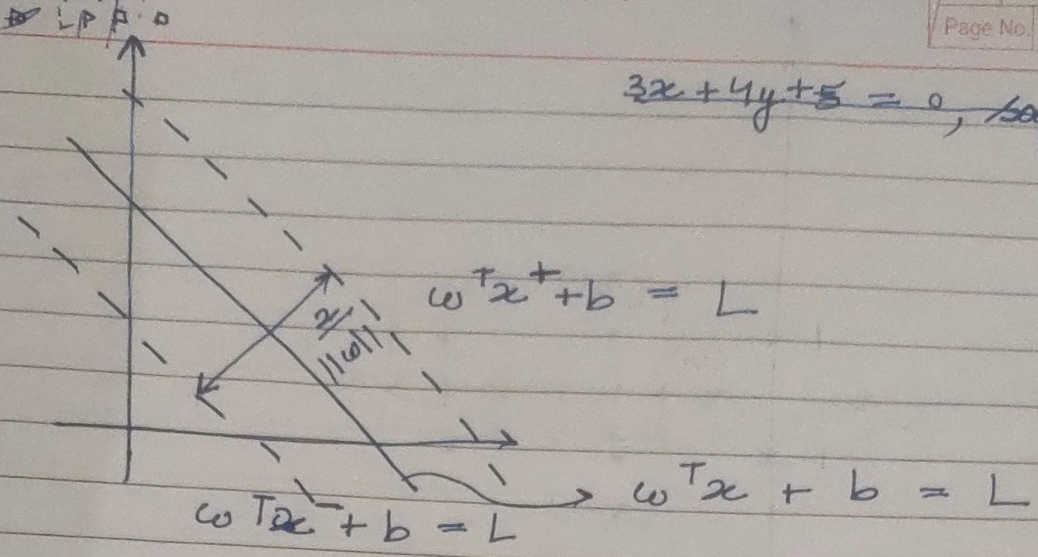
$y = \beta_1 x + \beta_0$

For multi var. $y = w_1 x_1 + w_2 x_2 + \dots + b$

$$\boxed{y = w_i^T x_j + b}$$

LPP :-

$3x + 4y + 5 = 0$, say



$$w^T (x^+ - x^-) = 2$$

$$x^+ - x^- = \frac{2}{\|w\|} = \text{Margin}$$

→ Loss function

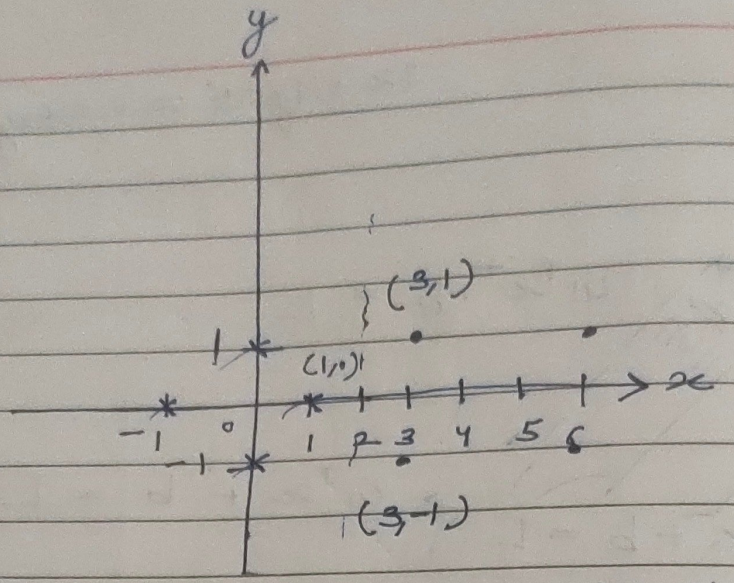
For multiclass, $\text{Margin} = \frac{2}{\|w\|} + \sum_{i=1}^{n} C_i \varepsilon_i$

(soft margin)

C_i → dist. of data pts. from margin plane

$(3,1), (3,-1), (6,1), (6,-1) \longrightarrow +1$

$(1,0), (0,1), (-1,0), (0,-1) \longrightarrow -1$



$S_1 \rightarrow (1,0)$, $S_2 = (3,1)$, $S_3 = (3,-1)$

(To find the line $y = mx + c$)

Bias

$\tilde{S}_1 (1,0,1)$, $\tilde{S}_2 (3,1,1)$, $\tilde{S}_3 (3,-1,1)$

$$\alpha_1 \tilde{S}_1 \tilde{S}_1 + \alpha_2 \tilde{S}_1 \tilde{S}_2 + \alpha_3 \tilde{S}_1 \tilde{S}_3 = -1$$
$$\alpha_1 \tilde{S}_2 \tilde{S}_1 + \alpha_2 \tilde{S}_2 \tilde{S}_2 + \alpha_3 \tilde{S}_2 \tilde{S}_3 = +1$$
$$\alpha_1 \tilde{S}_3 \tilde{S}_1 + \alpha_2 \tilde{S}_3 \tilde{S}_2 + \alpha_3 \tilde{S}_3 \tilde{S}_3 = +1$$

$$\alpha_1 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} + \alpha_3 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix}$$

$$\alpha_1 \begin{bmatrix} 1+0+1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 3+0+1 \end{bmatrix} + \alpha_3 \begin{bmatrix} 3+0+1 \end{bmatrix} = -1$$
$$= -1$$

$$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$$

Date
Page No.

Sim. $4\alpha_1 + 11\alpha_2 + 9\alpha_3 = 1$

and $4\alpha_1 + 9\alpha_2 + 11\alpha_3 = 1$

$$\begin{bmatrix} 2 & 4 & 4 \\ 4 & 11 & 9 \\ 4 & 9 & 11 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$$

$\alpha_1 = -3.5, \alpha_2 = \alpha_3 = 0.75$

$y = w^T x + c$

$w^T = \sum_i \alpha_i s_i$

$$= -3.5 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + 0.75 \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} + 0.75 \begin{bmatrix} +3 \\ -1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} -3.5 \\ 0 \\ -3.5 \end{bmatrix} + \begin{bmatrix} 0 \\ 1.5 \\ 1.5 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

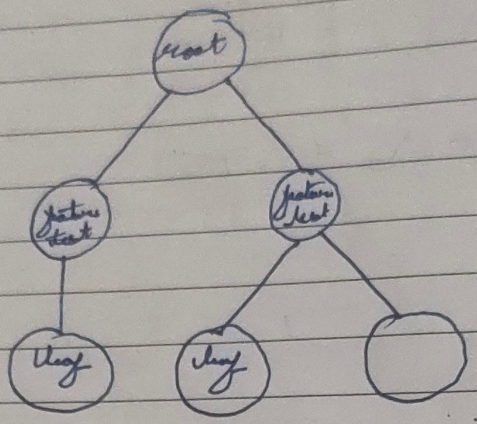
$$y = \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix} [x_1 \; x_2 \; x_3] + c$$

$\Rightarrow \boxed{y = 2}$

Decision Tree

Supervised algo

↳ regression + classification

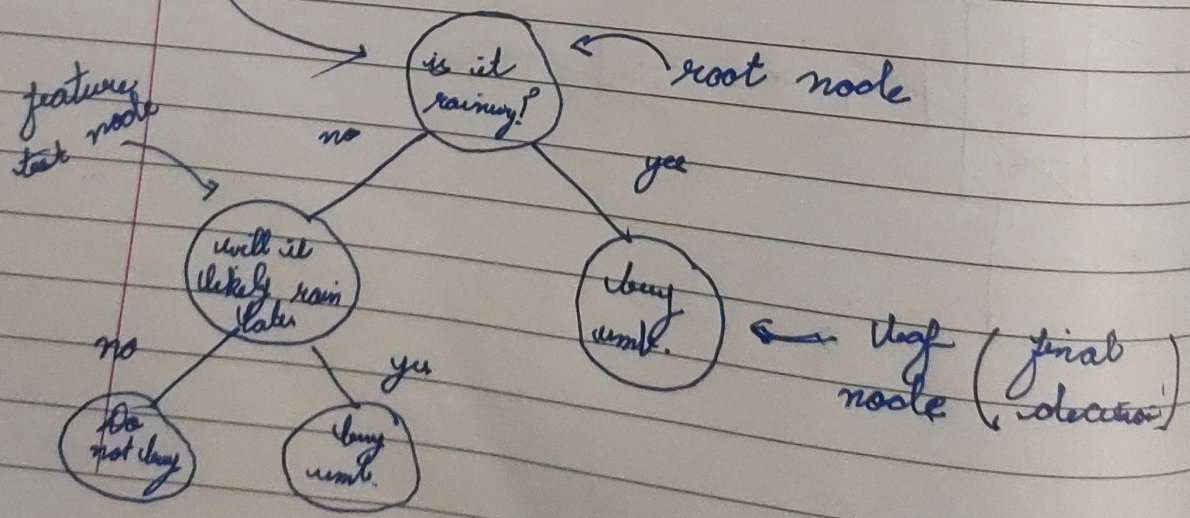



```
if (it is raining)
    print (Buy umbrella)

else if (it will likely rain later)
    print (Buy umbrella)

else
    print (Do not buy)
```

Eg of how dec. tree works



→ Information Gain
→ Gini Index

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

Eg. Values (outlook) = {sunny, overcast, rain}

$S = [9+, 5-]$

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

$$\Rightarrow \text{Entropy}(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

$$S_{\text{sunny}} = [2+, 3-] \Rightarrow -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

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

$$S_{\text{rain}} = [3+, 2-] \Rightarrow -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= 0.971$$

$$\text{Gain}(S, \text{outlook}) = 0.94 - \frac{5}{14}(0.971) - \frac{4}{14}(0) - \frac{5}{14}(0.971)$$

Info. Gain = 0.2464

Now, for attribute temp.

$S = [9+, 5-]$, $E(S) = -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14}$

$= 0.94$

$S_{hot} = [2+, 2-] \rightarrow E(S) = 1$

S_{mild}, $E(S) = 0.9183$

S_{cool}, $E(S) = 0.8113$

$Gain = 0.94 - \frac{4}{14} - \frac{6}{14}(0.9183) - \frac{4}{14}(0.811)$

$= 0.0289$

Gain for humidity → high, low

$S = [9+, 5-]$, $E(S) = 0.94$

$S_{high} = [3+, 4-]$, $E(S) = -\frac{3}{7}\log_2\frac{3}{7}$

$- \frac{4}{7}\log_2\frac{4}{7} = 0.985$

$S_{normal} = [6+, 1-]$, $E(S) = -\frac{6}{7}\log_2\frac{6}{7}$

$- \frac{1}{7}\log_2\frac{1}{7} = 0.591$

$Gain = 0.1516$

~~wind~~

Simi., for wind

$S = [9+, 5-]$, $E(S) = 0.94$

$S_{\text{Sunny}} = [3+, 3-]$, $E(S) = 1$

Gain = 0.0478

Info. gain of outlook is highest ← root node

Day	Temp	Hum.	Wind	Play Tennis
D1	H	H	W	N
D2	H	H	S	N
D8	M	H	W	Y
D9	C	N	W	Y
D11	M	N	S	Y

Table afterwards for sunny

Gain (S_S, temp) = 0.570
Gain (S_S, hum.) = 0.97 ⟶ next root node
Gain (S_S, W) = 0.0192

Now, for rainy

~~Gain (S_R, H) = 0, 0~~
Gain (S_R, T) = 0.0192
Gain (S_R, H) = 0.0192
Gain (S_R, H) = 0.97 ⟶ next node

Drawback of dec. tree ⟶ overfitting

Gini Index

$$G(S) = 1 - \sum_{i=1}^{n} P_i^2$$

* Root have min. gini index

Eg Finding $G(S)$ of decision

$$G(S) = 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)^2\right]$$
$$= 0.58$$

Now, for attribute money & case - poor

$$G(S) = 1 - \left(\frac{3}{3}\right)^2 = 0 \quad \left(\frac{3}{3} \text{ as all 3 have same o/p \& dec.}\right)$$

For rich

$$G(S) = 1 - \left[\left(\frac{2}{7}\right)^2 + \left(\frac{3}{7}\right)^2 + \left(\frac{1}{7}\right)^2 + \left(\frac{1}{7}\right)^2\right]$$
$$= 0.694$$

Weighted avg (money)

$$= 0\left(\frac{3}{10}\right) + 0.694\left(\frac{7}{10}\right) = 0.486$$

We for parent:-

→ Yes → No

$Gini(S) = 1 - \left(\frac{5}{5}\right)^2 = 0$

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

$= 1 - 0.8$

$= 0.72$

→

Weighted avg = 0.36

Now for weather :-

→ Sunny

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

$= 0.44$

→ Windy

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

$= 0.375$

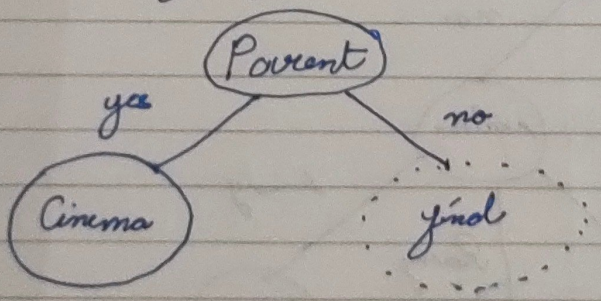
→ Rainy

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

$= 0.44$

Weighted avg $= 0.44 \cdot \frac{3}{10} + 0.44 \cdot \frac{3}{10} + 0.375 \cdot \frac{4}{10} = 0.414$

⇒ I nodes is for parent



Now, finding $Gini(S)$ for all those cases when parent said no.

Weather —

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

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

→ Windy
$1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$
$= 0.5$

Weighted avg $= 0.5 \times \frac{2}{5} = 0.2$

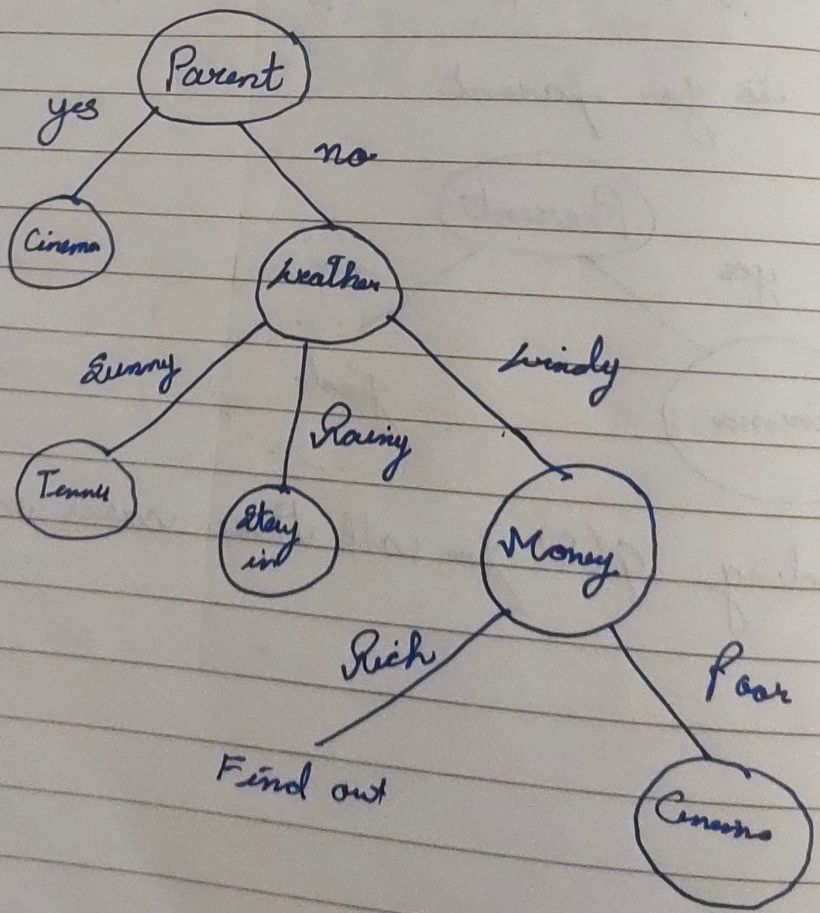
Money —

→ Rich
$G(S) = 1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{2}\right)^2\right]$
$= 0.375$ 625

Poor
$G(S) = 0$

Wt. avg $= 0.375$ 625 $\times \frac{4}{5} = 0.3$ 25

⇒ Next root is weather



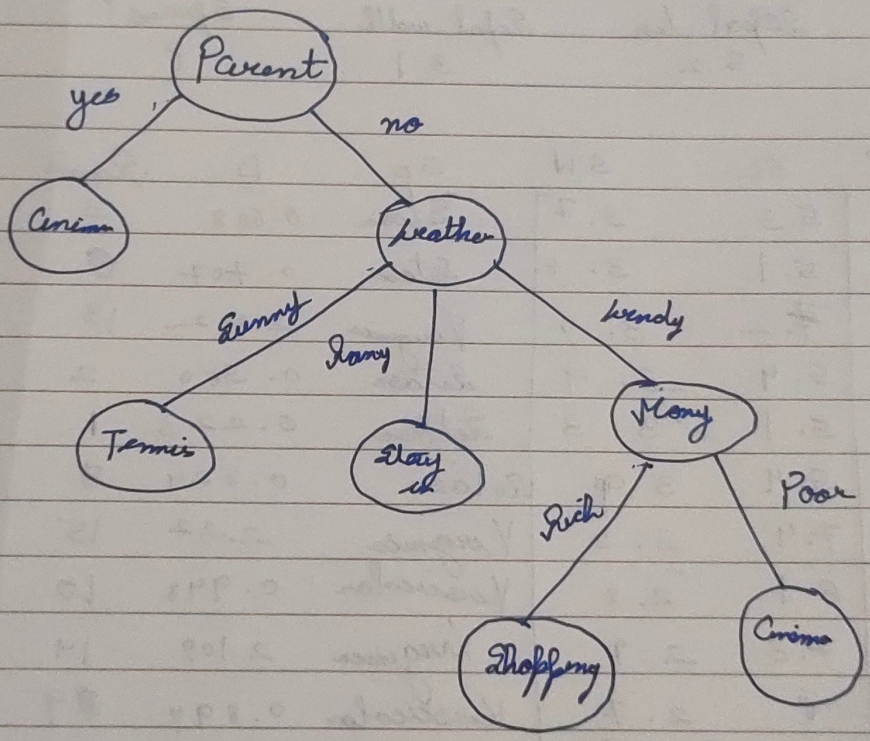
W2	S	R	T	
W5	R	R	S1	
W9	W5	R	S	✓
W10		R	T	

For weather →

→ Sunny → Rainy → Windy

$q(S) = 1 - 1 = 0$ $q(S) = 0$ $q(S) = 0$

⇒



KNN – K Nearest Neighbours

→ Supervised

→ Classif + regression

* Distance $(x, y) = \sqrt{(x-a)^2 + (y-b)^2}$

Given data —

Sepal len.	Sepal width	Species?
5.2	3.1	?

Given data	SL	SW	Sp	D	Rank
	5.3	3.7	Setosa	0.608	3
	5.1	3.8	Setosa	0.707	6
	7.2	3.0	Virginica	2.002	13
	5.4	3.4	Setosa	0.360	2
	5.1	3.3	Setosa	0.223	1
	5.4	3.9	Setosa	0.824	8
	7.4	2.8	Virginica	2.22	15
	6.1	2.8	Versicolor	0.948	10
	7.3	2.9	Virginica	2.109	14
	6	2.7	Versicolor	0.894	9
	5.8	2.8	Virginica	0.67	4
	6.3	2.3	Versicolor	1.36	12
	5.1	2.5		0.6708	
	6.3	2.5	"	0.6008	5
	5.5	2.4	"	1.25	11
			"	0.7615	7

If $k = 1$ ⟶ choose the class with the shortest Dist i.e Setosa

If $k = n$ ⟶ choose the n^{th} shortest Dist.