AI - Unit 5

Data Processing

Instructor: Saptarshi Jana

Learning Objectives

After studying this chapter, students will be able to:[a]

- Understand the concept and importance of data modeling.
- Identify different types of data models.
- Explain regression analysis and its applications.
- Understand simple linear regression and its equation.
- Solve linear equations and apply them to real-life problems.

[a] Unit 5: Theoretical and Practical Aspects of Data Processing

1 Introduction to Data Modeling

Data modeling is the process of representing real-world data in a structured and organized form so that it can be efficiently stored, processed, and analyzed. It helps in understanding relationships among data entities and ensures consistency in data handling.

Data modeling serves as the foundation for databases, data analytics, and machine learning systems. It converts raw data into a meaningful structure that supports decision-making.

1.1 Importance of Data Modeling

- Data modeling helps in organizing large volumes of data in a structured and systematic manner.
- It reduces data redundancy and minimizes inconsistencies across datasets.
- Data modeling improves data accuracy while maintaining data integrity throughout the system.
- It supports effective data analysis and facilitates accurate prediction and decision-making.

2 Types of Data Models

2.1 Relational Data Model

The **Relational Data Model** organizes data into tables (relations) consisting of rows and columns. Each row represents a record, while each column represents an attribute.

Example: A student table containing attributes such as ID, Name, and Age.

2.2 Entity–Relationship (ER) Model

The **Entity–Relationship Model** represents data using the following components:

- Entities represent real-world objects such as students or teachers.
- Attributes describe the properties of entities, such as name or age.
- Relationships define how entities are connected to each other, such as enrollment or teaching associations.

It is commonly used during the database design phase.

2.3 Dimensional Data Model

The **Dimensional Model** is mainly used in data warehousing and analytical applications. It consists of the following components:

- Fact tables store numerical and measurable data used for analysis.

- Dimension tables store descriptive data that provide context to the numerical values.

This model is optimized for fast data retrieval and efficient data analysis.

2.4 Other Types of Data Models (Optional)

There are several other data models are widely used in database systems and applications.

- **Hierarchical Data Model** The hierarchical data model organizes data in a tree-like structure, where each parent record can have multiple child records, but each child record has only one parent. This model is simple to implement but is not suitable for representing complex relationships.

- **Network Data Model** The network data model extends the hierarchical model by allowing a child record to have multiple parent records. It supports many-to-many relationships but is complex in design and maintenance.

- **Object-Oriented Data Model** The object-oriented data model represents data as objects and supports object-oriented programming concepts such as inheritance, encapsulation, and polymorphism. This model is suitable for handling complex data structures.

- **Object-Relational Data Model** The object-relational data model combines features of both relational and object-oriented models. It supports complex data types while maintaining a table-based structure.

- **Physical Data Model** The physical data model describes how data is physically stored in the database, including storage structures, indexes, and access paths. It focuses on performance and implementation details.

3 Regression Analysis

Regression analysis is a statistical technique used to study the relationship between variables and to predict the value of one variable based on another.

3.1 Dependent and Independent Variables

- **Independent Variable (X)**: The variable that influences the outcome. It is used to make predictions.

- **Dependent Variable (Y)**: It is the variable whose value is predicted.

Example: Study hours (X) and exam marks (Y).

3.2 Types of Regression

- *Linear regression* is used to model the relationship between a dependent variable and a single independent variable.
- *Multiple regression* analyzes the relationship between a dependent variable and two or more independent variables.
- *Polynomial regression* is used when the relationship between variables is non-linear and can be represented using polynomial terms.
- *Logistic regression* is used to predict categorical or binary outcomes based on independent variables.

4 Linear Regression

Linear regression is a statistical technique used to model the relationship between two variables by fitting a straight line to the observed data. It helps in predicting the value of one variable based on another.

4.1 Linear Regression Equation

The equation of a straight line is given by:

$$y = mx + c$$

where:

- y = dependent variable
- x = independent variable
- m = regression coefficient (slope)
- c = intercept

4.2 Least Squares Regression Line

The **least squares method** determines the best-fit line by minimizing the sum of squared differences between actual and predicted values.

4.3 Properties of Linear Regression

- Linear regression assumes a linear relationship between the independent and dependent variables.
- It minimizes the prediction error by fitting the best possible line to the given data.
- Linear regression is widely used for trend analysis and forecasting future values.

4.4 Regression Coefficient

The regression coefficient indicates the following characteristics:

- It represents the strength of the relationship between the independent and dependent variables.
- It shows the direction of the relationship, which may be positive or negative.

4.5 Types of Linear Regression

- Simple Linear Regression
- Multiple Linear Regression

5 Solving Linear Equations

Solving linear equations helps in understanding variable relationships and forms the mathematical basis of linear regression.Let's see an example of it.

A real estate company wants to estimate the price of a house based on its size in square feet. The relationship is given by:

$$\text{Price} = 1{,}000{,}000 + 4{,}000 \times \text{Area}$$

Calculate the estimated price of a house with an area of 120 square feet.

$$\text{Price} = 1{,}000{,}000 + 4{,}000 \times 120 \text{Price} = 1{,}000{,}000 + 480{,}000 = 1{,}480{,}000$$

Therefore, the estimated house price is **1,480,000**.

- Independent variable (X): Area of the house (in square feet)
- Dependent variable (Y): Price of the house
- The regression coefficient **4,000** indicates that for every additional square foot of area, the house price increases by 4,000.
- The intercept **1,000,000** represents the minimum or starting price of the house.

6 Applications of Linear Regression

- Linear regression is used to predict exam scores based on relevant input variables.
- It is widely applied in sales and demand forecasting to estimate future trends.
- Linear regression helps in estimating house prices by analyzing factors such as location, size, and amenities.
- It is used for trend analysis to identify patterns and changes over time.
- Linear regression supports decision-making systems by providing data-driven insights.