⇒ PCA:- Principal component Analysis:-
it is used for to reduced the dimensionality.
it is used in the unsoperised learning Algortham.
→ LDA Linear discrimintal Analysis Algortham.
which is used in the supervised learning Algorithm to reduce the dimension.

→ Definations:- Principal component Analysis (PCA) is a dimensionality reduction technique used in unsupervised learning to transform a data set with many variables (features) into a smaller set that still contains most of the essential information.
ex:- we have a data set.
and if we have to calculate the price of the house. then what we see mostly?? like the size of the plot, or number of rooms, or number of washrooms or number of floors like that:~ we take the multiple features.
so if we have less features we can get the good accuracy but if we have many features then there complexity increase there, and at that time accuracy become decrease. and prediction will no come goode
sothats the reason we have to reduce the dimentions.

→ steps of principal component Analysis (PCA)

1 → Calculate Mean of every feature.

2 Calculate the covariance Matrix.

$$Cov = \begin{bmatrix} var(x_1) & cov(x_2, x_1) \\ cov(x_1, x_2) & var(x_2) \end{bmatrix}$$

example we have data set-

Size (in carpet) x_1	No of rooms x_2
5	10
3	7
10	15
2	4
20 → sum	36 → sum of number
4 → Total Num	4 → Total no of x_2

Mean(x_1) = 5

Mean of x_2 = 9

so

$\bar{x}_1 = 5$

$\bar{x}_2 = 9$

step 2-

Calculate Covariance Matrix

$$= \begin{bmatrix} 12.67 & 16.33 \\ 16.33 & 22 \end{bmatrix}$$

(a_{11}: 12.67, a_{12}: 16.33, a_{21}: 16.33, a_{22}: 22)

so how can we calculate the value of a_{11}?

we have to calculate variance.

$$a_{11} = \frac{(5-5)^2 + (3-5)^2 + (10-5)^2 + (2-5)^2}{N-1}$$

- 5: value of n_1, first value
- 5: mean of n_1
- 3: value of n_1, second value
- mean of n_1
- $N-1$: where N = total number of values in n_1, which is 4, then $n-1 = 4-1 = 3$

we will get $\boxed{12.67} \rightarrow a_{11}$

same way find the var(n_2) $\rightarrow a_{22}$.

you will get 22

Note => What is variance? variance is a number that tells you how spread your data is.

→ Imagine you have test scores from two different classrooms:- class A and B. class A's scores 71, 72, 73, 74, 75

class B's scores: 50, 60, 73, 86, 96

→ The average (Mean) Test score for both class is 73.

But the two sets are of score are very different!

→ In class A, every student scored very close to the average(mean) 73, The scores are bunched together.
• In class B, some students did very poorly and some very well, The score are all over the place, They are spread out.
→ variance is the number that captures this "spread".
→ class A would have a low variance because the scores are close together.
• class B would have a high variance becuse the score are far apart.

→ Now let's see how can we calculate q_{12} ?? cov (x_2, x_1) and cov (x_1, x_2)

→ so for that,

$$\frac{(x_1 - \bar{x}_1)\ (x_2 - \bar{x}_2)}{n-1}$$

$$\frac{(5-5)^2\ (10-9)^2 + (3-5)\ (7-9) + (10-5)\ (15-9)\ (2-5)\ (4-9)}{3}$$

we get

$q_{12} = 16.33$ → both are same

$q_{21} = 16.33$ → because it is in the multiplications.

⇒

Step 3: Calculate the Eigenvalues:

$$|S - \lambda I| = 0$$

S → covariance matrix.

$$\begin{vmatrix} 12.67 - \lambda & 16.33 \\ 16.33 & 22 - \lambda \end{vmatrix} = 0$$

and after finding the determinats of it we will get.

$$\lambda_1 = 32.3$$

$$\lambda_2 = -0.4$$

↳ we have two Eigen values because we worked on two features x_1 and x_2

→ so from the both values we have to consider the largest value. for calculate the next step which is Eigen vectors

which is largest value is $\lambda_1 = 32.3$.

step 4: Calculate the Eigenvectors

$$\begin{bmatrix} 12.67-\lambda & 16.33 \\ 16.33 & 22-\lambda \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} -1.3256 \\ 1 \end{bmatrix}$$

here U_1 and U_2 are the vectors (original data)

Step 5: Project Data onto principal components

$$\begin{bmatrix} 5 & 10 \\ n_1 & n_2 \end{bmatrix} \begin{bmatrix} -1.3256 \\ 1 \end{bmatrix} \begin{matrix} U_1 \\ U_2 \end{matrix}$$

1×2 rows, columns — 2×1

we will get 1×1 Matrix

$$[5 \times -1.3256 + 10 \times 1]$$

$$= 3.3 \rightarrow P.C$$

same for others

n_1	n_2	PC
5	10	3.3
3	7	3
10	15	18
2	4	1.3

So it means we have reduce the 2 dimensions into one, n_1 and n_2 to PC.

Now we can use just PC feature to predict the data.