Principal Component Analysis

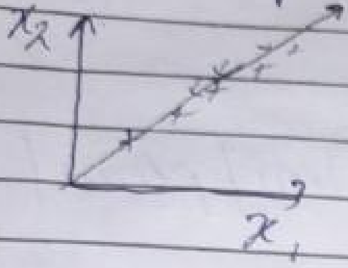
↳ For Feature Extraction

$x \in R^n \longrightarrow x \in R^k \quad k < n$

→ Application of PCA :-

① Data Compression — reduce no. of parameters.

example :-

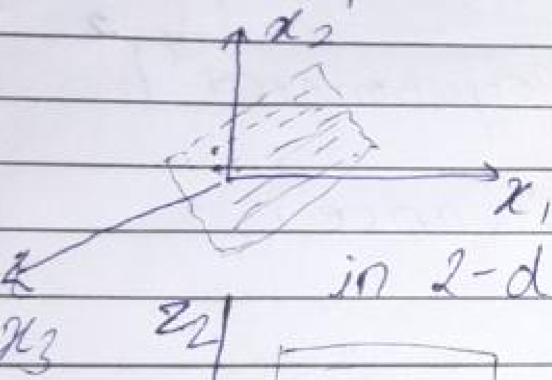


x_1	x_2

$x \in R^2 \longrightarrow x \in R$

① draw line that fits data best
② this line becomes our new x-axis
③ 2d now becomes 1 D

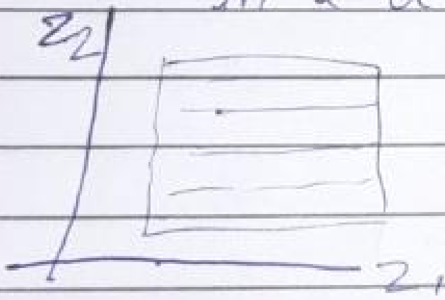
example :-



$x \in R^3$

3D → 2D plane
3D → 1D line

in 2-d



in 1-d we choose feature which has higher spread / variance
most of the variance should be retained

② Plotting
we cannot make plots for multiple features at once (scatterplots) hence it becomes much easier to convert it to 2-D and then plot.

③ Speeds up the Algorithm Computatio

PCA Objective

Objective:- $R^n \longrightarrow R^k$ $k < n$

steps:-

(1) Data Preprocessing - Standardization

$\mu = 0$ $\sigma = 1$

$x' = \frac{x - \bar{x}}{\sigma} = \frac{x - \mu}{\sigma}$

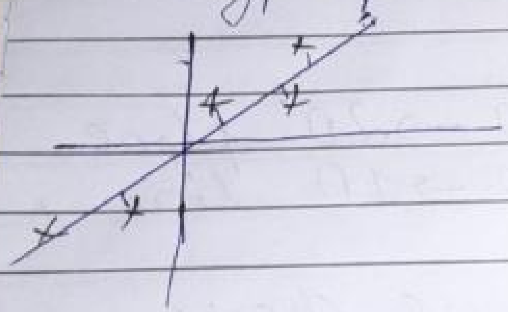
standardized

(2) Find

1D → line 2D - plane

3D → hyperplane

project you data on the line/plane/hyperplane



projection error

minimize $\sum_{i=1}^{m} (\text{projectionerror}^{i})^2$

maximize (variance)

PCA Algorithm

X data

training → (10,000 , 2)

(1) preprocessing (Standardization)

(2) compute Covariance Matrix

$$Var(X) = \sum_{i=1}^{m} \frac{(x^i - \bar{x})^2}{N-1}$$

Page No.
Date :

x_1	x_2

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

$$cov(x, y) \Rightarrow \sum_{i=1}^{m} \frac{(x-\bar{x})(y-\bar{y})}{N-1}$$

$$cov(x, x) = \sum_{i=1}^{m} \frac{(x-\bar{x})^2}{N-1}$$

(3) u^i → compute eigenvector of covariance matrix

svd – singular value decomposition

svd (covariance matrix) – u, s, v

(4) Projection

$U_{red} \rightarrow (n \times k)$ $\quad x \rightarrow (n \times 1)$

$z^i = U_{red}^T \cdot x^i$

$= k \times n \cdot n \times 1$ $\quad x^i \in R^n \quad k < n$

$z^i = k \times 1$ $\quad z^i \in R^k$