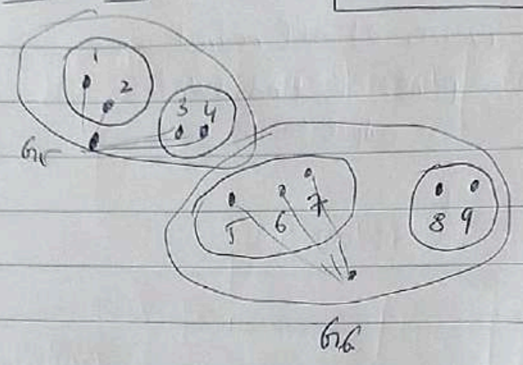
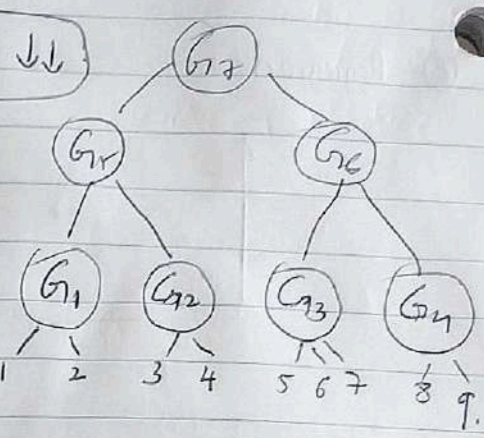
② Ball Tree

Time complexity ↓↓





PCA (Principal Component Analysis) [Dimensinality Reduction]

① Curse of Dimensinality — Dimensions

Dataset ≑ 500 features

Price of house
House size
No. of bedrooms
No of Bathrooms

3 features → m_1 ; 6 feature → m_2 ; 5 features → m_3 ; 50 features → m_4

Acc 1 ; Acc2 ↑↑ ; Acc3 ↑↑ ; Acc4 ↓

1000 features → m_5 ; 500 features → m_6

< Acc 5 ↓↓ < ; Acc 6 ↓↓↓

Model is over-fielded

② Model performance Degrade

Person ↓↓↓ Acc
↑ Confused

LocA ← 450k-500$
3Bhk ← 500k-600k
Beach ← ↑↑
Nex to celebrity ← ↑↑↑
School ← ↑↑

Two different ways to remove Curse of Dimensiality

① Feature Selection ⟶ Take Imp features

② Dimensinality Reduction (PCA) ⟹ Feature Extraction

<u>Feature Selection</u> vs <u>Feature Extraction</u>

↳ Dimensinality Reduction

① Why Dimensinaility Reduction)

$\boxed{3d}$ (or) $\boxed{2d}$

① Prevent → Curse of Dimensinality

② To improve the performance of the model

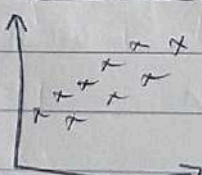
③ To visualize the data → understand the data

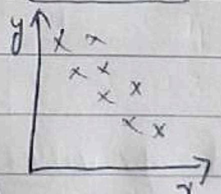
$\boxed{100d} \Downarrow \boxed{3d}$ (or) $\boxed{2d}$

<u>Feature selection</u>

x	y
–	–
–	–
–	–

+ve $\boxed{\begin{matrix} x\uparrow & y\uparrow \\ x\downarrow & y\downarrow \end{matrix}}$ $\boxed{\begin{matrix} x\downarrow & y\uparrow \\ X\uparrow & X\downarrow \end{matrix}}$ -ve





$$cov(x,y) = \frac{\sum_{i=1}^{n} (x_i-\bar{x}) * (y_i-\bar{y})}{N-1} \begin{matrix} = +ve \\ = -ve \end{matrix}$$

≈0 → no relationship b/w x & y.

$$\text{Pearson Correlation} = \frac{Cov(x,y)}{\sigma_x * \sigma_y} = \boxed{-1 \text{ to } 1}$$

The more towards value +1 the more corrected x & y

We are trying to extract a new feature from the old one which is already present

Feature extraction → Imp. features

2 features → 1 feature

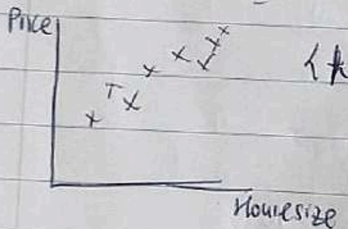
~~House~~ Room size	No. of rooms	Price

Transformation to extract new features

House size	Price
Independent	dependent O/P

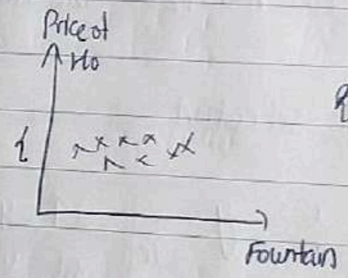
Feature selection

House size	Fountain size	Price
—	—	—
—	—	—



{Linear Relationship}

Cov = +ve
(or)
-ve



{ Not have more Relationship the changes so we can drop column from the table

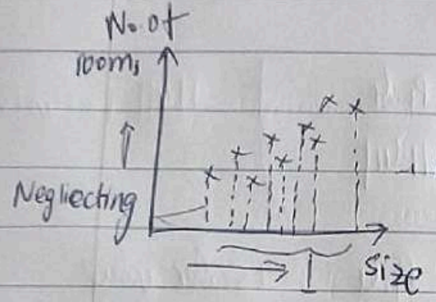
PCA Geometric Intuition {Dimensinality Reduction}

Housing Dataset

PCA

| 2dimension → 1dimension |

Size House	No. of Room	Price



· 2D → 1D

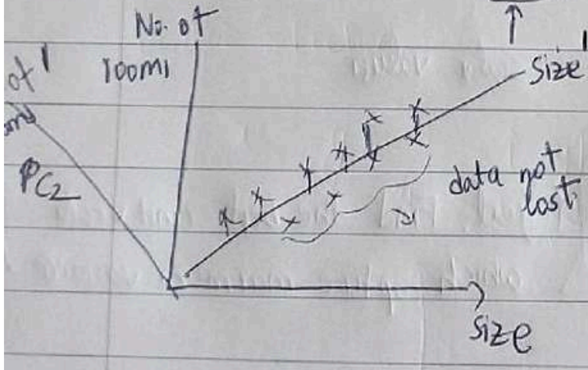
→ but no. of rooms of data was loost.

* loss of information (No of Rooms)

⇓

Feature extraction

$\boxed{PC1}$ → pirnciple component



| eigen decomposition on Matrix |

⇑

Transformation

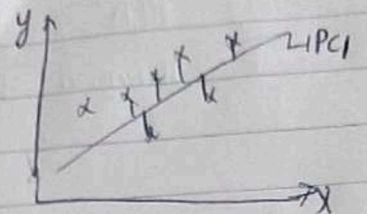
2D → 1D

much Information is not lost

3 Dimension	2 Dimension
PC_1 PC_2 PC_3	$PC1$ PC_2
$Var(PC_1) > Var(PC_2) > Var(PC_3)$	$Var(PC1) > Var(PC_2)$

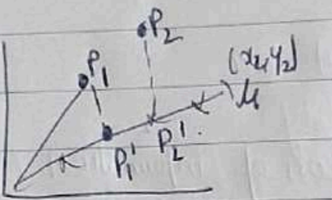
To get the best Pinciple component which captures maximum variance

Maths Intuition behind PCA Algorithm



① Projection

② Cost function → Variance



$$\boxed{Proj_{P_1} u = \frac{P_1 \cdot u}{\|u\|}}$$

$\|u\| = 1$ → unit vect

$$Proj_{P_1} u = P_1 \cdot u$$ → scalar value

$$P_0', P_1', P_2', P_3', P_4' \ldots P_n'$$

scalar value

$$x_0', x_2', x_3', x_4', \ldots x_n'$$

$$\underbrace{\boxed{\text{Varience} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}}_{\text{Cost function}}$$

{ Goal: Find the best unit vector which capture maximum variance }

Eigen vectors and eigen values

① Covariance matrix between features

② Eigen vector and eigen values will found out from this covarience matrix

③ Eigen vector → Eigen value → magnitud of eigen vector → capture the maximum variance