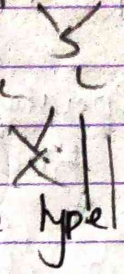
Unit 3

Clustering: It is technique of combining of datasets of similar type in groups.

Density based

- Clusters are formed at dense region.
- High accuracy
- Ability to merge two clusters
  Eg DBSCAN

Hierarchical based



Clusters are based formed as a tree type structure based on hierarchy

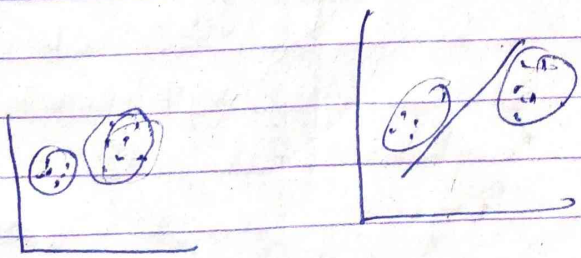
2 types → Bottom up
→ Top down

eg: CURE (Clustering using representatives)

Partitioning:

- Clusters are formed by partitioning into k clusters
- No of clusters = No of partitions.

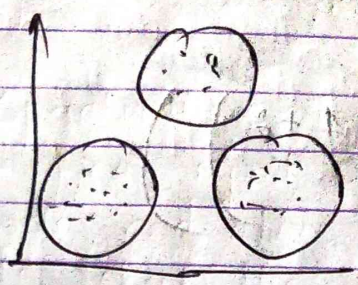
eg: k means



K means clustering is an unsupervised learning algorithm that is used to solve clustering problem in ML or data science.

* defins no of pre defined clusters, that need to be created in process.

Eg K = 2 2 clusters



DS		Sum
	x_i	y_i
C_1 =	1.0	1.0
C_2 =	5.0	7.0

Student	x Var 1.	y Var 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

K=2

Student	c1	c2
1	0	7.21
2	1.11	6.10
3	3.60	3.60
4	7.21	0
5	4.71	2.5
6	5.31	2.06
7	4.30	2.91

Now $c1 = \left(\frac{1}{2}(1+2.5), \frac{1}{2}(4+2)\right)$

$\rightarrow$ 1.25, 1.5 $\rightarrow$

new $c1 = (1.25, 1.5)$

Now $c2 = \left(\frac{1}{5}(3+5+3.5+4.5+3.5), \frac{1}{5}(4+7+5+5+4.5\right)$

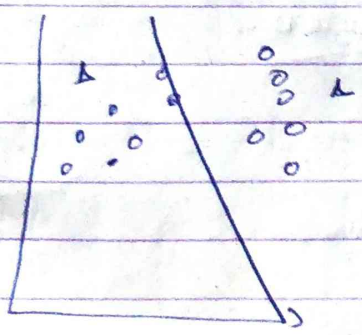
$\rightarrow (3.9, 5.2)$

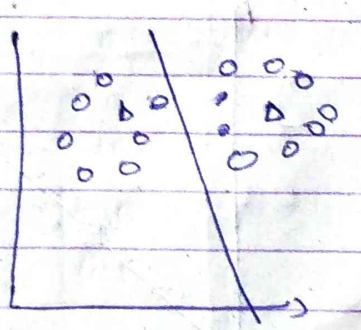
k=2

Student	C1	C2
1	0.55	5.02
2	0.55	3.92
3	3.05	1.42
4	6.65	2.19
5	4.16	0.41
6	4.77	0.60
7	3.75	0.72

Working

1. Select the number of k to divide no of clusters
2. Select random k point or centroids
3. Assign each dataset to their closest centroid
4. Apply Euclidean distance to determine grouping of clusters
5. Then calculate the new centroid value by taking avg of value
6. Repeat step 3 then 4 and it goes on till the desired k value
7. The model is ready.





Linear regression

- It is machine learning algo based on supervised learn
- Regression is mainly used to find the relationship between dependent & independent variables.
- Linear regression performs task to predict a dependent variable value (y) based on given independent variable (x),

Eg y → rapeseed x → Salary.
dependent.

(Single)

- $y = \alpha_0 + \alpha_1 x_i$

α_0 → ⁺ intercept.

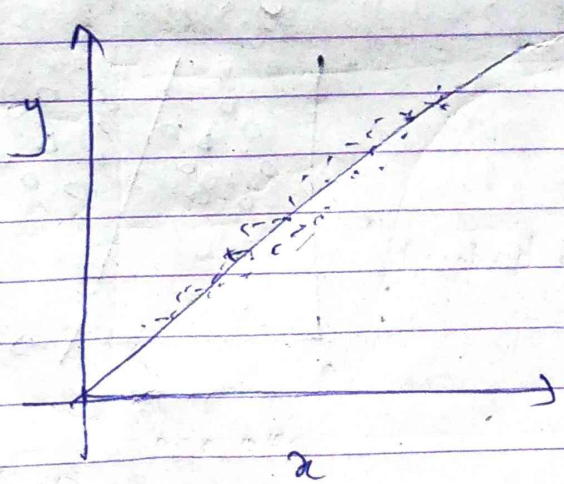
α_1 → coefficient of x.

(multiple)

$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots \alpha_n x_n$

α_0 → Intercept

α_i → coefficient of x_i



Higher value of α more weightage is given to data training.

Eg

x	y	xy	x^2	Predicted value	Error
1	3	3	1	2.8	0.2
2	4	8	4	4.1	0.1
3	5	15	9	5.4	0.4
4	7	28	16	6.7	0.3
$\Sigma x = 10$	$\Sigma y = 19$	$\Sigma xy = 54$	$\Sigma x^2 = 30$		

Find eqn $y = bx + a$

To find b and a.

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

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

$$a = \frac{(19)(30) - (10)(54)}{4(30) - (10)^2}$$

$$\Rightarrow \frac{570 - 540}{120 - 100} = \frac{30}{20} = 1.5$$

$$b = \frac{4(54) - (10)(19)}{4(30) - (10)^2}$$

$$\Rightarrow \frac{216 - 190}{120 - 100}$$

$$= \frac{26}{20} = \frac{13}{10} = 1.3$$

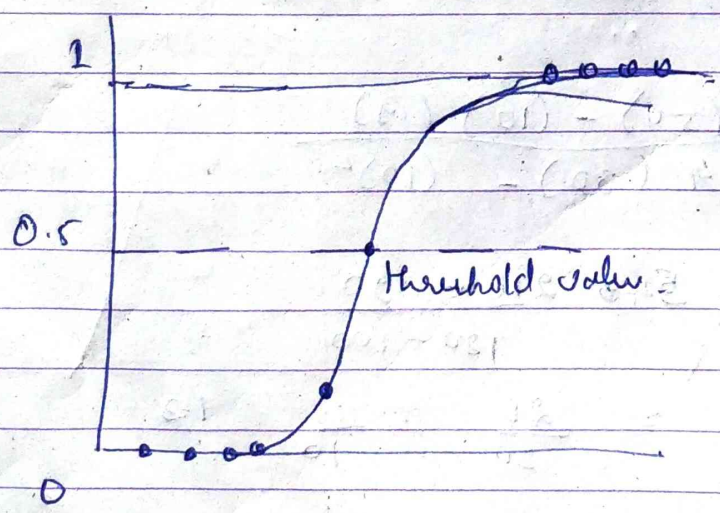
$$y = 1.3x + 1.5$$

What next ?

Sub given value of x & find ŷ
or find error in predicted & actual value

Logistic Regression

- It is a supervised algorithm.
- It is used for predicting the categorical dependent variable using given set of independent variables.
- As it just predicts the output of categorical dependent variable. Therefore the outcome must be categorical or discrete.
- Like it can be either 0 or 1, true or false etc.
- It basically gives probabilistic values which lie in b/w 0 & 1.
- Logistic regression is similar to linear regression except the type of data used.
- The graph or curve is fit in 'S' shape. It predicts 2 values (0 or 1)



Sigmoid function

It is function which helps us to map the values and get the outputs in b/w (0 or 1).

$$\Rightarrow y = \frac{1}{1+e^{-x}}$$

It is used for
- fraud detection
- disease diagnosis
- Spam no spam
- Emergency detection

Measures

- The dataset should be free of missing values
- Absence of multicollinearity
- Linearity in variables
- Irrelevant variable bias
- functional form
- Structural Break