PRINCIPAL COMPONENT ANALYSIS → Dimensionality Reduction.

* Curse of Dimensionality:

> CURSE OF DIMENSIONALITY:

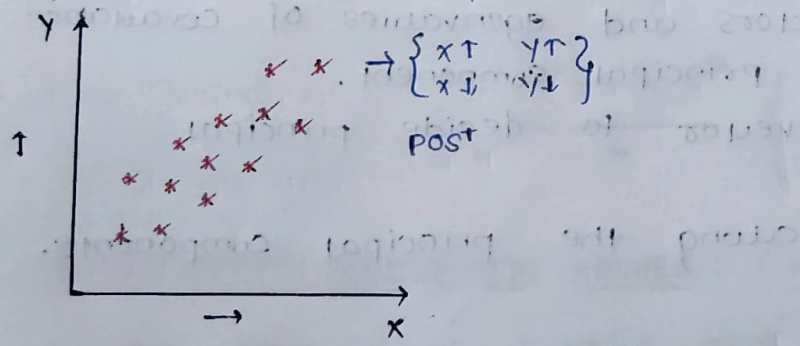
- Consider dataset with two feature x_1 and x_2.
- When two independent variable are very strongly interacting with each other, that is the correlation coefficient is close to 1 then we are providing the same information to algorithms in two dimension.
- which is nothing but redundancy. it can lead poor performance in model.
- In short when we have too many dimension more than required, which is nothing but called multicollineality in data.
- PCA also helps to reduce dimension without any lose huge information.

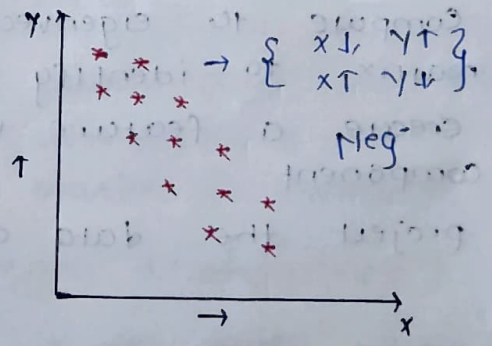
> TWO DIFFERENT WAY OF REMOVE CURSE OF DIMENSIONALITY

1.) FEATURE SELECTION

2.) PCA - FEATURE EXTRACTION

1.) FEATURE SELECTION





> $cov(x,y) = \frac{\sum_{i=1}^{N} (xi-\bar{x}) * (yi-\bar{y})}{N-1}$ = +ve or -ve → Any value

> pearson's corr = $\frac{cov(x,y)}{\sigma_x * \sigma_y}$. -1 to 1

> In pearso correlation the more value towards 1 more positive correlated or more toward -1 more negative correlated or value towards tw 0 no correlation.

2.) FEATURE EXTRACTION - PCA

terminology:

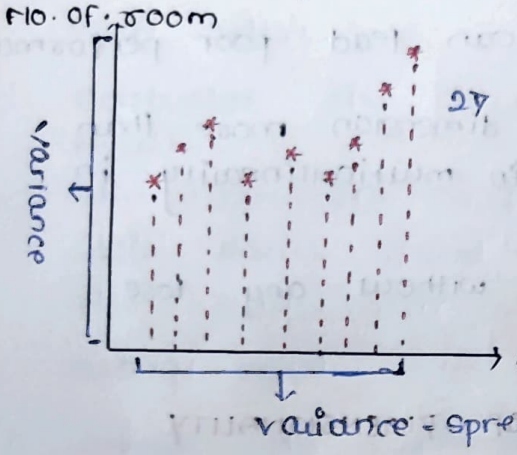
> variance: spread of datapoints from its means

> covariance: measure the relationship between x any y.

> GEOMETRIC INTUITION:

> consider having dataset with no. of rooms and size.

Aim: predict price based on no. of rooms and size.



> If you drop any one feature. it can make a huge loss of information.

> so we use PCA. in PCA we apply some transformation that give new feature with maximum variance.

> In PCA we reduce dimension 2D → 1D

STEPS:-

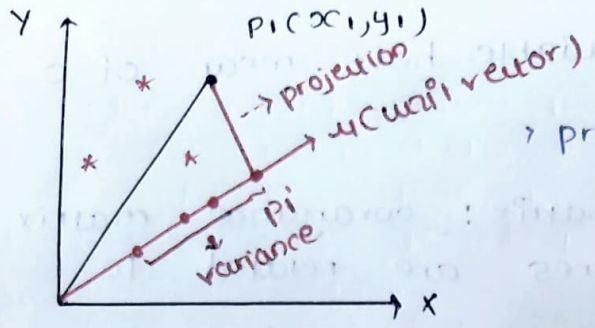
1.) Compute covariance matrix.

2.) Compute the eigenvectors and eigenvalues of covariance matrix to identify principal component.

3.) Create a feature vector to decide principal component

4.) project the data along the principal components.

* MATH INTUITION BEHIND PCA ALGORITHM *



› projection of p_1 on u

$$\text{proj}_{p_1}^{u} = \frac{p_1 \cdot u}{\|u\|}$$

$\|u\| = 1$ · unit vector

$$\boxed{\text{proj}_{p_1}^{u} = p_1 \cdot u}$$

› When we do all the projection we will get

$\{p_0', p_1', p_2', p_3', p_4', \ldots p_n'\}$ ⟶ scalar value

↳ Talking about distance from origin,

different notation

$\{x_0', x_1', x_2', x_3', x_4', \ldots x_n'\}$

Max variance. or Cost function $\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$

{ Aim: find the best unit vector which capture maximum variance

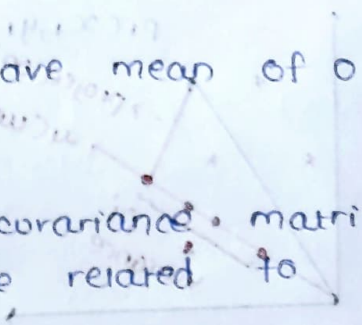
› **EIGEN VECTOR AND EIGEN VALUES** : (eigen decomposition)

Let say A is matrix and v is eigen vector and λ is eigenvalue

$\therefore \{Av = \lambda v\}$

* STEPS FOR PCA ALGORITHM *

1.) standardize the data : all variable have mean of 0 and standard deviation of 1.



2.) calculate the covariance matrix : covariance matrix gives you idea how features are related to each other.

3.) calculate the eigenvector and eigenvalue of covariance matrix.
- the eigenvector represent the direction in which the data varies most.
- the eigenvalue represent the amount of variation along each eigenvector.

4.) choose the principal component

5.) Transform the data into lower dimension.