Unit-I
Introduction to ML

Intro to Well proposed Machine Learning:-
- i) Intro to well proposed ML.
- ii) Intro to supervised & unsupervised algo.
- imp iii) Decision tree algorithm
- iv) K nearest neighbor algorithm (KNN)
- v) Applications of ML.

ML has 3 fields → Experience, Tasks, Performan

Supervised → Using labels (correct data is presented)
Unsupervised → No labels are provided hence we use clustering method.

imp * K nearest neighbour algorithm :- learning. → also called instance based

The most instance based method is K nearest neighbour learning. This algo assumes all the instances corresponds to the points in all n-dimensional space R^n. Where R is arbitrary axis btw the data points.

$$d(x_i, x_j) = \sum_{r=1}^{n} \sqrt{[a_1(x_i) - a_2(x_j)]^2}$$ Euclidian distance.

The nearest neighbour of an instance are defined in terms of standard Euclidian distance

Steps

i) Let an arbitrary instance 'x' described by feature vector (data points) $a_1(x)$, $a_2(x)$, $a_3(x)$ .... $a_n(x)$

ii) Where $a_r(x)$ denotes the value of the r^{th} attribute of instance 'x'. The distance btw two instances is calculated using Euclidian distance.

$$\left[d(x_i, x_j) = \sum_{r=1}^{n} \sqrt{[a_1(x_i) - a_2(x_j)]^2} \right]$$

Q1) Following dataset contains types of tissues produced by a company & its clients responds

Type of Tissue	Acid Durability	Strength	Class
Type 1	7	7	Bad
Type 2	7	4	Bad
Type 3	3	4	Good
Type 4	1	4	Good

Find the class label of the new instance (test data with acid durability = 3 and strength = 7). using K data point (min distance) ie K = 3 Using the instance based algorithm classify the above sample data set.

Soln:

Step 1:

Given: $K = 3$

acid durability $= 3$

Strength $= 7$

Step 1: Calculating the distance for the given test data $(3, 7)$

Type of Types	Acid durability	Strength	$(3,7)$ Distance
1	7	7	4
2	7	4	5
3	3	4	3
4	1	4	3.6

Distance $(3, 7)$:

1. $\sqrt{(3-7)^2 + (7-7)^2} = 4$
2. $\sqrt{(3-7)^2 + (7-4)^2} = \sqrt{16+9} = \sqrt{25} = 5$
3. $\sqrt{(3-3)^2 + (7-4)^2} = \sqrt{9} = 3$
4. $\sqrt{(3-1)^2 + (7-4)^2} = \sqrt{4+9} = \sqrt{13} = 3.6$

Step 2:

Using k^{th} minimum distance assign the rank for the given distance

→

Rank = k (For good or bad)

(k = 3)

Type	Acid	Strength	Distance (3, 7)	Rank (3)
1	7	7	4	3 Yes
2	7	4	5	4 X No
3	3	4	3	1 Yes
4	1	4	3.6	2 Yes

the highest rank to ignore for test

Step 3:

Recalculate using the same test data (3, 7)

(Replace the rank value '4' with acid durability to recalculate the nearest neighbour data using (3, 7))

Type	Acid	Strength	Distance	Rank
1	7	7	4	4 N
2	4	4	$\sqrt{10} = 3.16$	2 Y
3	3	4	3	1 Y
4	1	4	3.6	3 Y

Hence the test data (3, 7) is added into the class good.

→

IB.

imp 2)

Height	Weight	Class
167	51	Under Weight
182	62	Normal.
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Under weight
169	58	Normal
173	57	Normal
170	55	Normal
170	57	?

@ K = 1, 2, 5. [Check for all]
Find class for (170, 57).

For K = 1 :

Step 1 :

Height	Weight	Distance	Class	Rank
167	51	6.70	UW	5
182	62	13	N	
176	69	13.41	N	
173	64	7.61	N	
172	65	8.24	N	
174	56	4.12	UW	4
169	58	1.41	N	1
173	57	3.	N	3
170	55	2	N	2

For k = 1 class is Normal
k = 2 class is Normal

For k = 5 → 3 Normal, 2 UW
hence K = 5 is in class Normal.
Hence the point (170, 57) is classified as Normal

3) Calculate the class for the test data $x_1 = 28$, $x_2 = 26$ $K = 5$ for the given attribute set

x_1	x_2	Class : y
2	4	A
4	2	A
4	4	B
6	6	B
8	6	B
10	10	B
12	8	A
14	10	A
16	8	A
18	12	A
20	14	A
22	18	B
24	20	B
26	22	B
28	24	B
30	26	B
32	28	B
34	30	B
36	32	A
38	36	A

Step 1: (28, 26) K = 5

x_1	x_2	y	Distance	
2	4	A	34.05	
4	2	A	33.94	
4	4	B	32.55	
6	6	B	29.73	
8	6	B	28.28	
10	10	B	24.08	
12	8	A	24.08	
14	10	A	21.26	
16	8	A	21.63	
18	12	A	17.20	
20	14	A	14.42	
22	18	B	10	
24	20	B	7.21	
26	22	B	4.47	4
28	24	B	2	1
30	26	B	2	2
32	28	B	4.47	3
34	30	B	7.21	5
36	32	A	10	
38	36	A	14.14	

For K = 5 the dataset (28, 26) is classified ~~in~~ into B.

4) To the following data set give the classification by considering k=3. test data (66, 03)

Student no	mark1	mark2	grade
S1	87	17	A
S2	20	06	B
S3	25	12	B
S4	93	75	A
S5	91	52	B.

Step 1:-

Std No	Mark1	Mark 2	Distance	Grade	Rank
S1	87	17	25.23	A	1
S2	20	06	46.09	B	3
S3	25	12	41.97	B	2
S4	93	75	76.89	A	5
S5	91	52	55.00	B	4

Step 2:- Recalculate

Std No	Mark1	Mark 2	Distance	Grade	Rank
S1	87	17	25.23	A	1
S2	20	06	46.09	B	3
S3	25	12	41.97	B	2
S4	5	75	94.36	A	5
S5	91	52	55.00	B	4

$$\log_n m = \frac{\log m}{\log n}$$

9/2/24

* **Decision Tree Problem :**

if classification is diff change the sign to the classification

$$\text{Entropy} = -P_{\oplus} \log_2 P_{\oplus} - P_{\ominus} \log_2 P_{\ominus}$$

i) What is Entropy of this collection of training examples wrt the target function classification.

ii) What is the information gain of attribute a_1 & a_2 related to these training examples

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

iii) Draw the decision tree for the given dataset.

[If the classification numbers are equal i.e 3+ & 3- = 1]

Instance	classification	a_1	a_2
1	+	T	T
2	+	T	T
3	−	T	F
4	+	F	F
5	−	F	T
6	−	F	T

Entrop[3+, 3-] = 1 Equal classification then = 1

For a_1 attribute :

→

(From instances 1, 2, 3)

→ True

Entropy $[S_T] = [2+, 1-] = \frac{3}{6}$

$= -\left(\frac{2}{3}\right) \times \log_2(2/3) - \left(\frac{1}{3}\right) \times \log_2\left(\frac{1}{3}\right)$

$= 0.918$

Entropy $[S_F] = [1+, 2-]$

$= -\left(\frac{1}{3}\right) \times \log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right) \times \log_2(2/3)$

$= 0.918$

For Information gain (S, a_1)

→ Entropy of whole dataset

$= \text{Entropy}(S) - \sum \frac{|S_v|}{|S|} \text{Entropy}(S_T)$

$= 1 - \left[\frac{3}{6} \times \text{Entropy}(S_T) + \frac{3}{6} \text{Entropy}(S_F)\right]$

$= 1 - [0.918]$

$= 0.082$

For a2 attribute :

Entropy $[S_T] = [2+, 2-] = 1 - \left(\frac{2}{4}\right) \times \log_2(2/4) - \left(\frac{2}{4}\right) \log_2(2/4)$

Entropy $[S_F] = [1+, 1-] = 1$

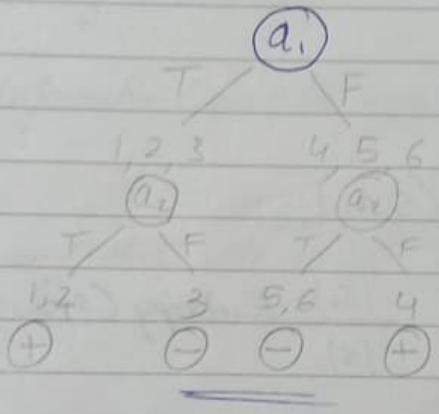
(S, a_2)

Information gain$_1$ = Entropy (S) − $\sum \frac{|S_v|}{|S|}$ Entropy (S_v)

$= 1 - \left[\frac{4}{6} \times 1 + \frac{2}{6} \times 1\right]$

$= 0$

Attribute with higher info gain is the root node. in this case a_1



2) Give the decision tree for the following set of training examples

Day	Outlook	Temp	Humidity	Wind	PlayTen
D1					
D2					
D3					
D4					
D5					
D6					
D					

attribute

classmate Date Page

Day	Outlook	Temp	Win Humidity	Wind	Play Tennis
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Wk	Yes
D_4	Rain	Mild	High	Wk	Yes
D_5	Rain	Cool	Normal	Wk	Yes
D_6	Rain	Cool	Normal	S	No
D_7	Overcast	Cool	Normal	S	Yes
D_8	Sunny	Mild	High	Wk	No
D_9	Sunny	Cool	Normal	Wk	Yes
D_{10}	Rain	Mild	Normal	Wk	Yes
D_{11}	Sunny	Mild	Norm	S	Yes
D_{12}	Overcast	Mild	High	S	Yes
D_{13}	Overcast	Hot	Norm	Wk	Yes
D_{14}	Rain	Mild	High	Strn	No
	0.248	0.029	0.152	0.048	

Ans: Entropy of whole data

Entropy[9yes, 5no] = 0.9402

Outlook → Sunny, Rain, Overcast.

For outlook:

i) Sunny

Entropy $[S_{sunny}]$ = [2 Yes, 3 No]

= 0.970

& Entropy $[S_{rain}]$ = [3y, 2no]

= 0.970

Entropy $[S_{overcast}]$ = [4yes, 0no]

= 0

Information Gain :

$$Ent(S) = \sum \frac{|S_v|}{|S|} Ent(S_T)$$

$$= 0.9402 - \left[\frac{5}{14} \times 0.97 + \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0\right]$$

$$= [0.248]$$

For Temp

Entropy $[S_{hot}] = [2Y, 2N] = 1$

Entropy $[S_{midd}] = [4Y, 2N] = 0.918$

Entropy $[S_{cool}] = [3Y, 1N] = 0.811$

Info gain :

$$= 0.9402 - \left[\frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.81\right]$$

$$[= 0.029]$$

For humidity :

Entropy $[S_{high}] = [3Y, 4N] = 0.985$

Entropy $[S_{normal}] = [6Y, 1N] = 0.591$

Info gain :-

$$0.9402 - \left[\frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.591\right]$$

$$= \boxed{0.1522}$$

For wind :-

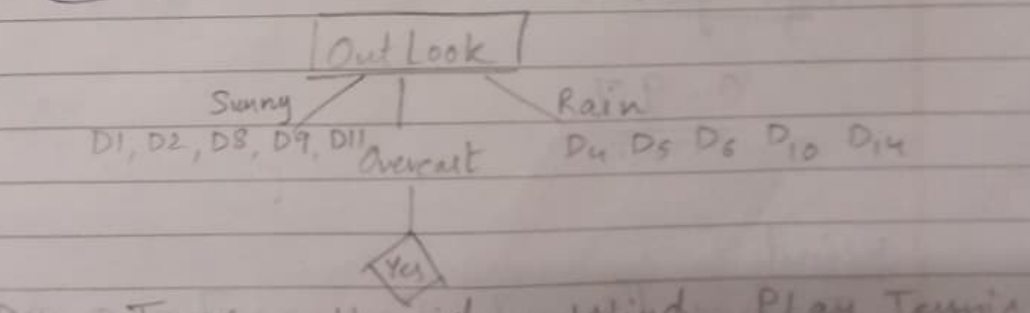
Entropy [Swk] = [6Y, 2N] = 0.8112

Entropy [Sstrong] = [3Y, 3N) = 1

Info gain :-

$$0.9402 - \left[\frac{8}{14} \times 0.8112 + \frac{6}{14} \times 1\right]$$

$$= \boxed{0.048}$$



Day	Temp	Humid	Wind	Play Tennis
D1	hot	High	Weak	No
D2	hot	high	Strong	No
D8	Mild	high	Wk	No
D9	cool	Normal	Wk	Yes
D11	mild	Normal	Strong	Yes

Entropy = 0.97

For Temp

Entrop [S_{hot}] = [0Y, 2No] = 0

Entropy [S_{mild}] = [1Y, 1No] = 1

Entropy [S_{cool}] = [1Y, 0No] = 0

Info gain

$$0.97 - \left[\frac{2}{5} \times 0 + \frac{2}{5} \times 1 + \frac{1}{5} \times 0\right]$$

0.57

Humid:

Entropy [S_{high}] = [3no] = 0

Entropy [S_{normal}] = [2Y] = 0

Info gain:

0.97

Wind:

Entropy [S_{weak}] = [1Y, 2N] = 0.918

Entropy [S_{strong}] = [1Y, 1No] = 1

Info gain: $0.97 - \left[\frac{3}{5} \times 0.918 + \frac{2}{5} \times 1\right]$

$= 0.0192$

For Temp

Entropy [Shot] = [0Y, 2No] = 0

Entropy [Smild] = [1Y, 1No] = 1

Entropy [Scool] = [1Y, 0No] = 0

Info gain

$$0.97 - \left[\frac{2}{5} \times 0 + \frac{2}{5} \times 1 + \frac{1}{5} \times 0\right]$$

0.57

Humid:

Entropy [Shigh] = [3no] = 0

Entropy [Snormal] = [2Y] = 0

Info gain:

0.97

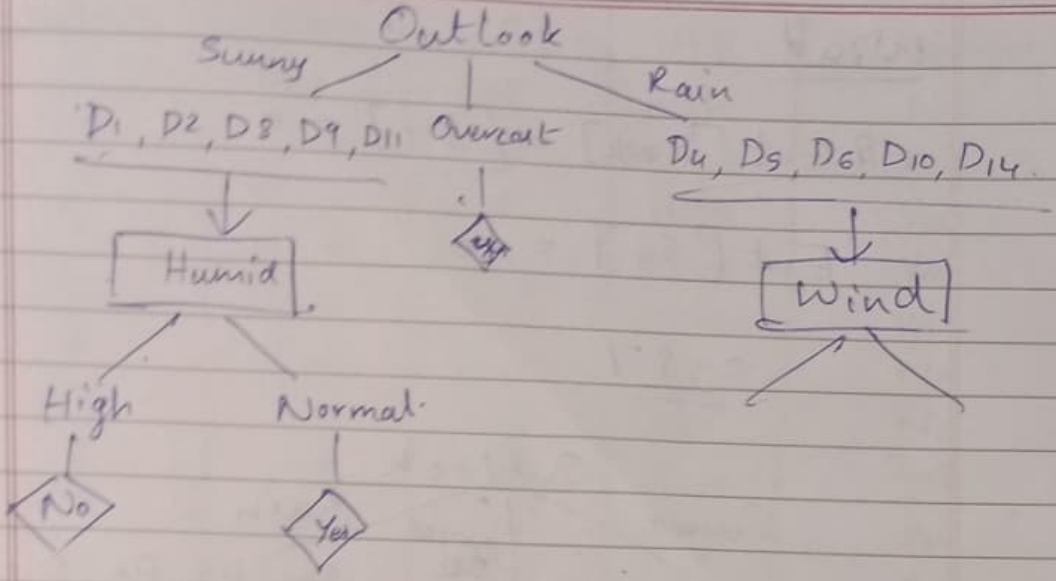
Wind:

Entropy [Sweak] = [1Y, 2N] = 0.918

Entropy [Sstrong] = [1Y, 1No] = 1

Info gain:- $0.97 - \left[\frac{3}{5} \times 0.918 + \frac{2}{5} \times 1\right]$

$= 0.0192$



Day	Temp	Wind	Play Tennis
D4	Mild	Wk	Y
D5	Cool	Wk	Y
D6	Cool	Strong	N
D10	Mild	Wk	Y
D14	Mild	Strong	N

∴ Entropy → 0.97.

Temp:

Entropy [Smild] = [2Y, 1N] = 0.918

Entrop [Scool] = [1Y, 1N] = 1

Info Gain:

$$0.97 - \left[\frac{3}{5} \times 0.918 + \frac{1}{5} \times 0.91^{0}\right]$$

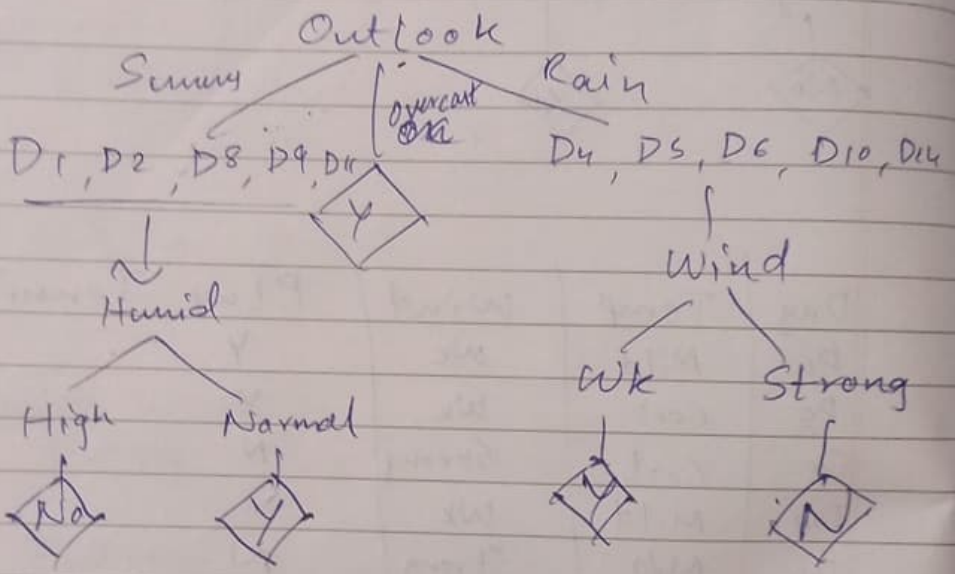
$$= 0.0192$$

Wind

Ent [S wk] = [3 Y] = 0

Ent [S s] = 0.

0.97



3)

Instances	A1	A2	A3	Classification
1	True	Hot	High	No
2	True	Hot	High	No
3	False	Hot	High	Yes
4	False	Cool	Norm	Yes
5	False	Cool	Norm	Yes
6	True	Cool	High	No
7	True	Hot	High	No
8	True	Hot	Norm	Yes
9	False	Cool	Norm	Yes
10	False	Cool	High	Yes

i) Give the entropy & info gain for the above dataset

ii) Construct Decision tree based on effectiveness of info gain value.

Ans: Entropy [S] = [6Y, 4No] = 0.97

For A1 :- True, False.

Entropy [True] = [1Y, 4No] = 0.72

Entropy [False] = [5Y] = 0

Info gain :-

$$= 0.97 - \left[\frac{5}{10} \times 0.72 + 0\right]$$

$$= 0.61$$

For A2 :-

Ent [Hot] = [2Y, 3N] = 0.97

Ent [cool] = [4Y, 1N] = 0.72

Info gain :

$$0.97 - \left[\frac{5}{10} \times 0.97 + \frac{5}{10} \times 0.72\right]$$

$$= 0.125$$

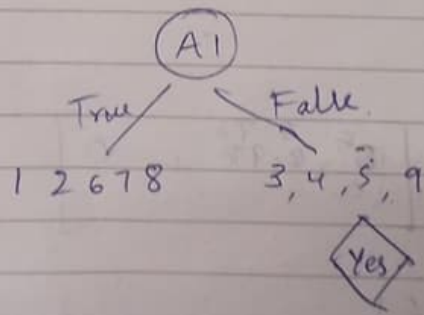
For A3

Ent [High] = [2Y, 4N] = 0.918

Ent [Norm] = [4Y] = 0.

Info gain :-

$$0.97 - \left[\frac{6}{10} \times 0.918\right] = 0.419.$$



Instance	A2	A3	Class
1	Hot	Hig	No
2	Hot	Hig	No
6	Cool	High	No
7	Hot	High	No
8	Hot	Norm.	Yes

Entropy[S] = [1Y, 4N] = 0.72.

For A2:

Ent[Hot] = [1Y, 3N] = 0.811

Ent[cool] = [1No] = 0.

Info gain:

$$0.72 - \left[\frac{4}{5} \times 0.811 + 0\right]$$

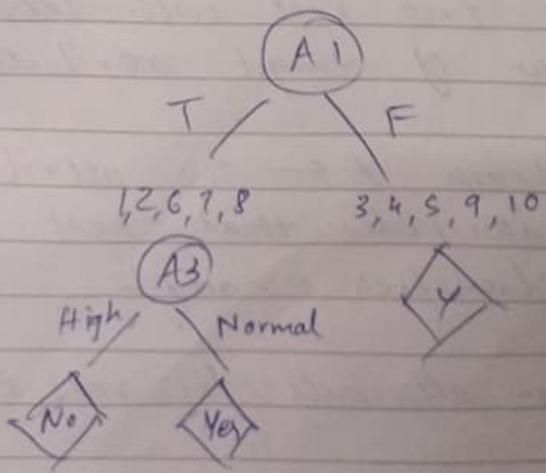
$$= 0.071$$

For A3:

Ent[High] = [4No] = 0

Ent[Nor] = 0.

Info = 0.72



Submission on Monday.

→ Assignment

1) Explain all 4 steps of Designing Learning System.
2) Explain any two application of machine learning.

Algorithm :-

imp

ID3(Examples, Target Attribute, Attributes)
Examples are the training examples
Target attribute is the attribute whose value is to be predicted by the tree and return a decision tree that correctly classifies as the given examples

i) Create a root node for the tree
ii) If all the examples are positive return the single node tree root, with lable=+
iii) If all the examples are negative return the single node tree root with label = '-' (Negative
iv) If the attribute is empty return the single node tree root with lable = most common value of target attributes in examples.
Otherwise begin A ← The attributes from the attributes that best multiplied with * classifies example.
v) The decision attribute for root ← A

vi) For the possible value ~~Vi(A)~~ Vi of A

Vi of A :

i) Add a new tree branch below the root corresponding to the test A = Vi

ii) Let examples Vi be the subset of examples that have value Vi of A.

If examples Vi is empty :-

i) Then below this new branch, add a leaf node with a label equals to most common value of target attribute in examples (ii) else below this new branch add the sub tree ID3(Examples Vi, Target_attribute, Attributes - {A})

→ END

→ return root

19/2/24

* How to select the value of K in KNN algorithm.

→ There is no particular way to determine the best value for K so we need to try some values to find the best out of them. The most preferred value for k is '5'. A very low value for 'k' such as K=1 or K=2 can be noisy & lead to the effects of outliers in the model.

→ Large values for 'k' are good but it may find some difficulties during training.

Advantages of KNN.

→ Simple to implement
→ It is robust to the noisy training data
→ It can be more effective for the training data. if it is large.

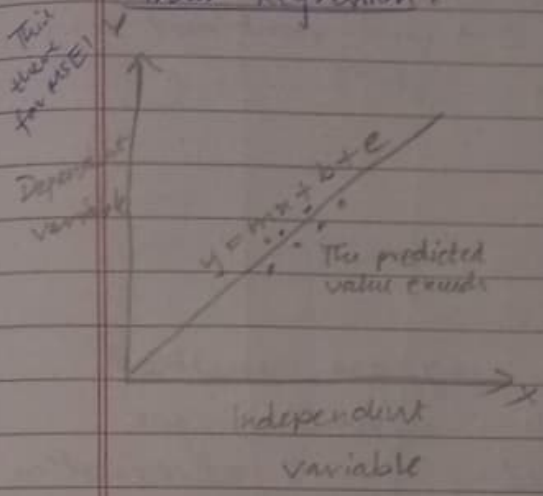
Disadvantages:

→ Always needs to determine value of k which may be complex sometime.
→ The computation cost is high because of calculating the distance btw the datapoints for all the training samples

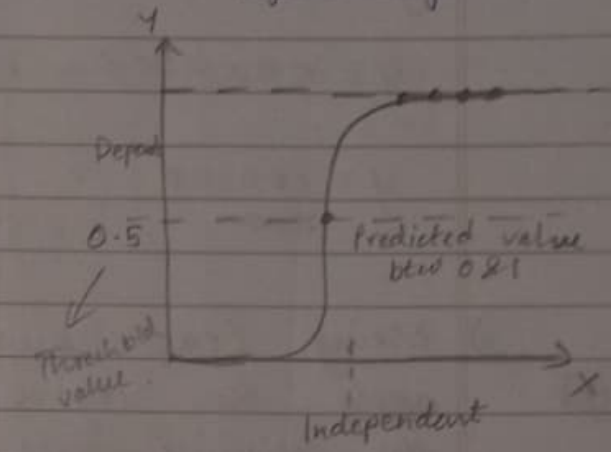
MSE1 Questions:

1) Explain Well proposed learning with an example.
2) Explain with a neat diagram concept of Machine Learning that summarizes concept learning.
3) Problems of KNN & Decision tree
4) Give the basic steps of ID3 algorithm
5) Explain ID3 algo in detail.
6) Explain ID3 algo with example (Draw tree)
7) Give algo for KNN.
8) Explain or list out the applications of ML
9) Give the difference between Supervised & unsup learning method.

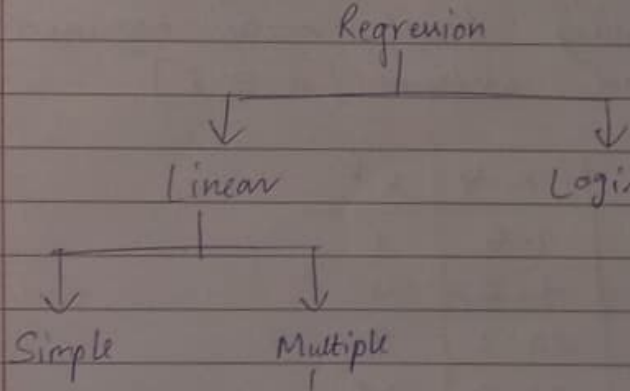
Linear Regression :



Logistic Regression



Regression :- Uses least mean square error method



MSE! Problem :-

1) Calculate the selling price (Y) of a house based on its size (X) and following are the model parameters :
Slope = 50, intercept = 100
Predict the sale price of a house with size 2000 sq ft.

UM

→

$y = mx + c + e$

$e \rightarrow$ is not mentioned hence zero.

$y = 50 \times 2000 + 100$

$y = 100100$

2) Using least mean square method apply the regression algorithm for the following data set of 7 observation on 2 variables X & Y & find the regression equation that relates X & Y variables or find the regression equation using least mean square error method where $[x \geq 2]$.

X	Y	X Y	X^2
1	1.5	1.5	1
2	3.8	7.6	4
3	6.7	20.1	9
4	9.0	36.0	16
5	11.2	56	25
6	13.6	81.6	36
7	16	112	49
$\Sigma x = 28$	$\Sigma Y = 61.8$	$\Sigma XY = 314.8$	$\Sigma x^2 = 140$

Step 1:- Take the combination of both X & Y variables and write sum of XY as well as the independent variable

$\Sigma XY = 1.5 + 7.6 + 20.1 + 36.0 + 56 + 81.6 + 112$

$\Sigma XY = 314.8$

$\Sigma x = 1+2+3+4+5+6+7 = 28$

$\Sigma y = 1.5+3.8+6.7+9+11.2+13.6+16 = 61.8$

$$m = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$m = \frac{7 \times 314.8 - 28 \times 61.8}{7 \times 140 - (28)^2}$$

$m = 2.414$

$$b = \frac{\Sigma y - m\Sigma x}{n}$$

$$b = \frac{61.8 - 2.414 \times 28}{7} \qquad b = -0.827$$

$y = mx + b + e$

x given in question ($x \geq 2$)

$y = 2.414 \times 2 + (-0.827)$

$y = 4.001$

if not given in question assume x btw (0–5)

3)

Observation	X	Y	XY	X^2
1	2.5	3.6	9	6.25
2	3.0	4.0	12	9
3	4.5	6.0	27	20.25
4	5.0	7.2	36	25
5	6.0	8.0	48.0	36
6	6.5	9.0	58.5	42.25
7	7.0	10.0	70.0	49
8	8.0	11.0	88.0	64
9	8.5	12.0	102	72.25
10	9.0	12.5	112.5	81
	$\Sigma x = 60$	$\Sigma Y = 83.3$	$\Sigma XY = 563$	$\Sigma x^2 = 405$

$$m = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{10 \times 563 - 60 \times 83.3}{10 \times 405 - (60)^2}$$

$$m = 1.404$$

$$b = \frac{\Sigma y - m\Sigma x}{n}$$

$$= \frac{83.3 - 1.404 \times 60}{10}$$

$$b = -0.094$$

$$y = mx + b \quad \text{let } x = 2$$

$$= 1.404 \times 2 + (-0.094)$$

$$y = 2.714$$