Data Science

Q) PCA for

x	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

$n = 10$

x	y	$x_i - \bar{x}$ (A)	$y_i - \bar{y}$ (B)	AB	A^2	B^2
2.5	2.4	0.69	0.49	0.3381	0.4761	0.2401
0.5	0.7	-1.31	-1.21	1.5851	1.7161	1.4641
2.2	2.9	0.39	0.99	0.3861	0.1521	0.9801
1.9	2.2	0.09	0.29	0.0261	0.0081	0.0841
3.1	3.0	1.29	1.09	1.4061	1.6641	1.1881
2.3	2.7	0.49	0.79	0.3871	0.2401	0.6241
2.0	1.6	0.19	-0.31	-0.4589	0.0361	0.0961
1.0	1.1	-0.81	-0.81	0.6561	0.6561	0.6561
1.5	1.6	-0.31	-0.31	0.0961	0.0961	0.0961
1.1	0.9	-0.71	-1.01	0.7171	0.5041	1.0201
				5.539	5.549	5.449

$\bar{x} = 1.81 \quad \bar{y} = 1.91$

Covariance matrix $= \begin{bmatrix} Cov(x,x) & Cov(x,y) \\ Cov(y,x) & Cov(y,y) \end{bmatrix}$ $\because$ here $Cov(x,y) = Cov(y,x)$

where; $Cov(x,x) = \sum_{n=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}$; $Cov(x,y) = \sum_{n=1}^{n} \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

$\Rightarrow$ Covariance matrix (A) $= \dfrac{\begin{bmatrix} 5.549 & 5.539 \\ 5.539 & 6.449 \end{bmatrix}}{9} = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$

now, we need to find eigen values & eigen vectors for the above covariance matrix.

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

$$\Rightarrow \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} (0.6166-\lambda) & 0.6154 \\ 0.6154 & (0.7166-\lambda) \end{bmatrix} = 0 \quad \text{------ eq(1)}$$

By finding det for above matrix, we will get an equation, which can be solved to get the eigen values:-

$\Rightarrow (0.6166-\lambda)(0.7166-\lambda) - (0.6154)^2 = 0$

$\Rightarrow 0.441 - 0.7166\lambda - 0.6166\lambda + \lambda^2 - 0.378 = 0$

$\Rightarrow \lambda^2 - 1.3332\lambda + 0.063 = 0$

now, we will find roots for above equation :- $\lambda = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$

$\Rightarrow \lambda_1 = 1.283$ & $\lambda_2 = 0.049$

now putting λ_1 in eq--(1) :- for eigen vectors

$$\begin{bmatrix} (0.6166-1.283) & 0.6154 \\ 0.6154 & (0.7166-1.283) \end{bmatrix}$$

(or)

eigen vectors can be found by using :- $CV = \lambda V$

$$\Rightarrow \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 1.283 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$\Rightarrow 0.6166x_1 + 0.6154y_1 = 1.283x_1 \Rightarrow x_1 = 0.923y_1$

$\Rightarrow 0.6154x_1 + 0.7166y_1 = 1.283y_1$

$\cancel{\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 0.923 \\ 1 \end{bmatrix}}$

$$\therefore \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} \frac{x_1}{\sqrt{x_1^2+y_1^2}} \\ \frac{y_1}{\sqrt{x_1^2+y_1^2}} \end{bmatrix} = \begin{bmatrix} \frac{0.923}{\sqrt{1.851}} \\ \frac{1}{\sqrt{1.851}} \end{bmatrix} = \begin{bmatrix} \frac{0.923}{1.36} \\ \frac{1}{1.36} \end{bmatrix} = \begin{bmatrix} 0.678 \\ 0.735 \end{bmatrix}$$

for $\lambda = 0.049$:- $CV = \lambda v$

$$\begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0.049 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$\Rightarrow 0.6166x_1 + 0.6154y_1 = 0.049x_1 \Rightarrow x_1 = \frac{-0.6154}{0.5676} y_1$

$0.6154x_1 + 0.7166y_1 = 0.049y_1$

$\Rightarrow x_1 = -1.084\, y_1$

$$\Rightarrow \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} \frac{-1.084}{\sqrt{2.175}} \\ \frac{1}{\sqrt{2.175}} \end{bmatrix} = \begin{bmatrix} \frac{-1.084}{1.474} \\ \frac{1}{1.474} \end{bmatrix} = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$$

SVD :- $A_{m\times n} \rightarrow$ rank (r)

$\Sigma_{m\times n}$; $U_{m\times m}$; $V_{n\times n}$

$\Rightarrow A = U\Sigma V^T$

$\Rightarrow$ from given "A" find $A^T.A$ & find eigen value for that matrix and also find σ values [like $\sigma_1 = \sqrt{\lambda_1}$; $\sigma_2 = \sqrt{\lambda_2}$, etc...].

$\Rightarrow$ from eigen values find eigen vectors and place beside each other by normalizing them to get "v".

$\Rightarrow$ find Σ, as it is a diagnol matrix of values "σ" & make its rank equal to A i.e $Rank(\Sigma) = Rank(A)$.

$\Rightarrow$ now $u_1 = \frac{1}{\sigma_1} Av_1$ & $u_2 = \frac{1}{\sigma_2} Av_2 \Rightarrow U = [u_1, u_2]$

$$A = \underbrace{U}_{AA^T} \; \underbrace{\Sigma}_{\text{diagnol matrix}} \; \underbrace{V^T}_{A^TA}$$

U = normalized eigen vector matrix AA^T

V^T = normalized eigen vector matrix A^TA.

Statistics : is the study of the collection, analysis, interpretation, presentation and organization of data.

Eg:- mean of marks obtained by 50 students in a class.

-> Basics of statistics include the measure of central tendency (mean, median, mode) and the measure of dispersion (variance and standard deviation)

Types of Statistics :-

i) Descriptive statistics :-

These statistics are used to summarize and describe a data set. Histograms, pie charts, bar and scatter plots are common ways to summarize data and present it in table and graphs.

Here, the summarization is done using :-

- Measures of central tendency : mean, median, mode
- Measures of dispersion : range, variance, standard deviation.
- Measures of shape : skewness, kurtosis.

For instance, if we have a data set containing the weights of 20 people, to summarize and describe the data, we could use descriptive statistics such as calculating the mean weight, the range of weights, or the skewness of the weight distribution.

ii) Inferential statistics :-

These statistics are used to draw conclusions or make predictions about a population based on a sample of data.

Here, the prediction is done using :- (t-tests, ANOVA, chi-square test & linear regression)

- Hypothesis testing : testing whether a sample mean is significantly different from population mean
- Confidence intervals : estimating the range of values within which a population parameter (i.e. mean) likely to fall,
- Regression analysis : analyzing the relationship between variables and predicting values based on that relationship.

For example, we might use inferential statistics to test whether a new drug reduces cholesterol levels in a population. we would take a sample of people and randomly assign them to a treatment group or a control group, and then use statistical tests to determine wheather there is a significant difference in cholesterol levels between the two groups.

Probability :-

Probability is a mathematical concept that predicts how likely events are to occur. The probability values are expressed between 0 & 1. The definition of probability is the degree to which something is likely to occur.

There are mainly two types of probability distributions :

i) Discrete probability :-

These distributions are associated with random variables that can only take on a finite or countable number of values, such as the outcome of a coin toss or the number of defective items in a batch of products.

ii) Continuous probability :-

These distributions on other hand are associated with random variables that can take on any value within a range, such as the height or weight of individuals in a population.

-> Examples of common probability distribution include the normal distribution (also known as Gaussian distribution), the binomial distribution, the poisson distribution and the exponential distribution.

-> Probability distributions are widely used in fields such as statistics, physics, engineering, finance, and many other areas of science and industry to model and analyze a wide range of phenomena.

SVD (singular value decomposition) :-

$$A_{m\times n} = \underset{\text{rotation}}{U_{m\times m}} \cdot \underset{\text{stretching}}{\Sigma_{m\times n}} \cdot \underset{\text{rotation}}{V^T_{nn}}$$

here U & V are orthogonal i.e $U.U^T = I$, $V.V^T = I$

& Σ is a rectangular diagonal matrix with singular values

$$A^T.A = (V.\Sigma^T.U^T)(U.\Sigma.V^T)$$
$$= \underline{V.\Sigma^T.\Sigma.V^T}$$

rotation : $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \rightarrow$ unitary transformation

$$A.A^T = (U.\Sigma.V^T)(V.\Sigma^T.U^T)$$
$$= \underline{U.\Sigma.\Sigma^T.U^T}$$

stretching : $\begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$

eg:- $A = \begin{vmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{vmatrix}_{2\times 3}$

$A^T.A = V$

$A.A^T = U$

$A^T = \begin{vmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{vmatrix}_{3\times 2}$

$A.A^T = \begin{vmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{vmatrix} \begin{vmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{vmatrix} = \begin{vmatrix} 11 & 1 \\ 1 & 11 \end{vmatrix}$

$\Rightarrow |A - \lambda I| x = 0$

$\Rightarrow \begin{vmatrix} 11-\lambda & 1 \\ 1 & 11-\lambda \end{vmatrix} = 0$

$\Rightarrow (11-\lambda)^2 - 1^2 = 0 \Rightarrow (11-\lambda)^2 = 1$

$\Rightarrow \pm(11-\lambda) = 1$

$\Rightarrow \underline{\lambda = 10}$; $\underline{\lambda = 12}$

Put λ values in above equation.

$\begin{vmatrix} 11-10 & 1 \\ 1 & 11-10 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = 0 \Rightarrow \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} \Rightarrow \begin{matrix} x_1 + x_2 = 0 \\ x_1 + x_2 = 0 \end{matrix}$ $\quad x_1 = -x_2$

$\begin{vmatrix} 11-12 & 1 \\ 1 & 11-12 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = 0 \Rightarrow \begin{vmatrix} -1 & 1 \\ 1 & -1 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} \Rightarrow \begin{matrix} -x_1 + x_2 = 0 \\ x_1 - x_2 = 0 \end{matrix}$ $\quad \therefore x_1 = x_2$

we need to write in decending order so, $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$

now calculating orthogonal for above $\Rightarrow \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{matrix} \rightarrow \sqrt{2} \\ \rightarrow \sqrt{2} \end{matrix} \Rightarrow \begin{vmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{vmatrix}$

now for V : $A^T.A$.

$$\Rightarrow \begin{vmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{vmatrix} \begin{vmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{vmatrix} \Rightarrow \begin{vmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{vmatrix} \longrightarrow \lambda^3 - S_1\lambda^2 + S_2\lambda - S_3 = 0$$

$S_1 = 22$

$S_2 = 100 + (+4) + 16 = 120$

$S_3 = 10(4) - 0 + 2(-20) = 0$

$\Rightarrow \lambda^3 - 22\lambda^2 + 120\lambda = 0$

$\Rightarrow \lambda(\lambda^2 - 22\lambda + 120) = 0$

$\Rightarrow \lambda^2 - 12\lambda - 10\lambda + 120 = 0 \Rightarrow \lambda(\lambda - 12) - 10(\lambda - 12) = 0 \Rightarrow (\lambda - 10)(\lambda - 12)$.

$\therefore \lambda = 0;\ \lambda = 10;\ \lambda = 12 \Rightarrow \lambda_1 = 12;\ \lambda_2 = 10;\ \lambda_3 = 0$

by cramers rule :-

$\therefore$ for $\lambda = 12$:- $\begin{vmatrix} (10-12) & 0 & 2 \\ 0 & (10-12) & 4 \\ 2 & 4 & (2-12) \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \\ x_3 \end{vmatrix} = 0 \quad \dfrac{x_1}{4} = \dfrac{-x_2}{-8} = \dfrac{x_3}{4}$

$$\begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix} = \begin{vmatrix} 1 \\ 2 \\ 1 \end{vmatrix}$$

for $\lambda = 10$:- $\begin{vmatrix} 0 & 0 & 2 \\ 0 & 0 & 4 \\ 2 & 4 & -2 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \\ x_3 \end{vmatrix} = 0 \Rightarrow \dfrac{x_1}{-16} = \dfrac{-x_2}{-8} = \dfrac{x_3}{0} \Rightarrow \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix} = \begin{vmatrix} 2 \\ -1 \\ 0 \end{vmatrix}$

for $\lambda = 0$:- $\begin{vmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \\ x_3 \end{vmatrix} = 0 \Rightarrow \dfrac{x_1}{4} = \dfrac{-x_2}{-8} = \dfrac{x_3}{-20} \Rightarrow \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix} = \begin{vmatrix} 1 \\ 2 \\ -5 \end{vmatrix}$

$V = \begin{vmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 0 & -5 \end{vmatrix} \Rightarrow$ orthogonal ~~V~~ $\quad V^T = \begin{vmatrix} 1 & 2 & 1 \\ 2 & -1 & 0 \\ 1 & 2 & -5 \end{vmatrix}$

orthogonal $V^T = \begin{vmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{5}{\sqrt{30}} \end{vmatrix}$

here λ_1 & λ_2 are 12 & 10 as they are common in both AA^T; A^TA and should be in decending orde

Σ = diognol matrix with same dimensionality as A $\Rightarrow \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 \\ 0 & \sqrt{\lambda_2} & 0 \end{bmatrix}$

$\begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix}$

$$\therefore A = \begin{vmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{vmatrix} \begin{vmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{vmatrix} \begin{vmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{vmatrix}$$

1. Data preprocessing :-

It is the process of transforming raw data into a format that can be easily analyzed by ML algorithms. It involves several steps, including data cleaning, integration, transformation and reduction.

i) Data cleaning :-

- Process of identifing and removing (or) correcting errors, inconsistencies and the missing values from data.
- Techniques used for data cleaning includes imputation, deduplication, outlier removal and handling missing values.
- Example : filling in missing values with estimated values when there is missing data in the 'age' coloum of dataset, this is called imputation.

ii) Data integration :-

- Process of combining data from multiple sources and resolving inconsistencies or conflicts.
- Techniques used includes data fusion, data alignment and data merging.
- Example: merging of customer data based on common identifier, such as customer ID and then align the information.

iii) Data transformation :-

- Process of converting data into a format suitable for machine learning algorithms.
- Techniques used includes normalization, feature scaling, encoding & data discretization
- Example : Converting categorical variable like colour into numerical value by encoding enables the ML algorithm to use in it directly.

iv) Data Reduction :-

- Process of reducing the amount of data while retaining as much relevant info as possible.
- Techniques used includes feature selection, dimensionality reduction and data sampling.
- Example : Selecting a subset of features that are most relevant to the analysis to increase computation speed of the model by dimensionality reduction.

Student T-test :-

Student T-test is a statistical hypothesis test that is used to compare the mean of two independent group of data. The t-test calculates a t-value, which is a ratio of the difference between the two means to the standard error of the difference.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{\Delta}}} \quad \text{where} \quad s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\& \quad s = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$$

where $\bar{x}_1$ = mean of first set of values.

$\bar{x}_2$ = mean of second set of values.

s_1 = standard deviation of first set of values.

s_2 = " " second " " "

n_1 = Total no. of values in first set.

n_2 = " " second ".