Previous Year Important Questions

Q. 1 Explore PCA. Given data = {4, 8,13 ; 11,4,5,14} . Compute the principal component using PCA algorithm. Also use PCA to reduce dimension from 2 to 1. **AKTU(2024-25)**

Q.2 Explain the concept of Outliers. **AKTU(2024-25)**

Q. 3 A = {(10,0.2) , (20,0.4),(25,0.7),(30.0.9),(40,1),(50.0.4)} **AKTU(2024-25)**

B = {(10,0.4) , (20,0.1),(25,0.9),(30.0.2),(40,0.6),(50.0.6)}

Apply Union , Intersection , Complement , bold Union , and Bold intersection operations on above listed Fuzzy sets. **AKTU(2024-25)**

Q. 4 For the dataset {12 , 15, 20, 22, 25, 30} , Calculate mean & Standard deviation and comment on data spread. **AKTU(2024-25)**

Q.5 Explain the concept of generalization in neural networks . How does it relate to the trade – off between bias and variance , & what strategies can be employed to enhance generalization performance?

AKTU(2023-24)

GW GATEWAY CLASSES

GW GATEWAY CLASSES

Q.6 Elaborate on the mathematical foundations of Support Vector Machines. **AKTU(2023-24)**

Q.7 The following data represent study hours(X) and scores(Y) : **AKTU(2025-26)**

X : 2,4,6,8

Y : 50,55,65,70,78

(1) Find the linear Regression Equation.

(2) Predict Y when X = 7

Q. 8 A dataset has eigenvalues : 4.5,2.8,1.2,0.4,0.1 **AKTU(2025-26)**

(1) Calculate percentage variance of each component

(2) Determine the number of components required to retain 90% variance.

GATEWAY CLASSES

Q.9 Describe PCA algorithm. Compute the Principal component using PCA algorithm on given data as :

AKTU(2023-24)

CLASS 1 : X = 2,3,4 Y = 1,5,3

CLASS 2 : X = 5,6,7 Y = 6,7,8

SIMPLE LINEAR REGRESSION

AKTU(2025-26)

- Equation of Linear Regression : **$Y = mX + b$**
- Y represents the dependent variable
- X represents the independent variable
- m is the slope of the line (how much y changes for a unit change in x)
- b is the intercept (the value of Y when X is zero)

NUMERICAL 1 : LINEAR REGRESSION

Question : the following data shows the number of hours studied(X) and marks obtained(Y).

Hours Studied (X)	**2**	**4**	**6**	**8**
Marks Obtained (Y)	30	40	50	60

Using the Least Square Method, find :

1. Regression formula
2. Predict the marks if a student studies 12hrs.

Solution :

Hours Studied(X)	Marks obtained(Y)	Mean($\bar{x}$)	Mean($\bar{y}$)	Deviations (X) ($X - \bar{x}$)	Deviations(Y) ($Y-\bar{y}$)	Product of Deviation	Sum of Product of deviations	Square of deviations for X
2	30	5	45	-3	-15	45	100	9
4	40			-1	-5	5		1
6	50			1	5	5		1
8	60			3	15	45		9

m : sum of product of deviations / sum of square of deviation for X

$m = 100/20 = 5$

GW
GATEWAY CLASSES

b = Mean of Y – (m * Mean of X)

= 45 – (5 *5) = 20

(1) Regression Equation : Y = mX +b

Y = 5X + 20

(2) Predict marks for 12 hrs :

Given X = 12

Substitute in the Equation : Y = 5*12 + 20 = 80

GW
GATEWAY CLASSES

MULTIPLE LINEAR EQUATION

- It is an extension of Simple Linear Regression.
- It models the relationship between a dependent variable(Y) and two or more independent variables (X1, X2, X3, ……Xn).
- The relationship is expressed by the Equation :

$$Y = b_0x + b_1x + b_2x \ldots\ldots b_nx$$

NUMERICAL 2 : MULTIPLE LINEAR REGRESSION

Q : A Company wants to predict sales (Y) based on Advertising budget ($X_{1)}$ and number of salespersons(X_2).

Size	Bedrooms	Price (in crores)
1	3	2
2	4	5
3	7	9

$$b = ((X^T X)^{-1} X^T)Y$$

$$X = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 2 & 4 \\ 1 & 3 & 7 \end{bmatrix} \quad Y = \begin{bmatrix} 2 \\ 5 \\ 9 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 3 & 4 & 7 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 3 & 6 & 14 \\ 6 & 14 & 32 \\ 14 & 32 & 74 \end{bmatrix}$$

$$b = ((X^T X)^{-1} X^T)Y$$

$$((X^T X)^{-1} = \begin{bmatrix} 3 & 1 & -1 \\ 1 & 6.5 & -3 \\ -1 & -3 & 1.5 \end{bmatrix}$$

$$((X^T X)^{-1} X^T) = \begin{bmatrix} 1 & 1 & -1 \\ -1.5 & 2 & -0.5 \\ 0.5 & -1 & 0.5 \end{bmatrix}$$

$$((X^T X)^{-1} X^T)Y = \begin{bmatrix} -2 \\ 2.5 \\ 0.5 \end{bmatrix}$$

NUMERICAL 3 : BAYESIAN THEOREM

TOTAL PROBABILITY + BAYE'S THEOREM

Q. A disease affects 2% of the population.

Given : $P(D) = 0.02$, $P(+|D) = 0.95$, $P(+|\bar{D}) = 0.05$

Find the probability that a person actually has the disease if the test result is positive.

Solution :

Step 1: Calculate the probability of not having the disease

$$P(\bar{D}) = 1 - P(D) = 1 - 0.02 = 0.98$$

Step 2: Apply Total Probability Theorem

$$P(+) = P(+|D) \times P(D) + P(+|\bar{D}) \times P(\bar{D}) = (0.95 \times 0.02) + (0.05 \times 0.98) = 0.019 + 0.049 = 0.068$$

Step 3: Apply Bayes Theorem

$$P(D|+) = [P(+|D) \times P(D)] / P(+) = (0.95 \times 0.02) / 0.068 = 0.019 / 0.068 = 0.2794 = 27.94\%$$

GATEWAY CLASSES

GATEWAY CLASSES

Final Answer :

P(D|+) = 27.94%

Conditional Probability :

P(D|+) is called Conditional Probability because it represents the probability of a person having the disease given that the test result is already positive.

Here,

Condition :

Test Result is Positive (+)

Required Probability = Person has the Disease (D)

Therefore , P(D|+) = 27.94%

Interpretation :

If a person's test result is positive, there is a 27.94% probability that the person actually has the disease

NUMERICAL 4 : BASED ON BAYE'S THEOREM

A factory has two machines. Machine A produces 60% of the items . Machine B produces 40% of the items . The probability of producing a defective item is :

P(D|A) = 0.03

P(D|B) = 0.08

If a randomly selected item is defective, find the probability that it was produced by Machine B.

Solution :

Given,

P(A) = 0.60P(B) = 0.40

P(D|A) = 0.03

P(D|B) = 0.08

Using Bayes' Theorem , **P(B|D) = [P(D|B) × P(B)] / P(D)**

GATEWAY CLASSES

GATEWAY CLASSES

Where ,

$P(D) = P(D|A) \times P(A) + P(D|B) \times P(B) = (0.03 \times 0.60) + (0.08 \times 0.40) = 0.018 + 0.032 = 0.050$

Now ,

$P(B|D) = (0.08 \times 0.40) / 0.050 = 0.032 / 0.050 = 0.64$

Final Answer : $P(B|D) = 0.64$

Therefore, the probability that the defective item was produced by Machine B is 64%.

NUMERICAL 5 : CONDITIONAL PROBABILITY TABLE

Q. Three factories A, B, and C produce light bulbs.

Factory A produces 30% of the bulbs, Factory B produces 40% , and factory C produces 30% .

The probability that a bulb is defective is :

Factory A	**P(D|A)**	**0.02**
Factory B	P(D|B)	0.05
Factory C	P(D|C)	0.03

A randomly selected bulb is found to be defective . Find the probability that it was produced by Factory B.

Step 1: Write the given probabilities :

P(A) = 0.30P P(D|A) = 0.02

(B) = 0.40 P(D|B) = 0.05

P(C) = 0.30 P(D|C) = 0.03

Step 2: Construct the Conditional Probability Table (CPT).

Factory	P(Factory)	P(Defective | Factory)
A	0.30	0.02
B	0.40	0.05
c	0.30	0.03

Step 3: Apply Total Probability Theorem:

$P(D) = P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C) = (0.02 \times 0.30) + (0.05 \times 0.40) + (0.03 \times 0.30) = 0.006 + 0.020 + 0.009 = 0.035$

Step 4: Apply Bayes Theorem:

$P(B|D) = [P(D|B) \times P(B)] / P(D)$

$= (0.05 \times 0.40) / 0.035$

$= 0.020 / 0.035 = 0.5714 = 57.14\%$

Final Answer :

$P(B|D) = 57.14\%$

Conditional Probability :

$P(B|D)$ is called Conditional Probability because it represents the probability that the defective bulb was produced by Factory B, given that the selected bulb is already defective .

Here , Condition = Bulb is Defective (D)

Required Probability = Bulb is produced by Factory B interpretation :

If a randomly selected bulb is defective, then there is a 57.14% probability that it was manufactured by Factory B.

NUMERICAL 5 : CONDITIONAL PROBABILITY TABLE

Q. Suppose the probability of rain on a day is 0.30.The probability that the grass is wet if it rains is 0.90.The probability that the grass is wet even if it does not rain is 0.20.

Find the probability that it actually rained given that the grass is wet.

Given:

$P(R) = 0.30$

$P(W|R) = 0.90$

$P(W|\bar{R}) = 0.20$

Solution :

Step 1: Calculate the probability of no rain :

$P(\bar{R}) = 1 - P(R) = 1 - 0.30 = 0.70$

GW GATEWAY CLASSES

Step 2 : Apply Total Probability Theorem :

$P(W) = P(W|R) \times P(R) + P(W|\bar{R}) \times P(\bar{R}) = (0.90 \times 0.30) + (0.20 \times 0.70) = 0.27 + 0.14 = 0.41$

Step 3: Apply Bayes Theorem :

$P(R|W) = [P(W|R) \times P(R)] / P(W) = (0.90 \times 0.30) / 0.41 = 0.27 / 0.41 = 0.6585 = 65.85\%$

Final Answer :

$P(R|W) = 65.85\%$

Conditional Probability :

P(R|W) is called Conditional Probability because it represents the probability that it rained given that the grass is already wet . Here, Condition = Grass is Wet (W)Required Probability = Rain occurred (R)

Interpretation : If the grass is wet, there is a 65.85% probability that it rained.

NUMERICAL 6 : FUZZY LOGIC : MEMBERSHIP FUNCTION

Q. Consider the fuzzy set "Hot Temperature" defined as : At 20°C, μ = 0 , At 40°C μ = 1Find the membership value of temperature 30°C using the Membership Function.

Solution :

Formula (Triangular/Linear Membership Function) = μ(x) = x-a/b − a

Where ,

a = 20°C

b = 40°C

x = 30°C

μ(x) = Membership value of x

x = Input value (temperature whose membership is to be found)

a = Lower limit (where membership value is 0)

b = Upper limit (where membership value is 1)

Substitute the values : μ(30) = (30 − 20) / (40 − 20)= 10 / 20= 0.5

Final Answer:

μ(30) = 0.5

Interpretation :

The membership value of 30°C in the fuzzy set "Hot Temperature" is 0.5.This means that 30°C is 50% Hot and 50% Not Hot .

Hence, the degree of membership of 30°C is 0.5.

NUMERICAL 7 : FUZZY LOGIC : FUZZIFICATION

Q. The fuzzy set "Low Temperature" is defined as follows :

At 20°C, the membership value is 1.At 40°C, the membership value is 0.

Find the fuzzy membership value of temperature 28°C using the Fuzzification process.

Solution :

Step 1: Identify the given values.

a = 20°C

b = 40°C

x = 28°C

Since the membership value decreases linearly from 1 to 0, use the decreasing membership function.

Formula : $\mu(x) = (b - x) / (b - a)$

GW
GATEWAY CLASSES
GW
GATEWAY CLASSES

Step 2: Substitute the values

$\mu(28) = (40 - 28) / (40 - 20) = 12 / 20 = 0.6$

Final Answer : **$\mu(28) = 0.6$**

Fuzzification : The crisp input value (28°C) is converted into a fuzzy value of 0.6.

Interpretation : The temperature 28°C belongs to the fuzzy set "Low Temperature" with a membership value of 0.6.This means that 28°C is 60% Low Temperature and 40% Not Low Temperature

REMEMBER POINTS

INCREASING MEMBERSHIP FUNCTION

When the membership value increases from 0 to 1.

μ(x) = Membership value

a = Lower limit

b = Upper limit

Formula : μ(x) = x-a/b − a

DECREASING MEMBERSHIP FUNCTION

When the membership value decreases from 1 to 0.

μ(x) = Membership value

a = Lower limit

b = Upper limit

Formula : μ (x) = (b − x) / (b − a)

FUZZY DECISION TREE

A Fuzzy Decision Tree (FDT) is a decision tree that combines Decision Tree and Fuzzy Logic. It uses fuzzy rules and membership values (0 to 1) to make decisions instead of only Yes/No conditions.

Example

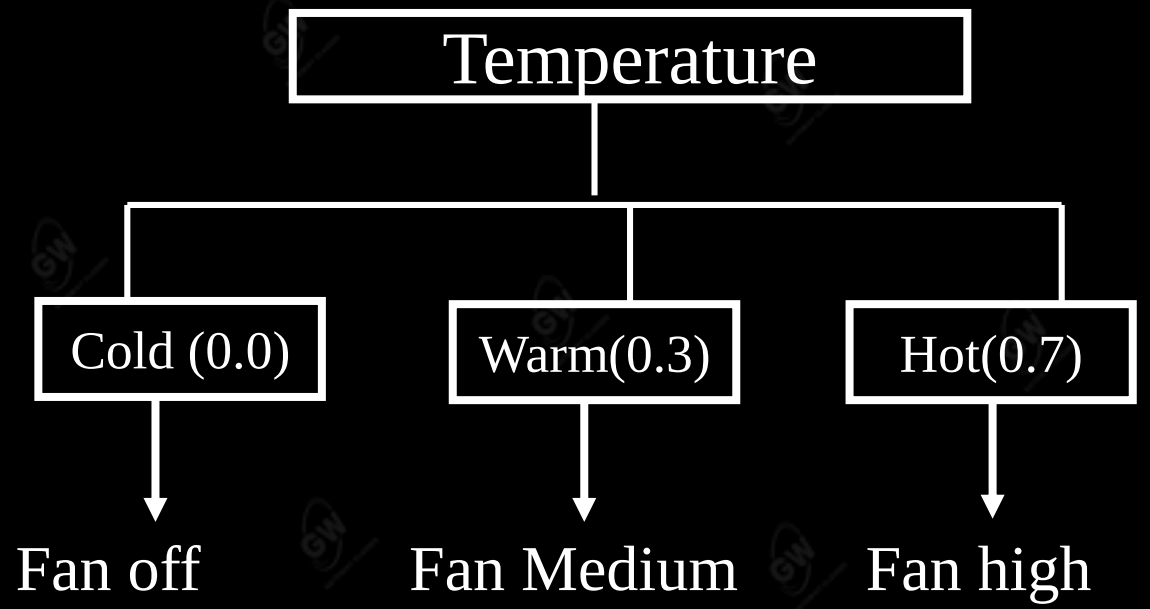
A smart fan decides its speed based on Temperature.

Fuzzy Sets: Cold , Warm, Hot

Suppose : Temperature = 35°C

Membership values : Cold = 0.0 , Warm = 0.3 , Hot = 0.7

This means 35°C is mostly Hot but slightly Warm.



Apply Union, Intersection, Complement, Bold Union and Bold Intersection Operations on the Above Listed Fuzzy Sets

AKTU(2024-25)

Given Fuzzy Sets:

$A = \{ (10, 0.2), (20, 0.4), (25, 0.7), (30, 0.9), (40, 1.0), (50, 0.4) \}$

$B = \{ (10, 0.4), (20, 0.1), (25, 0.9), (30, 0.2), (40, 0.6), (50, 0.6) \}$

Membership Values of A and B

$x \in U$	10	20	25	30	40	50
$A(x)$	0.2	0.4	0.7	0.9	1.0	0.4
$B(x)$	0.4	0.1	0.9	0.2	0.6	0.6

1. Union ($A \cup B$)

Formula: $\mu_{A \cup B}(x) = max\,(\mu_A(x), \mu_B(x))$

$x \in U$	10	20	25	30	40	50
$\mu_{A \cup B}(x)$	0.4	0.4	0.9	0.9	1.0	0.6

$$A \cup B = \{ (10, 0.4), (20, 0.4), (25, 0.9), (30, 0.9), (40, 1.0), (50, 0.6) \}$$

GATEWAY CLASSES

2. Intersection $(A \cap B)$

Formula: $\mu_{A\cap B}(x) = min(\mu_A(x), \mu_B(x))$

$x \in U$	10	20	25	30	40	50
$\mu_{A\cap B}(x)$	0.2	0.1	0.7	0.2	0.6	0.4

$$A \cap B = \{ (10,0.2), (20,0.1), (25,0.7), (30,0.2), (40,0.6), (50,0.4) \}$$

3. Complement of A $(\bar{A})$

Formula: $\mu_{\bar{A}}(x) = 1 - \mu_A(x)$

$x \in U$	10	20	25	30	40	50
$\mu_{\bar{A}}(x)$	0.8	0.6	0.3	0.1	0.0	0.6

$$\bar{A} = \{ (10,0.8), (20,0.6), (25,0.3), (30,0.1), (40,0.0), (50,0.6) \}$$

4. Bold Union $(A \cup_b B)$

Formula: $\mu_{A\cup_b B}(x) = \mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x)$

$x \in U$	10	20	25	30	40	50
$\mu_{A\cup_b B}(x)$	0.52	0.46	0.97	0.92	1.00	0.76

$$A \cup_b B = \{ (10,0.52), (20,0.46), (25,0.97), (30,0.92), (40,1.00), (50,0.76) \}$$

GATEWAY CLASSES

Universe of Discourse (U):

$U = \{10, 20, 25, 30, 40, 50\}$

Fuzzy Operations Used

Union $(A \cup B)$: $\mu_{A\cup B}(x) = max(\mu_A(x), \mu_B(x))$

Intersection $(A \cap B)$: $\mu_{A\cap B}(x) = min(\mu_A(x), \mu_B(x))$

Complement $(\bar{A})$: $\mu_{\bar{A}}(x) = 1 - \mu_A(x)$

Bold Union $(A \cup_b B)$: $\mu_{A\cup_b B}(x) = \mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x)$

Bold Intersection $(A \cap_b B)$: $\mu_{A\cap_b B}(x) = \mu_A(x) \cdot \mu_B(x)$

---

5. Bold Intersection $(A \cap_b B)$

$x \in U$	10	20	25	30	40	50
$\mu_{A\cap_b B}(x)$	0.08	0.04	0.63	0.18	0.60	0.24

$$A \cap_b B = \{ (10,0.08), (20,0.04), (25,0.63), (30,0.18), (40,0.60), (50,0.24) \}$$

For the dataset {12, 15, 20, 22, 25, 30} calculate mean and standard deviation and comment on data spread. **AKTU(2024-25)**

1. Given Data

X = {12, 15, 20, 22, 25, 30}
Number of observation (N) = 6

2. Calculate Mean

Formula:

$\{\bar{x}\} = \Sigma X / n$

$\Sigma X = 12 + 15 + 20 + 22 + 25 + 30 = 124$

$\bar{x} = 124 / 6 = 20.67$

Mean ($\bar{x}$) = 20.67

3. Calculate $(X - \bar{x})$ and $(X - \bar{x})^2$

$\Sigma(X - \bar{x})^2 = 215.36$

X	$X - \bar{x}$ (20.67)	$(X - \bar{x})^2$
12	-8.67	75.17
15	-5.67	32.11
20	-0.67	0.45
22	1.33	1.77
25	4.33	18.75
30	9.33	87.11
Total	=	215.36

4. Calculate Standard Deviation

(A) Population Standard Deviation (σ)

$\sigma = \sqrt{[\Sigma(X - \bar{x})^2 / n]}$

$= \sqrt{(215.36 / 6)}$

$= \sqrt{35.89}$

$= 5.99$

Standard Deviation (Population) σ = 5.99

(B) Sample Standard Deviation (s)

$s = \sqrt{[\Sigma(X - \bar{x})^2 / (n-1)]}$

$= \sqrt{(215.36 / 5)}$

$= \sqrt{43.07}$

$= 6.56$

Standard Deviation (Sample) s = 6.56

5. Comment on Data Spread

- Mean = 20.67
- Standard Deviation (Population) = 5.99 (approximately 6)
- The standard deviation is about 6, indicating that the data values are moderately spread around the mean.
- The observations are neither too close together nor too far apart, so the dataset shows moderate variability.

MEAN, MEDIAN, MODE & RANGE

For the following dataset {4, 7, 7, 9, 12}, calculate Mean, Median, Mode and Range.

1. Mean (Average)

Mean is the sum of all observations divided by the total number of observations.

Formula:

Mean ($\bar{X}$) = $\Sigma X / n$

where,
ΣX = Sum of all observations
n = Number of observations

Solution:

Given data: {4, 7, 7, 9, 12}

n = 5

$\Sigma X = 4 + 7 + 7 + 9 + 12 = 39$

Mean ($\bar{X}$) = $39 / 5 = 7.8$

Mean = 7.8

GATEWAY CLASSES

2. MEDIAN (Middle Value)

Median is the middle value of the data when arranged in ascending or descending order.

Formula / Rule:

- If n is odd: Median is the value at position (n + 1) / 2
- If n is even: Median is the average of the values at positions n / 2 and (n / 2) + 1

Solution:

Step 1: Arrange the data in ascending order.

4, 7, 7, 9, 12

Step 2: Since n = 5 (odd)

Median is the value at position (n + 1) / 2

= (5 + 1) / 2
= 3rd position

Median = 7

GATEWAY CLASSES

3. MODE (Most Frequent Value)

Mode is the value that appears most frequently in the dataset.

Formula / Rule:

The observation with the highest frequency is the mode.

Solution:

Given data: {4, 7, 7, 9, 12}

Frequencies:

Value (X)	Frequency (f)
4	1
7	2
9	1
12	1

The highest frequency is 2 for value 7

Mode = 7

4. RANGE (Spread of Data)

Range is the difference between the largest and the smallest value in the dataset.

Formula:

Range = Maximum value − Minimum value

Solution:

Given data: {4, 7, 7, 9, 12}

Maximum value = 12

Minimum value = 4

Range = 12 − 4

Range = 8

STANDARD DEVIATION & VARIANCE

For the dataset {4, 7, 9, 10, 12}, calculate Variance and Standard Deviation.

1. VARIANCE

Variance is the average of the squared deviations from the mean. It measures how much the data values spread out from the mean.

Formula:

Population Variance (σ^2):

$\sigma^2 = \Sigma(Xi - \mu)^2 / N$

where,

Xi = each observation

μ = mean of the data

N = total number of observations

Sample Variance (s^2):

$s^2 = \Sigma(Xi - \bar{x})^2 / (n - 1)$

where,

$\bar{x}$ = sample mean

n = total number of observations

We will use SAMPLE formulas in this example.

GW
GATEWAY CLASSES

2. STANDARD DEVIATION

Standard Deviation is the square root of variance. It tells us the average distance of data values from the mean.

Formula:

Population Standard Deviation (σ):

$\sigma = \sqrt{\sigma^2}$

Sample Standard Deviation (s):

$s = \sqrt{s^2}$

Interpretation:

- Smaller SD → data points are closer to mean.
- Larger SD → data points are more spread out.

3. SOLUTION

Given data: {4, 7, 9, 10, 12}

Step 1: Calculate Mean ($\bar{x}$)

$\bar{x} = (4 + 7 + 9 + 10 + 12) / 5$

$= 42 / 5$

$= 8.4$

Step 2: Create the Table

Xi	$Xi - \bar{x}$	$(Xi - \bar{x})^2$
4	$4 - 8.4 = -4.4$	19.36
7	$7 - 8.4 = -1.4$	1.96
9	$9 - 8.4 = 0.6$	0.36
10	$10 - 8.4 = 1.6$	2.56
12	$12 - 8.4 = 3.6$	12.96
Total	=	37.20

Step 3: Calculate Sample Variance (s^2)

$s^2 = \Sigma(Xi - \bar{x})^2 / (n - 1)$

$= 37.20 / (5 - 1)$

$= 37.20 / 4$

$= 9.30$

$s^2 = 9.30$

Step 4: Calculate Sample Standard Deviation (s)

$s = \sqrt{s^2}$

$= \sqrt{9.30}$

$= 3.05$

$s = 3.05$

GW
GATEWAY CLASSES

4. FINAL ANSWER

Sample Variance (s^2) = 9.30

Sample Standard Deviation (s) = 3.05

KEY POINTS

- Variance is in squared units.
- Standard Deviation is in original units.
- Lower SD → less variability.
- Higher SD → more variability.

GW
GATEWAY CLASSES

Explain the concept of Outliers. AKTU(2024-25)

An outlier is a data value that is significantly different from the other values in a dataset. It lies far away from the majority of the observations. Outliers may occur due to measurement errors or natural variations and can affect the accuracy of statistical analysis and machine learning models.

Principal component Analysis (PCA) : Mean Centering

Q : Consider the following dataset :

Find the mean – centered Data.

X	Y
2	4
4	6
6	8

Solution :

Step 1 : Calculate the Mean of X

Mean(X) = (2+4+6)/3 = 12/3 = 4

Step 2 : Calculate the Mean of Y

Mean(Y) = (4 + 6 + 8) /3 = 18/3 = 6

Step 3 : Perform Mean Clustering

Formula : Mean - Centered value = Original value - Mean

X	Y	X – Mean(X)	Y – Mean(Y)
2	4	2 – 4 = -2	4 - 6 = -2
4	6	4 – 4 = 0	6 - 6 = 0
6	8	6 – 4 = 2	8 – 6 = 2

Final Mean – Centered Data :

X	Y
-2	-2
0	0
2	2

The Mean – Centered dataset is :

(-2, -2)

(0,0)

(2,2)

Mean Centering shifts the data so that the mean of each feature becomes zero.

Principal component Analysis (PCA) : Covariance Matrix

Q : Consider the following dataset :

Find the mean – centered Data.

X	Y
2	4
4	6
6	8

Solution :

Step 1 : Calculate the Mean

Mean(X) = (2+4+6)/3 = 12/3 = 4

Mean(Y) = (4 + 6 + 8) /3 = 18/3 = 6

Step 2 : Find the Mean – Centered Values

Formula : Mean - Centered value = Original value - Mean

X	Y	X – Mean(X)	Y – Mean(Y)
2	4	2 – 4 = -2	4 - 6 = -2
4	6	4 – 4 = 0	6 - 6 = 0
6	8	6 – 4 = 2	8 – 6 = 2

Step 3 : Calculate Variance & Covariance

$Var(X) = \sum(X - Mean)^2 / (n-1)$

$Var(Y) = \sum(Y - Mean)^2 / (n-1)$

Formula for covariance :

$Cov(X,Y) = \sum[X - Mean)(Y-Mean)]/(n-1)$

Where ,

n = 3

Step 4 : Calculate Variance of X

$\text{Var}(X) = [(-2)^2 + (0)^2 + (2)^2] / (3-1) = (4 + 0 + 4) / 2 = 8 / 2 = 4$

Step 5 : Calculate Variance of Y

$\text{Var}(Y) = [(-2)^2 + (0)^2 + (2)^2] / (3-1) = (4 + 0 + 4) / 2 = 8 / 2 = 4$

Step 6 : Calculate Covariance

$\text{Cov}(X,Y) = [(-2 \times -2) + (0 \times 0) + (2 \times 2)] / (3-1) = (4 + 0 + 4) / 2 = 8 / 2 = 4$

Step 7 : Form the Covariance Matrix :

$$\textbf{Cov} = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}(Y) \end{bmatrix}$$

Covariance matrix : $\begin{bmatrix} 4 & 4 \\ 4 & 4) \end{bmatrix}$

Q : A dataset has eigenvalues : 4.5,2.8,1.2,0.4,0.1 **AKTU(2025-26)**

(1) Calculate percentage variance of each component

(2) Determine the number of components required to retain 90% variance.

Formula : Percentage Variance = Eigenvalue / Sum of Eigenvalues * 100

Step 1 : Find total Eigenvalues

Given Eigenvalues : 4.5, 2.8, 1.2, 0.4,0.1

Total = 4.5 + 2.8 + 1.2 + 0.4 + 0.1 = 9.0

Step 2 : Calculate Percentage Variance

Principal Component	Eigen value	Percentage variance
PC1	4.5	(4.5/9) *100 = 50.00%
PC2	2.8	(2.8/9) *100 = 31.11%
PC3	1.2	(1.2/9) *100 = 13.33%
PC4	0.4	(0.4/9) *100 = 4.44%
PC5	0.1	(0.1/9) *100 = 1.11%

Step 3 : Calculate cumulative Variance

Principal Component	Eigen value	Cumulative variance
PC1	50.00%	50.00%
PC2	31.11%	81.11%
PC3	13.33%	94.44%
PC4	4.44%	98.88%
PC5	1.11%	99.99% ≈ 100%

GATEWAY CLASSES

GATEWAY CLASSES

Step 4 : Retain 90% Variance : PC1 + PC2 +PC3 = 94.44%

3 principal components are required to retain 90% variance .

Q: Elaborate on the mathematical foundations of Support Vector Machines **AKTU(2023-24)**

Ans :

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression. It finds the best hyperplane that separates different classes with the maximum margin.

The mathematical foundation of SVM is based on:

Hyperplane – Decision boundary between two classes.

Support Vectors – Nearest data points to the hyperplane.

Maximum Margin – Maximizes the distance between classes for better accuracy.

Kernel Function – Used to classify non-linear data.

Hyperplane Equation: $w^T x + b = 0$, where , w = Weight vector

x = Input feature vector

b = Bias (intercept)

PCA ALGORITHM – SOLUTION

Given Data:

Class	X	Y
Class 1	2	1
	3	5
	4	3
	5	6
Class 2	6	7
	7	8

Step 1: Calculate the Mean

Mean of X:

$$\bar{X} = \frac{2+3+4+5+6+7}{6} = \frac{27}{6} = 4.5$$

$$\text{Mean}(X) = 4.5$$

Mean of Y:

$$\bar{Y} = \frac{1+5+3+6+7+8}{6} = \frac{30}{6} = 5$$

$$\text{Mean}(Y) = 5$$

Step 2: Mean Centering

Formula:

$$X_{new} = X - \bar{X} \quad , \quad Y_{new} = Y - \bar{Y}$$

X	Y	X – 4.5	Y – 5
2	1	−2.5	−4
3	5	−1.5	0
4	3	−0.5	−2
5	6	0.5	1
6	7	1.5	2
7	8	2.5	3

Step 3: Calculate Covariance Matrix

Formula: $Cov(X, Y) = \frac{\sum (X_i Y_i)}{n-1}$

Here, $n = 6 \Rightarrow n - 1 = 5$

$$\text{Cov(X, X)} = \frac{(-2.5)^2 + (-1.5)^2 + (-0.5)^2 + (0.5)^2 + (1.5)^2 + (2.5)^2}{5}$$

$$\text{Cov(Y, Y)} = \frac{(-4)^2 + 0^2 + (-2)^2 + (1)^2 + (2)^2 + (3)^2}{5} = 6.8$$

$$\text{Cov(X, Y)} = \frac{(-2.5)(-4) + (-1.5)(0) + (-0.5)(-2) + (0.5)(1) + (1.5)(2) + (2.5)(3)}{5} = 4.4$$

Since, Cov(Y, X) = Cov(X, Y)

$$\text{Covariance Matrix} = \begin{bmatrix} 3.5 & 4.4 \\ 4.4 & 6.8 \end{bmatrix}$$

Step 4: Calculate Eigenvalues

Formula: $|A - \lambda I| = 0$

$$\begin{vmatrix} 3.5 - \lambda & 4.4 \\ 4.4 & 6.8 - \lambda \end{vmatrix} = 0$$

$$(3.5 - \lambda)(6.8 - \lambda) - 4.4 \times 4.4 = 0$$

$$\lambda^2 - 10.31\lambda + 4.46 = 0$$

Eigenvalues are

$$\lambda_1 = 9.84, \quad \lambda_2 = 0.46$$

Step 5: Percentage Variance

Total Eigenvalues = 9.84 + 0.46 = 10.30

- **For PC1:**

$$\frac{9.84}{10.30} \times 100 = 95.5\%$$

- **For PC2:**

$$\frac{0.46}{10.30} \times 100 = 4.5\%$$

PC1 explains 95.5% variance
PC2 explains 4.5% variance

Step 6: Principal Component

Since PC1 has the highest variance (95.5%),

Principal Component = PC1

Final Result (Summary)

- Mean(X) = 4.5
- Mean(Y) = 5
- Covariance Matrix = $\begin{bmatrix} 3.5 & 4.4 \\ 4.4 & 6.8 \end{bmatrix}$
- Eigenvalues = 9.84, 0.46
- Percentage Variance = 95.5%, 4.5%
- Principal Component = PC1

Conclusion:

The first principal component (PC1) explains 95.5% of the total variance, so it is selected as the principal component for dimensionality reduction.

: SOLUTION – PCA

Given data = {4, 8, 13, 7, 11, 4, 5, 14}.

Compute the principal component using PCA algorithm.
Also use PCA to reduce dimension from 2 to 1.

Step 1: Arrange the Data

Sample (i)	1	2	3	4
X	4	8	13	7
Y	11	4	5	14

Number of samples (n) = 4

Step 2: Calculate the Mean

Mean of X:

$$\bar{X} = \frac{4+8+13+7}{4} = \frac{32}{4} = 8$$

Mean(X) = 8

Mean of Y:

$$\bar{Y} = \frac{11+4+5+14}{4} = \frac{34}{4} = 8.5$$

Mean(Y) = 8.5

Step 3: Mean Centering

Formula:

$$X_{new} = X - \bar{X}, \quad Y_{new} = Y - \bar{Y}$$

Sample (i)	X	Y	X – 8	Y – 8.5
1	4	11	–4	2.5
2	8	4	0	–4.5
3	13	5	5	–3.5
4	7	14	–1	5.5

Step 4: Compute Covariance Matrix

Formula: $\text{Cov}(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$

Here, $n = 4 \Rightarrow n - 1 = 3$

$$\text{Cov}(X, X) = \frac{(-4)^2 + 0^2 + 5^2 + (-1)^2}{3} = \frac{16+0+25+1}{3} = \frac{42}{3} = 14$$

Cov(Y, Y):

$$= \frac{(2.5)^2 + (-4.5)^2 + (-3.5)^2 + (5.5)^2}{3}$$

$$= \frac{6.25 + 20.25 + 12.25 + 30.25}{3} = \frac{69}{3} = 23$$

Cov(X, Y):

$$= \frac{(-4)(2.5) + (0)(-4.5) + (5)(-3.5) + (-1)(5.5)}{3}$$

$$= \frac{-10 + 0 - 17.5 - 5.5}{3} = \frac{-33}{3} = -11$$

Since, $\text{Cov}(Y, X) = \text{Cov}(X, Y) = -11$

$$\text{Covariance Matrix} = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

Step 5: Calculate Eigenvalues

Let, $A = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$

Formula: $|A - \lambda I| = 0$

$$\begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} = 0$$

$$\Rightarrow (14-\lambda)(23-\lambda) - (-11)(-11) = 0$$

$$\Rightarrow (14-\lambda)(23-\lambda) - 121 = 0$$

$$\Rightarrow \lambda^2 - 37\lambda + 201 = 0$$

$$\boxed{\lambda = \frac{37 \pm \sqrt{37^2 - 4(1)(201)}}{2}}$$

$$\lambda = \frac{37 \pm \sqrt{1369 - 804}}{2} = \frac{37 \pm \sqrt{565}}{2}$$

Eigenvalues:

$$\boxed{\lambda_1 = 30.32, \quad \lambda_2 = 6.68}$$

$(\lambda_1 > \lambda_2)$

Step 6: Calculate Percentage Variance

Total Eigenvalues = 30.32 + 6.68 = 37.00

For PC1:

$$\frac{30.32}{37.00} \times 100 = 81.95\%$$

For PC2:

$$\frac{6.68}{37.00} \times 100 = 18.05\%$$

PC1 explains 81.95% variance

PC2 explains 18.05% variance

Step 7 : Dimension Reduction (2 to 1)

Since PC1 has the highest variance (81.95%), we select the first principal component and discard the second.

Reduced dimension = 1

(Keep PC1, Discard PC2)

Final Answer

- Mean(X) = 8
- Mean(Y) = 8.5
- Covariance Matrix = $\begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$
- Eigenvalues: 30.32, 6.68
- PC1 explains 81.95% variance
- PC2 explains 18.05% variance
- Dimension reduced from 2 to 1 by selecting PC1.

Previous Year Important Questions

Q. 1 Explore PCA. Given data = {4, 8,13 ; 11,4,5,14} . Compute the principal component using PCA algorithm. Also use PCA to reduce dimension from 2 to 1. **AKTU(2024-25)**

Q.2 Explain the concept of Outliers. **AKTU(2024-25)**

Q. 3 A = {(10,0.2) , (20,0.4),(25,0.7),(30.0.9),(40,1),(50.0.4)} **AKTU(2024-25)**

B = {(10,0.4) , (20,0.1),(25,0.9),(30.0.2),(40,0.6),(50.0.6)}

Apply Union , Intersection , Complement , bold Union , and Bold intersection operations on above listed Fuzzy sets. **AKTU(2024-25)**

Q. 4 For the dataset {12 , 15, 20, 22, 25, 30} , Calculate mean & Standard deviation and comment on data spread. **AKTU(2024-25)**

Q.5 Explain the concept of generalization in neural networks . How does it relate to the trade – off between bias and variance , & what strategies can be employed to enhance generalization performance?

AKTU(2023-24)

GW GATEWAY CLASSES

GW GATEWAY CLASSES

Q.6 Elaborate on the mathematical foundations of Support Vector Machines. **AKTU(2023-24)**

Q.7 The following data represent study hours(X) and scores(Y) : **AKTU(2025-26)**

X : 2,4,6,8

Y : 50,55,65,70,78

(1) Find the linear Regression Equation.

(2) Predict Y when X = 7

Q. 8 A dataset has eigenvalues : 4.5,2.8,1.2,0.4,0.1 **AKTU(2025-26)**

(1) Calculate percentage variance of each component

(2) Determine the number of components required to retain 90% variance.

GATEWAY CLASSES

GATEWAY CLASSES

Q.9 Describe PCA algorithm. Compute the Principal component using PCA algorithm on given data as :

AKTU(2023-24)

CLASS 1 : X = 2,3,4 Y = 1,5,3

CLASS 2 : X = 5,6,7 Y = 6,7,8

Gateway Classes

Empowering Learners, Transforming Futures

THANK YOU

GET IT ON Google Play

Download on the App Store

Helpline No- 7819 0058 53 , 9259833200