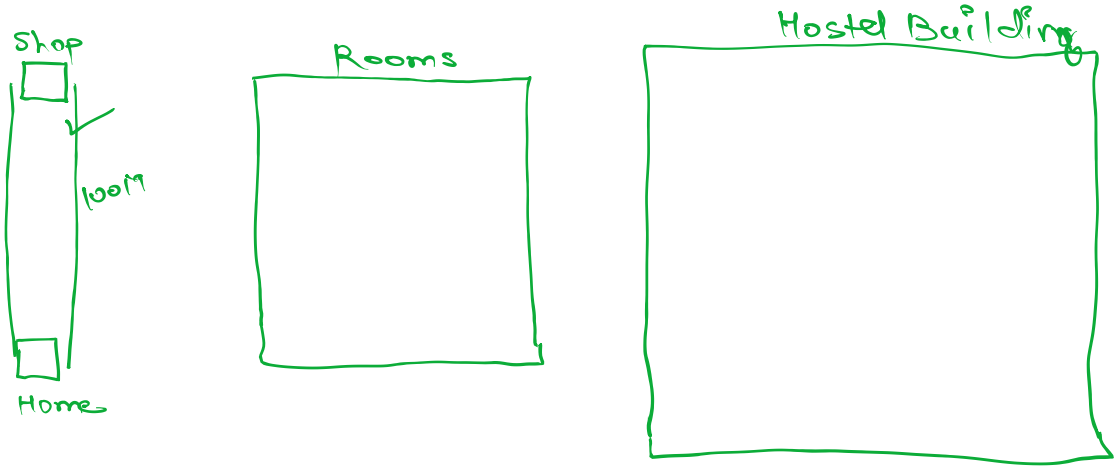
Principal Component Analysis(PCA) 2 Aug

Saturday, 2 August 2025 09:54

Curse of Dimensionality

No of cols = No of Dimensions



House Price Pred

10 cols

100 cols yes

→ Unnecessary cols → Overfitting

→ Multicollinearity → High Dimension

Solution ?-

Dimensionality reduction Techniques (DRT)

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)

Applied on all types of data

↓ only on Classification data

What are the option available to us for reducing dimensions →

(1) Feature Elimination / feature selection

→ Hyperparameter Tuning.

→ VIF (variance Inflation Factor) for dealing with multicollinearity.

→ RFE (Recursive feature Elimination).

→ Tree Based Algorithms like DT, RF, XGBoost have inbuilt feature selection techniques.

(2) Feature Extraction

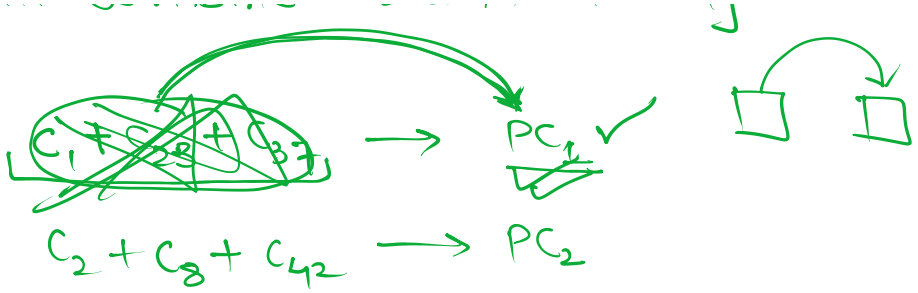
→ PCA

→ LDA

Principal Component Analysis (PCA)

$C_1, C_2 \ldots\ldots C_{100} \xrightarrow{PCA} PC_1\ PC_2\ PC_3 \ldots\ldots PC_n$

To create Principal components, PCA will combine columns linearly.



Mathematical working of PCA →

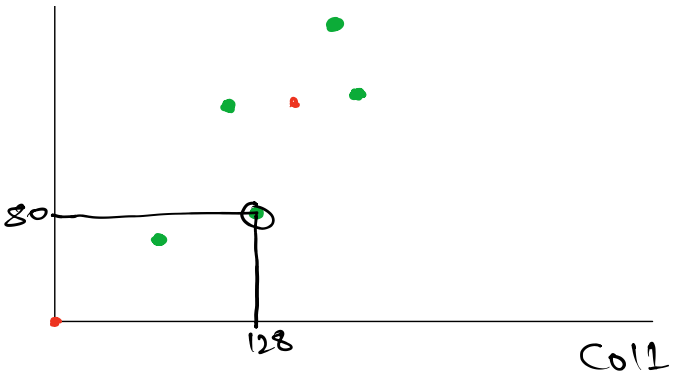
Steps in PCA →

1. Mean centering of the data.
2. Calculate the covariance matrix.
3. Calculate eigenvalues & eigenvectors of the covariance matrix.
4. Arrange the eigenvectors in descending order of eigenvalues.
5. Calculate the principal components.

Example calculation

Col 1	Col 2
126	78
128	80
128	82
130	82
130	84
132	86
→ (129	, 82)

Col 2



① $X - \text{mean}(X)$

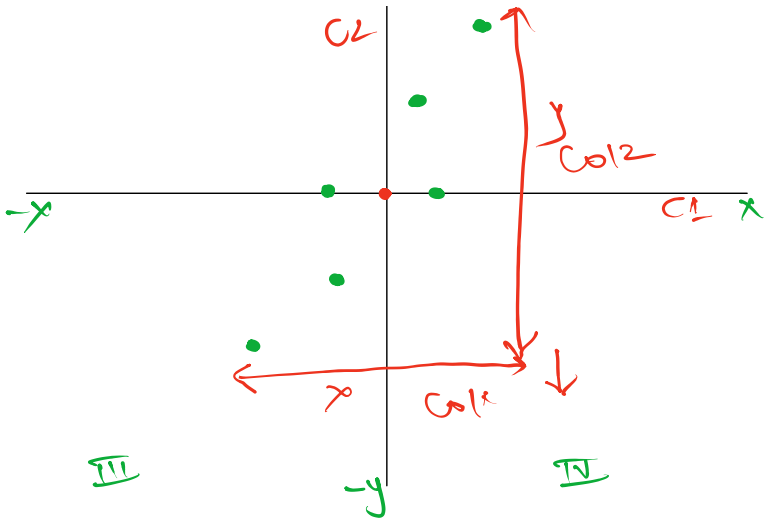
Data − mean

Col 1	Col 2
126 − 129	78 − 82
128 − 129	80 − 82
128 − 129	82 − 82
130 − 129	82 − 82
130 − 129	84 − 82
132 − 129	86 − 82

Var(c1)

Col1	Col2	
x_1 −3	−4	y_1
x_2 −1	−2	y_2
x_3 −1	0	y_3
x_4 1	0	y_4
x_5 1	2	y_5
x_6 3	4	y_6
(0 ,	0)	
$\bar{x}$	$\bar{y}$	

II I I



Dir
strength
cov ⇒ Dire

② Calculate the covariance matrix.

	Col 1	Col 2
Col 1	Var(C1)	Cov(C1, C2)
Col 2	Cov(C2, C1)	Var(C2)

$$Cov(C_1, C_2) = Cov(C_2, C_1)$$

Direction in which the variance of the data is more, the amount of information present in direction is more.

$$Var(C_1) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(-3-0)^2 + (-1-0)^2 + (-1-0)^2 + (1-0)^2 + (1-0)^2 + (3-0)^2}{6-1}$$

$$= 4.4$$

By similar calculation,

$Var(C_2) = 8.0$

Now,

$$Cov(C_1, C_2) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = (-3 \times -4) + (-1 \times -2) + (-1 \times 0) + (1 \times 0) + (1 \times 2) + (3 \times 4)$$

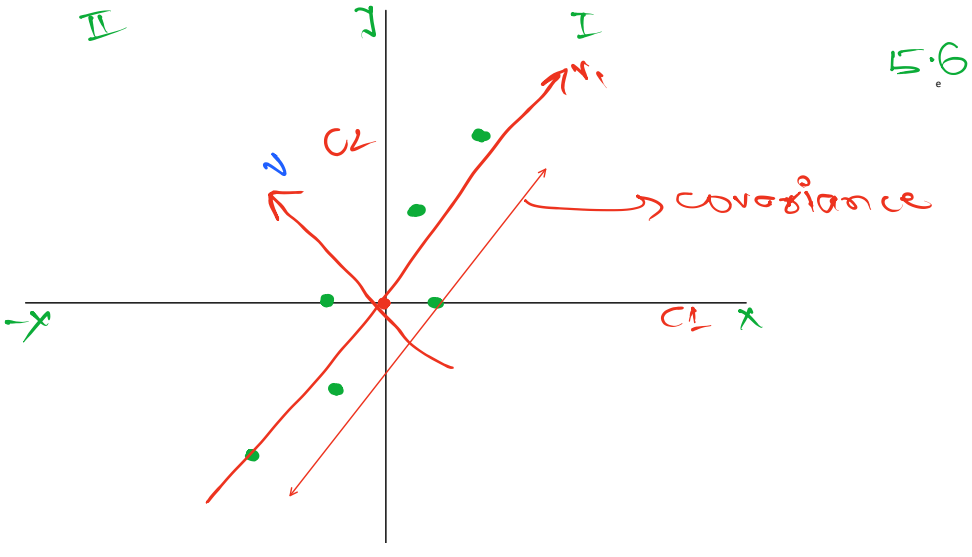
$$\frac{}{6-1}$$

$Cov(C_1, C_2) = 5.6$

$$\text{Cov Matrix} = \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix}$$

As we see from the covariance matrix, that variance of col2 is more than col1.

And the covariance value is somewhere between these two variances



III -y IV

(3) Now we will calculate the eigenvalue and eigen vectors of covariance matrix. And for eigenvector we use this formula → → Cov Matrix

$$\det|A - \lambda I| = 0$$

$$\begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix}$$

2

2,2

2,2

$$\det\left|\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right| = 0$$

$$\lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

Here the identity matrix should have some shape as covariance matrix.

$$\det\left|\begin{bmatrix} (4.4-\lambda) & 5.6 \\ 5.6 & (8.0-\lambda) \end{bmatrix}\right| = 0$$

$$(4.4-\lambda)(8.0-\lambda) - 5.6 \times 5.6 = 0$$

$$3.84 - 12.4\lambda + \lambda^2 = 0$$

$f(\lambda)$