DS IMP QUES

2M

UNIT 1

1. DEFINE ALGEBRAIC VIEW OF 2D AND 3D

Algebraic View of 2D:

Describes geometric objects in a plane using coordinates (x, y) and equations.

Point: (x, y)

Line: $y = mx + c$

Circle: $(x - a)^2 + (y - b)^2 = r^2$

Distance: $\sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$

Algebraic View of 3D:

Describes spatial objects using coordinates (x, y, z) and equations.

Point: (x, y, z)

Line: Vector/parametric form

Plane: $ax + by + cz = d$

Sphere: $(x - a)^2 + (y - b)^2 + (z - c)^2 = r^2$

Distance: $\sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2]}$

2. CALCULATE RANK AND NULLITY OF A MATRIX

- The rank of a matrix is the maximum number of linearly independent rows or columns in the matrix.
- Convert the matrix to Row Echelon Form (REF) or Reduced Row Echelon Form (RREF) using Gaussian elimination. Count the number of non-zero rows in that form. rank(A)

Find the rank of the matrix $A=\begin{pmatrix}0&1&2&1\\1&2&3&2\\3&1&1&3\end{pmatrix}$

Solution:

The order of A is 3 × 4.

$\therefore \rho(A) \leq 3$.

Let us transform the matrix A to an echelon form

Matrix A	Elementary Transformation
$A=\begin{pmatrix}0&1&2&1\\1&2&3&2\\3&1&1&3\end{pmatrix}$	
$A=\begin{pmatrix}1&2&3&2\\0&1&2&1\\3&1&1&3\end{pmatrix}$	$R_1 \leftrightarrow R_2$
$\sim\begin{pmatrix}1&2&3&2\\0&1&2&1\\0&-5&-8&-3\end{pmatrix}$	$R_3 \rightarrow R_3 - 3R_1$
$\sim\begin{pmatrix}1&2&3&2\\0&1&2&1\\0&0&2&2\end{pmatrix}$	$R_3 \rightarrow R_3 + 5R_2$

The number of non zero rows is 3. $\therefore \rho(A) = 3$.

2. Nullity of a Matrix

- **Definition:**
  The **nullity** of a matrix is the **dimension of the null space**, i.e., the number of free variables in the solution of the homogeneous system $A\mathbf{x} = 0$.
- **Formula (Rank-Nullity Theorem):**

$$\text{nullity}(A) = n - \text{rank}(A)$$

where n is the number of **columns** in matrix A.

Example:

Given matrix

$$A = \begin{bmatrix}1&2&3\\2&4&6\\3&6&9\end{bmatrix}$$

- Convert to row echelon form:

$$\begin{bmatrix}1&2&3\\0&0&0\\0&0&0\end{bmatrix} \Rightarrow \text{Rank} = 1$$

- Number of columns $n = 3$
- $$\text{Nullity} = 3 - 1 = 2$$

3. WHAT HAPPENS WHEN 2 VECTORS ARE PERPENDICULAR

When 2 Vectors Are Perpendicular:

- They form a 90° angle.
- Their dot product is 0.

Example:

If
A = (2, 3)
B = (3, –2)

Then
$A \cdot B = 2\times3 + 3\times(-2) = 6 - 6 = 0$
→ So, they are perpendicular.

Key Rule:
If $A \cdot B = 0$, the vectors are perpendicular.

UNIT 2

1. DEFINE PDF

A Probability Density Function (PDF) is a mathematical function that shows how likely a continuous random variable is to take on a certain value.

Key Points:

- Used for continuous data (like height, weight, time, etc.)
- The total area under the curve of a PDF is 1 (100% probability)

Example:

- In a normal distribution (bell curve), the PDF tells you values near the mean are more likely than values far away.

2. DIFFERENTIATE NULL HYPOTHESIS AND FEATURE SCALING

Aspect	**Null Hypothesis**	**Feature Scaling**
Field	Statistics / Hypothesis Testing	Machine Learning / Data Pre-processing
Purpose	To test assumptions about data	To normalize/standardize feature values
What it does	Assumes no effect or no difference	Adjusts data to a common scale
Example	"There is no difference between group A and B"	Changing ages from 5–90 to a 0–1 range
Used in	Statistical tests (t-test, ANOVA, etc.)	Algorithms like KNN, SVM, Gradient Descent

UNIT 3

1. DEFINE EDA AND ITS TYPES

Exploratory Data Analysis (EDA) is the process of analysing and visualizing data to:

- Understand its main characteristics
- Detect patterns, trends, and outliers
- Prepare data for further modeling or analysis

1. Univariate Analysis

- Examines one variable at a time and helps understand the distribution, central tendency, and spread.
- Tools: Histogram, Box plot, Mean, Median, Mode

2. Bivariate Analysis

- Examines the relationship between two variables and helps detect correlation, association, or patterns.
- Tools: Scatter plot, Correlation coefficient, Line plot, Crosstab

2. GIVE THE TOOLS OF DATA VISUALISATION

1. Tableau

- Type: Drag-and-drop software (no coding)
- Use: Creating interactive dashboards and reports
- Best for: Business Intelligence (BI), presentations

2. Power BI

- Type: Microsoft tool (low-code)
- Use: Data analysis and business dashboards
- Best for: Business users, Excel integration

3. Matplotlib

- Type: Python library
- Use: Creating basic and custom plots (line, bar, scatter, etc.)
- Best for: Programmers and data scientists

UNIT 4

1. DEFINE ML AND DIFFERENTIATE ITS TYPES

Machine Learning is a branch of Artificial Intelligence (AI) that enables computers to learn from data and make decisions or predictions without being explicitly programmed.

Types of ML:

Type	Description	Example
Supervised Learning	Model learns from labeled data	Spam detection, price prediction
Unsupervised Learning	Model finds patterns in unlabeled data	Customer segmentation, clustering
Reinforcement Learning	Model learns by interacting with environment and rewards	Game playing, robotics

2. DIFFERENTIATE PCA AND NN

Aspect	PCA (Principal Component Analysis)	NN (Neural Network)
Type	Dimensionality Reduction Technique	Machine Learning Model
Purpose	Reduces number of features while keeping information	Learns patterns to make predictions
Input Requirement	Works on numeric, continuous features	Can handle various types of data
Learning	Unsupervised	Can be supervised or unsupervised
Complexity	Simple, linear	Complex, nonlinear

3. DEFINE OVER FITTING AND UNDER FITTING

Term	**Definition**	**Behavior**	**Example**
Overfitting	Model learns too much, including noise and irrelevant details.	High accuracy on training, poor on testing	A student memorizes answers without understanding; good in mock tests, fails real exam
Underfitting	Model learns too little, fails to capture the underlying trend.	Poor on both training and testing data	A student reads only the chapter titles; performs poorly in all exams

12M

1. FIND THE EIGEN VALUE AND EIGEN VECTOR FOR THE FOLLOWING MATRIX

https://www.youtube.com/watch?v=C21DPLcmCTE 2*2

https://www.youtube.com/watch?v=wiA5bZ1kVxU 3*3

2. FIND THE PSEUDO INVERSE FOR A MATRIX

Pseudo inverse or Moore – Penrose inverse is the generalization of the matrix inverse that may not be invertible. If the matrix is invertible then its inverse will be equal to pseudo inverse and denoted by A+.

- If the columns of a matrix A are linearly independent, so $A^T \cdot A$ is invertible and we obtain with the following formula the pseudo inverse:

$$\mathbf{A^+ = (A^T \cdot A)^{-1} \cdot A^T}$$

- Here A^+ is a left inverse of A, what means: $A^+ \cdot A = E$.
- However, if the rows of the matrix are linearly independent, we obtain the pseudo inverse with the formula:

$$\boldsymbol{A^+ = A^T \cdot (A \cdot A^T)^{-1}}$$

- This is a right inverse of A, what means: $A \cdot A^+ = E$.
- If both the columns and the rows of the matrix are linearly independent, then the matrix is invertible and the pseudo inverse is equal to the inverse of the matrix.

If A has rank deficient, then the Pseudo inverse of A id defined as

$$\boldsymbol{A^+ = (U\Sigma V^T)^{-1} = (V^T)^{-1}\Sigma^{-1}U^{-1} = V\Sigma^{-1}U^T}$$

$$\text{If } \Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix} \text{ then } \boldsymbol{\Sigma^{-1}} = \begin{bmatrix} \frac{1}{\sigma_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sigma_2} & \mathbf{0} \end{bmatrix}$$

1. *Find the pseudo inverse of* $A=\begin{bmatrix}1 & 2 & 1 & 3\\ 4 & 3 & 2 & 1\end{bmatrix}$

Sol:

$$Given\ A=\begin{bmatrix}1 & 2 & 1 & 3\\ 4 & 3 & 2 & 1\end{bmatrix}$$

$$Here\begin{vmatrix}1 & 2\\ 4 & 3\end{vmatrix}=3-8=-5\neq 0$$

$$rank(A)=2$$

Then the pseudo inverse of A is $A^{+}=A^{T}(AA^{T})^{-1}$

$$A=\begin{bmatrix}1 & 2 & 1 & 3\\ 4 & 3 & 2 & 1\end{bmatrix}\quad A^{T}=\begin{bmatrix}1 & 4\\ 2 & 3\\ 1 & 2\\ 3 & 1\end{bmatrix}$$

$$AA^{T}=\begin{bmatrix}1 & 2 & 1 & 3\\ 4 & 3 & 2 & 1\end{bmatrix}\begin{bmatrix}1 & 4\\ 2 & 3\\ 1 & 2\\ 3 & 1\end{bmatrix}=\begin{bmatrix}15 & 15\\ 15 & 30\end{bmatrix}$$

$$|AA^{T}|=15(30-15)=225$$

$$(AA^{T})^{-1}=\frac{1}{225}\begin{vmatrix}30 & -15\\ -15 & 15\end{vmatrix}=\begin{bmatrix}2/15 & -1/15\\ -1/15 & 1/15\end{bmatrix}$$

$$A^{+}=A^{T}(AA^{T})^{-1}=\begin{bmatrix}1 & 4\\ 2 & 3\\ 1 & 2\\ 3 & 1\end{bmatrix}\begin{bmatrix}2/15 & -1/15\\ -1/15 & 1/15\end{bmatrix}=\begin{bmatrix}-2/15 & 3/15\\ 1/15 & 1/15\\ 0 & 1/15\\ 5/15 & -2/15\end{bmatrix}$$

$$=\frac{1}{15}\begin{bmatrix}-2 & 3\\ 1 & 1\\ 0 & 1\\ 5 & -2\end{bmatrix}$$

3. EXPLAIN UNIVARIATE AND MULTIVARIATE DISTRIBUON

Definition:

The Univariate Gaussian Distribution, also known as the Normal Distribution, is a continuous probability distribution involving only one variable. It is represented by a bell-shaped curve and is one of the most commonly used distributions in statistics and data science.

The probability density function (PDF) for a univariate Gaussian is:

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Where:

- μ is the **mean** (center of the distribution)
- σ^2 is the **variance**, and σ is the **standard deviation**
- The total area under the curve is **1**, representing total probability

Key Properties:

1. **Symmetric** around the mean μ
2. **Mean = Median = Mode**
3. **Spread** is controlled by the standard deviation σ
4. Most values lie within:
    - 68% within 1σ
    - 95% within 2σ
    - 99.7% within 3σ

Example:

- Heights of adult humans
- Test scores in a large population
- The distribution helps understand how values are spread around a central point and is essential for statistical modeling and hypothesis testing.

Importance:

- Easy to work with mathematically
- Foundation for many statistical methods
- Supports the Central Limit Theorem (CLT)

Definition:

The Multivariate Gaussian Distribution is an extension of the univariate normal distribution to two or more variables. It models multiple correlated continuous variables simultaneously and is useful in high-dimensional statistical analysis.

The PDF for a **k-dimensional** multivariate normal distribution is:

$$f(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Where:

- x is a vector of variables
- μ is the **mean vector**
- Σ is the **covariance matrix**
- $|\Sigma|$ is the determinant of Σ

Key Concepts:

- Mean vector (μ): Indicates the center for each variable
- Covariance matrix (Σ): Measures how variables vary together (correlation)
- The shape of the distribution is elliptical in 2D, ellipsoid in higher dimensions
- If variables are uncorrelated, the ellipses are circular or aligned to axes
- If variables are correlated, the ellipses are tilted

Example:

- Modeling height and weight of people together
- Financial risk analysis (e.g., return on different stocks)

Importance:

- Easy to work with mathematically
- Central to multivariate statistical methods
- Supports a multivariate version of the Central Limit Theorem

4. EXPLAIN HYPOTHESIS AND HYPOTHESIS TESTING

1. Definition of Hypothesis

A hypothesis is a tentative statement or assumption made about a population parameter such as the mean, proportion, or variance. It serves as the basis for statistical testing and is used to make inferences from sample data.

Example: "The average weight of apples is 150g." This can be tested using statistical methods.

2. What is Hypothesis Testing

Hypothesis testing is a formal procedure in statistics used to test claims or assumptions about a population parameter by analysing sample data.

It helps determine whether the sample data provides sufficient evidence to reject the null hypothesis in favour of the alternative hypothesis.

It is commonly used to compare:

- A single group with a known value or standard.
- Two or more groups with each other.

3. Types of Hypotheses

Type	Symbol	Meaning
Null Hypothesis	H_o	Assumes no effect or no difference exists.
Alternative Hypothesis	H_1 or Ha	Assumes there is an effect or a difference.

H_o always includes equality (=), while Ha includes inequality (≠, >, <).

Example:

- H_o: $\mu_1 = \mu_2$ (No difference in means)
- Ha: $\mu_1 \neq \mu_2$ (Difference exists)

4. Key Concepts in Hypothesis Testing

Term	Explanation
Level of Significance (α)	The threshold probability of making a Type I error. Commonly 0.05 (5%).
Test Statistic	A standardized value (e.g., z, t, F) used to evaluate H_o.
P-value	The probability of observing a result as extreme as the test statistic, under H_o.
Confidence Level	Probability of correctly accepting a true H_o (e.g., 95%).

Decision Rule:

If p-value < α, reject H_o.

If p-value ≥ α, fail to reject H_o.

5. Steps of Hypothesis Testing

- State H_o and Ha clearly.
- Set α (significance level), usually 0.05.
- Collect and analyse sample data.
- Calculate the test statistic and p-value.
- Compare p-value with α.
- Decide: Reject or fail to reject H_o.
- Interpret the result in the context of the problem.

6. Types of Errors in Hypothesis Testing

Error	Explanation	Symbol
Type I Error	Rejecting a true H_o (False positive)	α
Type II Error	Failing to reject a false H_o (False negative)	β

Power of the test = $1 - \beta \rightarrow$ Ability to detect true effects.

7. Types of Tests (0.5 Mark)

Test	Used For
Z-test	Comparing means when population SD is known
T-test	Comparing means when SD is unknown
Chi-square	Categorical data and independence tests
ANOVA	Comparing more than two means

8. Example to Illustrate

A company claims the average salary is ₹30,000.

H_o: $\mu = 30,000$

Ha: $\mu \neq 30,000$

After collecting data and testing, if $p < 0.05$, we reject H_o and conclude the average salary is not ₹30,000.

Advantages of Hypothesis Testing

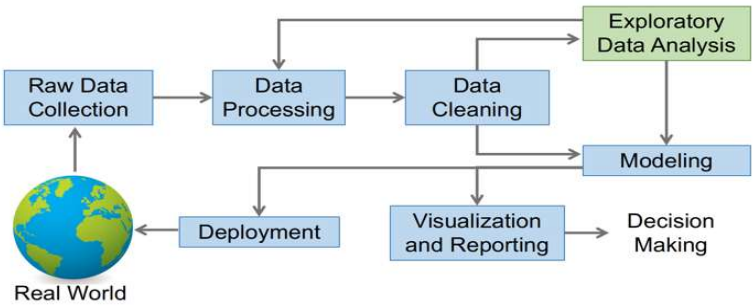
- Objective Decision-Making
- Helps Validate Assumptions
- Quantifies Risk (α, β)

Disadvantages of Hypothesis Testing

- Misuse of P-values
- Highly Sensitive to Sample Size
- Assumptions Required

5. EXPLAIN THE DATA SCIENCE PROCESS IN DETAIL

The Data Science Process is a structured framework that guides data scientists through solving real-world problems using data. It involves a series of steps from understanding the problem to deploying machine learning models.



The key steps involved in Data Science Modelling are:

Step 1: Understanding the Problem

- The first and most critical step.
- Involves communicating with stakeholders to understand the business challenge clearly.
- Data scientists identify what data is needed and what kind of solution is expected.
- Helps determine if a data science or AI approach is appropriate.
- Example: A business wants to predict customer churn; the data scientist must understand why customers are leaving and what data is available.

Step 2: Data Extraction

Relevant raw data is gathered from multiple sources:

- Internal databases
- APIs
- Web scraping
- Online repositories (e.g., Kaggle, UCI)
- Data can be structured (tables) or unstructured (text, images, etc.).
- The goal is to collect as much relevant and quality data as possible.

Step 3: Data Cleaning & Pre-processing

One of the most time-consuming steps in the data science process.

Includes:

- Handling missing values
- Removing duplicates
- Fixing incorrect formats
- Encoding categorical variables
- Also involves feature scaling:
- Standardization (mean = 0, std = 1)
- Normalization (scales between 0 and 1)
- Helps ensure that models are not misled by dirty or biased data.

Step 4: Exploratory Data Analysis (EDA)

Involves visualizing and summarizing the dataset to understand patterns.

Tools like histograms, boxplots, and correlation matrices are used.

Identifies:

- Trends
- Outliers
- Relationships between variables
- Example: A scatter plot might show a linear relationship between salary and experience.

Step 5: Feature Selection

- Helps identify the most relevant variables for the model.
- Removes redundant or irrelevant features that could reduce accuracy.

Improves:

- Model performance
- Training speed
- Interpretability
- Can be done using:
- Statistical tests
- Correlation analysis

Step 6: Applying Machine Learning Algorithms

Models are built based on the nature of the problem:

1. Supervised Learning:

Uses labelled data (with known outcomes)

Examples:

- Linear Regression for predictions
- Random Forest, SVM for classification

2. Unsupervised Learning:

No labelled output; finds structure in data

Examples:

- K-Means Clustering
- Hierarchical Clustering
- Anomaly Detection

Step 7: Model Evaluation and Tuning

- After training, models are tested using metrics:
- Accuracy, Precision, Recall, F1-Score, etc.
- Hyperparameter tuning is done to improve performance (e.g., using GridSearchCV).
- Final model is selected based on performance on validation/test data.

Step 8: Deployment and Monitoring

- The final model is deployed to a production environment where it can make real-time decisions.
- Ongoing monitoring ensures the model remains accurate over time.

Basic Tools in the Data Science Process

Step	Tool(s)
1. Understanding the Problem	– Notebooks (Jupyter, Colab) – Business tools (Excel, Google Sheets)
2. Data Extraction	– SQL, APIs – Python libraries: pandas, requests, beautifulsoup4
3. Data Cleaning & Pre-processing	– Python: pandas, numpy – Excel – R
4. Exploratory Data Analysis (EDA)	– matplotlib, seaborn, plotly, pandas-profiling
5. Feature Selection	– scikit-learn, statsmodels, SelectKBest, correlation heatmaps
6. Applying ML Algorithms	– scikit-learn, xgboost, tensorflow, keras, lightgbm
7. Model Evaluation & Tuning	– scikit-learn (metrics, GridSearchCV), mlflow, optuna
8. Deployment & Monitoring	– Flask, FastAPI, Docker, Streamlit, cloud (AWS, GCP, Azure), MLflow, Airflow

These tools are often used in combination through the data science lifecycle. Let me know if you'd like platform-specific tools (like R, Excel-based, or cloud tools only).

6. EXPLAIN PCA IN DETAIL

** IF THEORY ATTEND

** IF SUM DON'T ATTEND

Definition and Purpose

Principal Component Analysis (PCA) is an unsupervised machine learning technique used for dimensionality reduction. It transforms a large set of correlated variables into a smaller set of uncorrelated variables, called Principal Components, while retaining most of the information (variance) from the original data.

Why PCA is Used

- To reduce dimensionality in high-dimensional datasets.
- To remove redundancy and correlated features.
- To improve visualization, model performance, and computational efficiency.
- Commonly used in fields like image processing, recommender systems, and bioinformatics.

Key Concepts

Term	Meaning
Dimensionality	Number of features or columns in the dataset.
Variance	Measure of how much values in a feature differ from the mean.
Covariance	Measures how much two variables change together.
Eigenvectors	Directions of the new feature space (axes).
Eigenvalues	Magnitudes (importance) of those directions.
Orthogonal	Uncorrelated directions (zero correlation).
Principal Components	New variables formed from the linear combinations of original ones.

Properties of Principal Components

- Linearly uncorrelated and orthogonal to each other.
- The first principal component (PC1) captures the maximum variance.
- Each successive component captures the remaining variance under the constraint of being orthogonal to previous ones.
- The number of principal components ≤ number of original features.

Steps in PCA Algorithm

1. **Obtain the Dataset**

   Collect data and divide into **features (X)** and labels (Y), if applicable.

2. **Standardize the Data**

   Standardize features so they contribute equally using:

$$Z = \frac{X - \text{mean}}{\text{standard deviation}}$$

3. **Calculate Covariance Matrix**

   Find relationships between variables:

$$\text{Covariance Matrix} = Z^T \cdot Z$$

4. **Compute Eigenvalues and Eigenvectors**

   Determine the directions (eigenvectors) and their strength (eigenvalues).

5. **Sort Eigenvectors**

   Rank them by descending eigenvalues — the higher the eigenvalue, the more information it holds.

6. **Form the Principal Components**

   Multiply the top K eigenvectors with standardized data to get:

$$Z^* = Z \cdot P^*$$

7. **Reduce Dimensions**

   Keep only top **K principal components** that explain most of the variance.

Benefits of PCA

- Simplifies models by reducing the number of variables.
- Removes noise and multicollinearity.
- Improves training speed and sometimes accuracy.
- Aids in data visualization (e.g., 2D or 3D scatter plots).

Applications

- Face recognition
- Image compression
- Finance and portfolio analysis
- Genetics and biology
- Movie/music recommendation systems