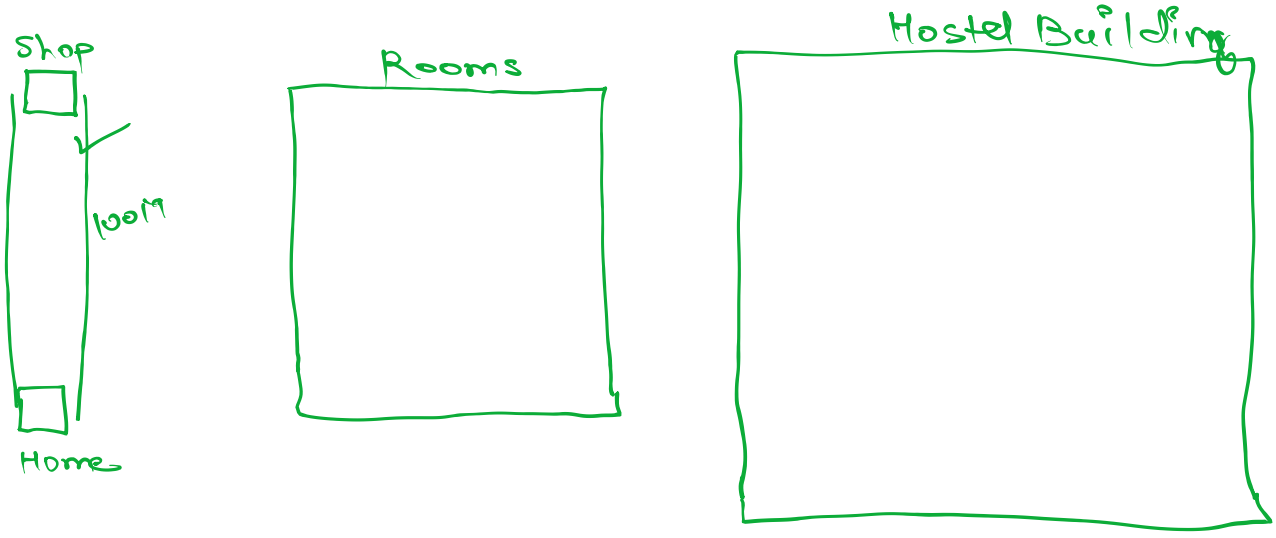
Principal Component Analysis(PCA) 2 Aug

Saturday, 2 August 2025 09:54

Curse of Dimensionality

No of cols = No of Dimensions



House Price Pred

10 cols

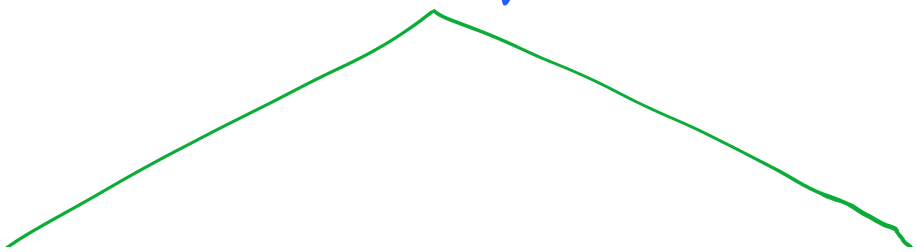
100 cols yes

→ Unnecessary cols → overfitting

→ Multicollinearity → High Dimension

Solution ?

Dimensionality reduction Techniques (DRT)



Principal Component Analysis (PCA)
↓
Applied on all types of data

Linear Discriminant Analysis (LDA)
↓
only on classification data

What are the option available to us for reducing dimensions →

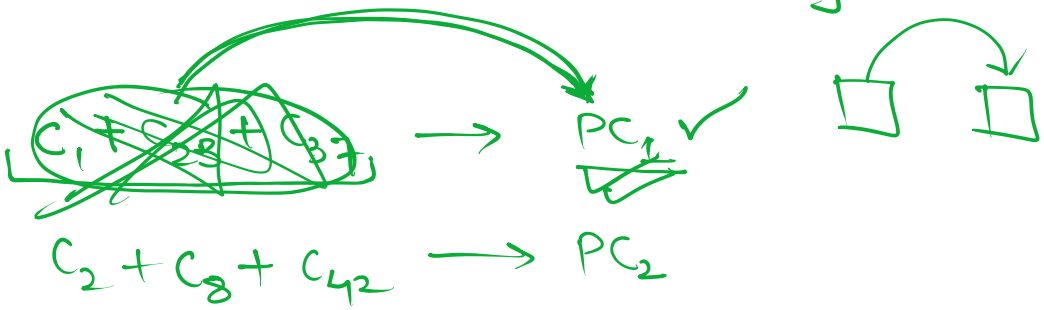
(1) Feature Elimination / feature selection
→ Hyperparameter Tuning.
→ VIF (variance Inflation Factor) for dealing with multicollinearity.
→ RFE (Recursive feature Elimination).
→ Tree Based Algorithms like DT, RF, XGBoost have inbuilt feature selection techniques.

(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.



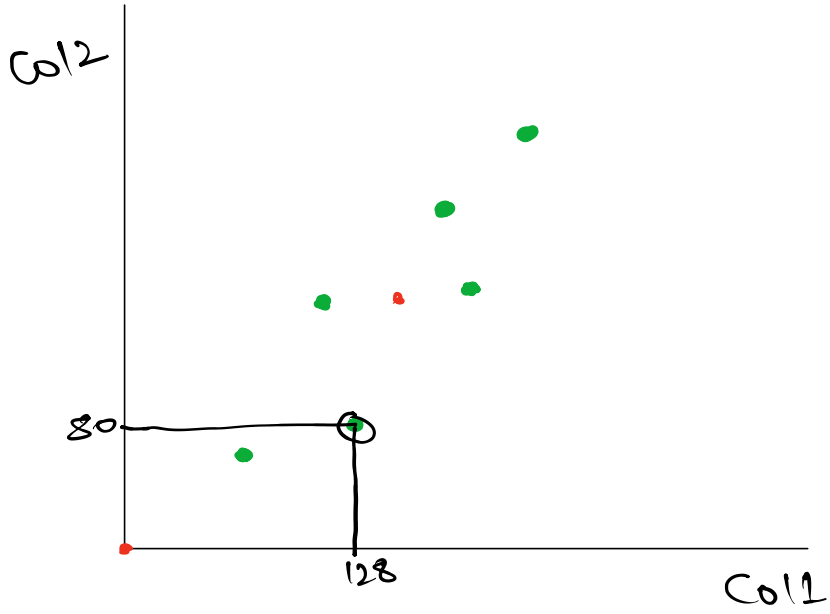
Mathematical working of PCA →

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)



2D

① X − Mean(X)

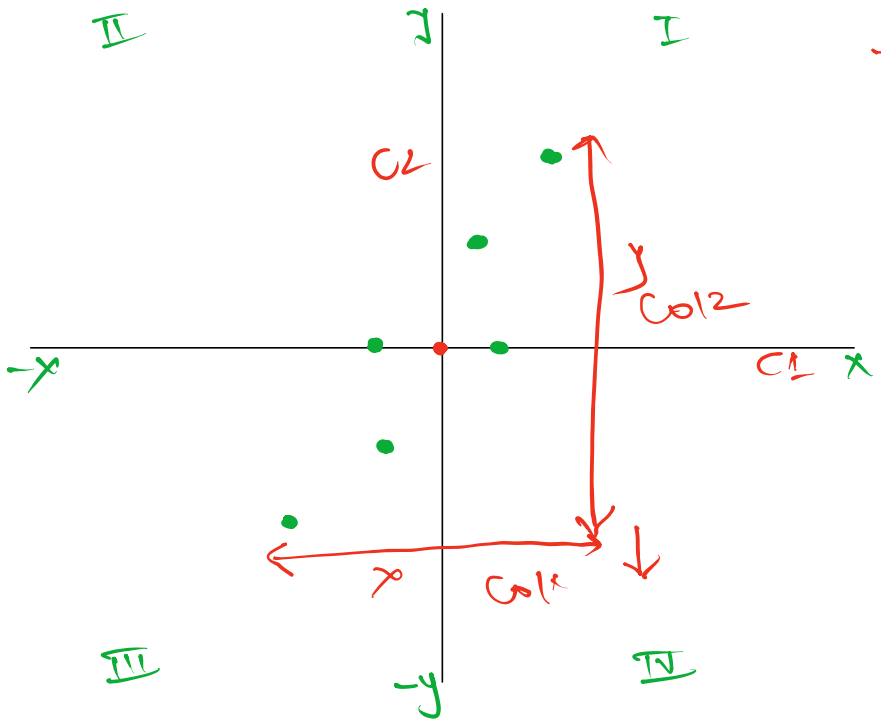
Data − Mean

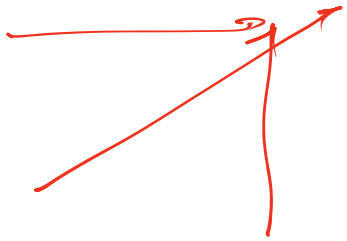
Col 1	Col 2
126 − 129	78 − 82
128 − 129	80 − 82
128 − 129	82 − 82

130 − 129	82 − 82
130 − 129	84 − 82
132 − 129	86 − 82

var(Col)

	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 ($\bar{x}$)	0 ($\bar{y}$)	





Dir
strength

Cov ⇒ Dir

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.

$$Var(C_1) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(-3-0)^2 + (-1-0)^2 + \cdots}{5}$$

$$= 4.4$$

$$Cov(C_1, C_2) = \frac{1}{n-1}\sum_{i=1}^{6}(x_i - \bar{x})(y_i - \bar{y}) =$$

	Col 1	Col 2
Col 1	4.4	5.6
Col 2	5.6	8.0

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