Chapter-04

Dimension Reduction Method

* Dimentionality Reduction:

→ It is a process of reducing no. of input variables or features in a dataset while retaining meaning information.

→ It is important step in data mining because in real-time application, dataset contains many attributes not all of which are useful or relevant

* Multi-colinearity:-

It occurs when two or more independent variables in a dataset are highly correlated with each other.

Ex- Predicting house price using Area, types of, locality. Among these area and types are highly correlated and process multicolinearity.

Problems:

(i) Unstable coefficient in regression Model.
(ii) Reduces Model interpretability.
(iii) Decrease Model performance.

Handle:

(i) Use PCA (principle component Analysis).
⇒ Other two methods are:-
(a) Ridge (b) Lasso.

2mark

* Need of Dimentionality Reduction:-

(i) To remove irrelevant and redundant features.
(ii) To improve model performance.
(iii) To reduce storage and memory requirement.
(iv) To improve visualization and interpretation.

* Principle Component Analysis:-

Steps:-

(1) Standarised the data using Z-score Normalization.

$$\text{Z-score} = \frac{x_i - \bar{x}_i}{S_i}$$

where $\bar{x}_i$, mean of x_i

S_i, standard deviation of x_i

(2) Compute co-variance matrix

(3) Compute eigen value & eigen vector.

(4) Compute PCA.

(Q) Suppose we have variables exam1, exam2 of students.

Student	Exam1	Exam2
A	90	85
B	70	65
C	80	78
D	65	60
E	95	92

Apply PCA to compute principle component.

Sol^n Step-1

$$\text{Mean of Exam1} = \frac{90+70+80+65+95}{5} = \cancel{800}\ 80$$

$$\text{Mean of Exam2} = \frac{85+65+78+60+92}{5} = 76$$

$$\text{S.D of Exam1} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}{n-1}}$$

$$= \sqrt{\frac{(90-80)^2+(70-80)^2+(80-80)^2+(65-80)^2+(95-80)^2}{4}}$$

$$= \sqrt{\frac{(10)^2+(-10)^2+(0)^2+(-15)^2+(15)^2}{4}}$$

$$= \sqrt{\frac{100+100+225+225}{4}} = \cancel{100}\ 12.748$$

$= \cancel{12.748}$

$$S.D \text{ of Exam2} = \sqrt{\frac{(85-76)^2+(65-76)^2+(78-76)^2+(60-76)^2+(92-76)^2}{4}}$$

$$= 13.398$$

$$Z\text{-score}_A \text{ (Exam1)} = \frac{90-80}{12.748} = \frac{10}{12.748} = 0.78.$$

$$Z\text{-score}_B \text{ (Exam1)} = \frac{70-80}{12.748} = \frac{-10}{12.748} = -0.78.$$

$$Z\text{-score}_C \text{ (Exam1)} = \frac{80-80}{12.748} = \frac{0}{12.748} = 0.$$

$$Z\text{-score}_D \text{ (Exam1)} = \frac{65-80}{12.748} = \frac{-15}{12.748} = -1.17.$$

$$Z\text{-score}_E \text{ (Exam1)} = \frac{95-80}{12.748} = \frac{15}{12.748} = 1.17.$$

$$Z\text{-score}_A \text{ (Exam2)} = \frac{85-76}{13.398} = 0.671$$

$$Z\text{-score}_B \text{ (Exam 2)} = \frac{65-76}{13.398} = -0.821$$

$$Z\text{-score}_C \text{ (Exam 2)} = \frac{78-76}{13.398} = 0.149$$

$$Z\text{-score}_D \text{ (Exam 2)} = \frac{60-76}{13.398} = -1.194$$

$$Z\text{-score}_E \text{ (Exam 2)} = \frac{92-76}{13.398} = 1.194$$

<u>Step-2</u> Compute covarriance matrix for standardize data.

<u>Covarriance matrix</u>:-
It is a square matrix that contain covarriance between each pair in the dataset.

$$\text{COV (Exam1, Exam2)} = \frac{\sum_{i=1}^{n} x_i^2 \text{ (exam1)}}{n-1}$$

$$= \frac{(0.784)^2+(-0.784)^2+(0)^2+(-1.177)^2+(1.177)^2}{4}$$

$$= \frac{0.614+0.614+0+1.385+1.385}{4} = 0.995 \approx 1.$$

$$\text{cov}(\text{Exam2}) = \frac{\sum_{i=1}^{n} x_i^2 \,(\text{exam2})}{n-1}$$

$$= \frac{(0.672)^2 + (-0.821)^2 + (0.149)^2 + (-1.194)^2 + (1.194)^2}{4}$$

$$\approx 1$$

$$\text{cov}(\text{Exam1}, \text{Exam2}) = \frac{\sum_{i=1}^{n} x_i(\text{exam1}) \cdot x_i(\text{exam1})}{n-1}$$

$$= \frac{(0.784)(0.671) + (-0.784)(-0.821) + (0)(0.149) + (-1.17)(-1.194) + (1.17)(1.194)}{4}$$

$$= \frac{0.526 + 0.643 + 0 + 1.396 + 1.396}{4} = 0.99$$

$$S = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix} \text{ --- } 1$$

<u>Step-3</u>

Eigen value $= |S - \lambda I| = 0.$

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

$$S - \lambda I = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 1-\lambda & 0.99 \\ 0.99 & 1-\lambda \end{bmatrix} \text{ --- } 2$$

$$\Rightarrow \lambda_1 = 1.995, \; \lambda_2 = 0.005.$$

For eigen vector, $(S-\lambda I)\vec{v} = 0$

$$\Rightarrow \lambda_1 = 1.995$$

Now,

$$\vec{v} = \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow \begin{bmatrix} 1-1.995 & 0.995 \\ 0.995 & 1-1.995 \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\cancel{\vec{v} = \begin{bmatrix} x \\ y \end{bmatrix}}$$

$$\Rightarrow \begin{bmatrix} -0.995 & 0.995 \\ 0.995 & -0.995 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} -0.995x + 0.995y \\ 0.995x - 0.995y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Now,

put $y = x$,

$$\vec{v} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ x \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Then,

$$\|\vec{v}\| = \sqrt{1^2 + 1^2} = \sqrt{2} = 1.414$$

$$\vec{v_1} = \frac{1}{1.414} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix}$$

$$\vec{v_2} = \begin{bmatrix} 0.707 \\ -0.707 \end{bmatrix}$$

Step-4

Select principle component.

$$\text{Varriance captured by PC1} = \frac{1.995}{\lambda_1 + \lambda_2} = \frac{1.995}{1.995 + 0.005}$$

$$= 0.9975 \times 100$$

$$= 99.75\%$$

$$PC2 = \frac{0.005}{2} = 0.25\%.$$

$\therefore$ PC1 captures all the structures in the dataset than PC2.

Step-5 Transform the data using PC.

$$\vec{V_1} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} \text{ PC}$$

$$PC1_A = 0.707 \times \text{Z-score}(\text{Exam1}) + 0.707 \times \text{Z-score}(\text{Exam2}).$$

$$= 0.707 \times 0.78 + 0.707 \times 0.671$$

$$= 0.551 + 0.474 = 1.025.$$

$PC1_B =$