Data Engineering Techniques (AID30010) — Unit I

Introduction to Data Engineering and Preprocessing

Structured Question Bank

PART B: LONG ANSWER QUESTIONS (5 marks)

Introduction to Data Engineering

Q1: What is Data Engineering? Explain the Data Engineering Ecosystem.

Answer:

1. Data Engineering — Definition

- Data engineering is the development, implementation, and maintenance of systems and processes that take in raw data and produce high-quality, consistent information
- This information supports downstream use cases such as analysis and machine learning
- Data engineering is the intersection of data management, data architecture, and software engineering
- Data engineers manage the data engineering lifecycle — starting from ingestion and ending with serving data for use cases like analysis or machine learning

2. Key Concepts

- **Data Governance (DG):** Process of managing the availability, usability, integrity, and security of data in enterprise systems
- **Data Wrangling:** Process of removing errors and combining complex data sets to make them more accessible and easier to analyze

3. Roles in the Data Engineering Ecosystem

- **Data Architect:** Designs, creates, and manages an organization's overall data infrastructure
- **Data Engineers:** Deal with raw data that may contain human, machine, or instrument errors
- **QA Engineer:** Ensures quality by testing, identifying bugs, and verifying products meet standards
- **Product Managers (PMs):** Define a product's vision, strategy, and roadmap — act as a central hub connecting business goals, user needs, and technology
- **Data Analysts:** Focus on past and present data to find insights (what happened and why)
- **Data Scientists:** Build predictive models for the future — they receive cleaned data
- **Stakeholders:** Owners, employees, managers, customers, shareholders, suppliers

4. Data Engineering Lifecycle

- The lifecycle begins with data ingestion and ends with serving data for analysis or ML
- Engineers ensure that data is clean, accessible, and usable at every stage

---

Q2: Explain the Evolution of Database Technology and the concept of Data Mining (KDD).

Answer:

1. Evolution of Database Technology

- **1960s:** Data collection, database creation, IMS and network DBMS were developed
- **1970s:** Relational data model introduced; relational DBMS implemented
- **1980s:** RDBMS widely adopted; advanced data models developed (object-oriented, deductive); application-oriented DBMS introduced (spatial, scientific, engineering)
- **1990s:** Data mining, data warehousing, multimedia databases, and web databases emerged
- **2000s:** Stream data management, data mining applications, and web technology became prominent

2. What is Data Mining?

- Data mining is also called **Knowledge Discovery in Databases (KDD)**
- It is the extraction of interesting, non-trivial, implicit, previously unknown, and potentially useful patterns or knowledge from a huge amount of data
- Alternative names: knowledge extraction, data/pattern analysis, data archeology, information harvesting, business intelligence

3. What Counts as Data Mining?

- Simple search and query processing is NOT data mining
- Expert systems are NOT data mining
- Patterns discovered must be: **valid, novel, potentially useful, and understandable**

4. Data Mining Algorithms — Three Parts

- **Model:** To be fit on data
- **Preference:** Criteria to select one model over another
- **Search:** Techniques to evaluate data points or search through data

5. Goal of Data Mining

- To provide efficient tools and techniques for Knowledge Discovery in Databases (KDD)
- Example: from a class's student scores, extract useful patterns about student performance

6. Why is Data Mining Important?

- Rapid computerization produces huge amounts of data
- Used to discover patterns and relationships to help make better business decisions
- Automated prediction of trends and behaviors
- Automated discovery of previously unknown patterns
- Data is growing from terabytes to petabytes to zettabytes

Q3: Explain the KDD (Knowledge Discovery in Databases) Process with all steps.

Answer:

1. KDD Process — Overview

- KDD is the overall process of converting raw data into useful knowledge
- Data mining plays an essential role inside the KDD process

2. Step 1 — Understanding the Domain

- Understand the application domain and relevant prior knowledge
- Understand the goals of the end-user
- Create a target dataset: select a data set or focus on a subset of variables

3. Step 2 — Data Cleaning and Preprocessing

- Remove noise or outliers
- Collect information to model or account for noise
- Handle missing data fields
- Account for time sequence information and known changes

4. Step 3 — Data Reduction and Projection

- Find useful features to represent the data depending on the goal
- Use dimensionality reduction or transformation methods to reduce the number of variables

5. Step 4 — Choosing the Data Mining Task

- Decide whether the goal is classification, regression, clustering, etc.

6. Step 5 — Choosing the Algorithm

- Select the method to be used for searching for patterns
- Decide which models and parameters are appropriate

- Match the data mining method with the overall KDD process criteria

7. Step 6 — Data Mining

- Search for patterns of interest in a particular form (classification rules, decision trees, regression, clustering, etc.)

8. Step 7 — Interpreting and Consolidating

- Interpret the mined patterns
- Consolidate discovered knowledge and present it to the user

9. Sequence Summary

- Data Cleaning → Data Integration → Data Selection → Data Transformation → Data Mining → Pattern Evaluation → Knowledge Presentation

---

Types of Data

Q4: Explain the types of data: Structured, Semi-structured, and Unstructured.

Answer:

1. Structured Data

- Data with a high degree of organization, typically stored in a spreadsheet or table
- Excel sheets and relational database tables are prime examples
- Easily machine-readable; can be analysed without major pre-processing
- Around 20% of the world's data is structured
- Example formats: Excel (.xlsx), CSV, relational database tables

2. Semi-Structured Data

- Data which has some degree of organization but is not as rigidly structured as structured data
- Achieved using tags or other elements that define hierarchy and system within a file
- The order and amount of structuring tags may vary
- Needs some pre-processing before analysis
- Example formats: HTML, JSON, XML files
- Gained importance with the emergence of the World Wide Web

3. Unstructured Data

- Data with no pre-defined organizational form or specific format

- Essentially anything that is not structured or semi-structured
- Needs major pre-processing before it can be analysed by a computer
- Often easily consumable for humans (pictures, videos, plain text)
- Most of the data created today is unstructured
- Example formats: images (.jpeg, .png), videos (.mp4), audio files (.mp3, .wav), plain text, Word files, PDF files

4. Key Differences

Feature	Structured	Semi-Structured	Unstructured
Organization	High	Some	None
Format	Spreadsheet/Table	HTML/JSON/XML	Any format
Pre-processing needed	Minimal	Some	Major
Examples	Excel, CSV	JSON, XML	Images, Videos

Q5: Explain Discrete and Continuous data types with examples and differences.

Answer:

1. Discrete Data

- The term discrete means distinct or separate
- Contains values that fall under integers or whole numbers
- Data cannot be broken into decimal or fraction values
- Discrete data is **countable** and has **finite values**; subdivision is not possible
- Represented mainly by bar graphs, number lines, or frequency tables
- Examples: total number of students in a class, cost of a cell phone, number of siblings

2. Continuous Data

- Data in the form of fractional numbers (decimals)
- Represents information that can be divided into smaller levels
- The continuous variable can take any value within a range
- Represented in the form of a histogram
- Examples: height of a person, temperature of a room, length of an object, version number of an app

3. Differences — Discrete vs. Continuous

Discrete Data	Continuous Data
Countable and finite; whole numbers or integers	Measurable; in the form of fractions or decimals
Represented mainly by bar graphs	Represented by histograms
Values cannot be subdivided into smaller pieces	Values can be divided into smaller subdivisions
Spaces between values	Continuous sequence — no gaps
Example: number of students in a class	Example: temperature of a room

---

Q6: Explain the types of data: Nominal, Ordinal, Binary, Interval, and Ratio.

Answer:

1. Qualitative (Categorical) Data

- Data that cannot be measured or counted in numbers
- Sorted by category, not by number
- Consists of audio, images, symbols, or text
- Example: gender of a person (male/female/other)

2. Nominal Attributes

- Values are names or symbols representing categories or states
- Also called categorical attributes
- **No order or rank** among the values
- Example: hair color, blood type, profession
- Cannot perform arithmetic tasks on nominal data

3. Binary Attributes

- Data has only 2 values/states
- Example: yes or no, true or false
- **Symmetric binary:** both values are equally important — no preference for coding as 0 or 1 (Example: gender)
- **Asymmetric binary:** both values are NOT equally important — most important outcome coded as 1

4. Ordinal Attributes

- Values that have a meaningful sequence or ranking (order) between them

- But the exact magnitude/difference between values is not known
- Example: customer satisfaction ratings (low, medium, high), academic ranks

5. Quantitative (Numeric) Data

- Data that can be expressed in numerical values
- Can be used for statistical manipulation
- Also known as Numerical data

6. Interval-Scaled Attributes

- Have order
- Values can be added and subtracted but **cannot be multiplied or divided**
- Do **not** have a true zero
- Example: Temperature in Celsius or Fahrenheit — 10°C is NOT twice as hot as 5°C

7. Ratio-Scaled Attributes

- Have all properties of interval-scaled attributes
- Also have a **true zero**
- Values can be added, subtracted, multiplied, and divided
- Example: weight, height, salary

Q7: Explain Time Series Data and Geographical Data.

Answer:

1. Time Series Data — Definition

- When data is collected over equal intervals of time, it is called time series data
- The time intervals can be as small as seconds or as large as years
- The only condition is the time period should be of equal intervals and clearly defined
- Examples: stock prices of a company throughout the year, sales data over 3 years, daily expenses over 6 months

2. Characteristics of Time Series Data

- **Equal Interval:** Data must be collected at equal time intervals — not uneven
- **Seasonality:** When the data is influenced by seasonal factors such as rainfall or temperature
- **Cyclicity:** When data rises and falls at intervals that are not at fixed periods (e.g., business cycles)
- **Trend:** When there is a long-term increase or decrease in the data — called uptrend or downtrend

3. Geographical Data — Definition

- A type of information that captures the physical features of a place
- Includes location of a city or town, layout of streets and highways, and natural features like rivers and mountains

4. Uses of Geographical Data

- Mapping out areas for planning purposes
- Social media networks use it to provide information about users' location
- Construction companies use it to map out locations of roads and highways
- Presented in two-dimensional format (paper maps) or three-dimensional digital means (virtual globes)
- **GIS (Geographic Information Systems):** Software applications that allow users to view, edit, and analyze geographical data in three dimensions

---

Measures of Central Tendency and Dispersion

Q8: Explain the Measures of Central Tendency: Mean, Median, and Mode.

Answer:

1. What are Measures of Central Tendency?

- Used to numerically summarize data by identifying the central position within that data set
- Also called summary statistics
- Three main measures: Mean, Median, Mode

2. Mean

- Represents the average value of the dataset
- Calculated by taking the sum of all values and dividing by the number of values
- The most popular and well-known measure of central tendency
- Can be used with both discrete and continuous data, but most often with continuous data
- Includes every value in the dataset as part of the calculation
- **Limitation:** Highly susceptible to the influence of outliers — unusually small or large values can distort the mean

3. Median

- Median is the middle value of the dataset when arranged in ascending or descending order
- When dataset has an even number of values, median = mean of the two middle values
- Example: 2, 5, 6, 7, 9, **10**, 12, 13, 15, 16, 18, 21, 23 → Median = 12 (6 values above, 6 below)

- **Advantage:** Less affected by outliers and skewed data than the mean — preferred when distribution is not symmetrical
- **Limitation:** Unsuitable for fractions and percentages; not suitable for further algebraic calculations

4. Mode

- Represents the most frequently occurring value in the dataset
- A dataset may have multiple modes or no mode at all
- Example: 5, 4, 2, 3, 2, 1, 5, 4, 5 → Mode = 5 (appears 3 times)
- **Advantage:** Can be found for both numerical and categorical (non-numerical) data
- **Limitation:** In some distributions, the mode may not reflect the center very well

5. How to Select the Right Measure

- Symmetrical distribution of continuous data → all three measures are valid; mean is preferred
- Skewed distribution → median is best
- Categorical data → mode is the best choice
- Original/raw data available → both median and mode work well

---

Q9: Explain Measures of Dispersion: Range, Variance, and Standard Deviation.

Answer:

1. What are Measures of Dispersion?

- Give information on the spread, variability, or dispersion of data values
- Help understand how far values are spread out from the center
- Two sets of data may have the same central value but differ in variability

2. Range

- Difference between the highest value and the lowest value in a data set
- Formula: Range = Maximum value − Minimum value
- Very sensitive to outliers — a single data value can greatly affect the result
- **Limitation:** Extremely coarse measure; does not tell us about the distribution between max and min
- Example: race finishing times = {16, 21, 28, 12, 30, 21, 24} → Range = 30 − 12 = **18**

3. Standard Deviation

- Tells us how spread out the observed values are from the mean
- Standard Deviation is always a **positive value**

- It is the square root of the variance
- If data points are farther from the mean → higher standard deviation (more spread)
- Calculated as: SD = √(Variance)
- Where Variance = sum of squared differences of each observation from the mean

4. Variance

- The average deviation of all data values from the mean
- The **square of standard deviation** is called variance
- Difference between SD and variance: only in terms of unit of measurement
- Example: If height is in meters and SD = 0.67 m, then Variance = 0.45 m^2

5. Relationship Between the Three

- Range is simple but crude (only uses max and min)
- Variance uses all data points but gives squared units
- Standard deviation uses all data points and gives same units as the original data

---

Data Preprocessing

Q10: What is Data Preprocessing? Why is it important? Explain the major tasks.

Answer:

1. What is Data Preprocessing?

- Process of transforming raw data into an understandable, usable format
- An important step in data mining because we cannot work with raw data directly
- Data preprocessing is the first and crucial step when creating a machine learning model
- The goal is to improve the quality of data and make it suitable for the specific data mining task

2. Why is it Important?

- Real-world data generally contains noise, missing values, and may be in an unusable format
- "No quality data, no quality mining results!"
- Quality decisions must be based on quality data
- Data preparation, cleaning, and transformation comprises about **90% of the work** in a data mining application
- Duplicate or missing data may cause incorrect or even misleading statistics

3. Data Quality Parameters

- **Accuracy:** Whether the data entered is correct or not
- **Completeness:** Whether the data is available or recorded
- **Consistency:** Whether the same data matches across all places
- **Timeliness:** Whether the data is updated correctly
- **Believability:** Whether the data is trustable
- **Interpretability:** Whether the data is understandable

4. Four Major Tasks in Data Preprocessing

- **Data Cleaning:** Remove noise and correct inconsistencies in the data
- **Data Integration:** Merge data from multiple sources into a coherent data store (e.g., a data warehouse)
- **Data Reduction:** Reduce the data size by aggregating, eliminating redundant features, or clustering
- **Data Transformation:** Normalize or scale data to fall within a smaller, specified range (e.g., 0.0 to 1.0)

Data Cleaning

Q11: Explain Data Cleaning in detail — handling Missing Values, Noisy Data, and Outliers.

Answer:

1. What is Data Cleaning?

- Process of fixing or removing incorrect, incomplete, duplicate, or otherwise unusable data
- Three main issues handled: Missing Values, Noisy Data, Inconsistencies

2. Missing Values

- Missing data appears as empty cells, NULL, N/A, or special characters like "?"
- **Causes of missing data:**
  - Equipment malfunction
  - Data inconsistent with other records and thus deleted
  - Data not entered due to misunderstanding
  - Certain data not considered important at the time of entry
  - No history or changes recorded

3. Methods to Handle Missing Values

- **Ignore the tuple:** Simply ignore records with missing values (only works when many attributes are missing)
- **Fill in manually:** Human fills in the missing value — impractical for large datasets
- **Use a global constant:** Fill with "Unknown" or a fixed value like 0

- **Use measures of central tendency:** Fill missing value with the mean, median, or mode of the attribute
- **Use class-based mean/median:** Fill using mean or median for all samples belonging to the same class
- **Use the most probable value:** Use regression, Bayesian formalism, or decision tree induction to predict the missing value

4. Noisy Data

- Noise is a random error or variance in a measured variable
- **Binning Method:** Smooth a sorted data value by consulting its "neighborhood"
    - Sorted values are distributed into a number of "buckets" or bins
    - **Smoothing by bin means:** Replace values in each bin with the bin's mean
    - **Smoothing by bin boundaries:** Replace values in each bin with the nearest boundary value

5. Regression-Based Imputation

- Data smoothing by conforming data values to a function — this is called regression
- **Linear regression:** Finds the best line to fit two attributes so one can predict the other
- **Multiple linear regression:** Extension of linear regression involving more than two attributes
- Example: Missing age values in the Titanic dataset can be predicted from Pclass, SibSp, Parch, and Fare

6. Outliers

- Values in a dataset that differ significantly from other observations
- Example: most people hold breath for 15–75 seconds; someone holding for 10 minutes is an outlier
- **Outlier Detection Methods:**
    - Values more than 3 times the standard deviation from the mean are considered outliers
    - Z-score greater than +3 or less than −3 is an outlier
    - Using K-means algorithm or DBSCAN clustering

7. Data Cleaning as a Process

- First step: **Discrepancy Detection** — finding inconsistencies in the data
- Discrepancies are caused by: poorly designed data entry forms, human error, deliberate errors, data decay, inconsistent use of codes
- **Tools used:**
    - Data scrubbing tools: use domain knowledge (postal addresses, spell-checking) to detect and correct errors
    - Data auditing tools: find discrepancies by analyzing data and detecting rule violations using statistical analysis or clustering
    - ETL (Extract/Transform/Load) tools: allow users to specify transforms through a GUI

Data Integration

Q12: Explain Data Integration and the problems associated with it.

Answer:

1. What is Data Integration?

- Combines data from multiple sources into a coherent store (like a data warehouse)
- Careful integration reduces redundancy and inconsistencies
- Improves accuracy and speed of subsequent data mining
- Data integration techniques: schema matching, instance conflict resolution, source selection, result merging, quality composition

2. Issues in Data Integration

Entity Identification Problem:

- How to match equivalent real-world entities from multiple data sources?
- Example: "Bill Clinton" in one database and "William Clinton" in another refer to the same person
- Schema integration: e.g., A.cust-id ≡ B.cust-number
- Use metadata (meaning, data type, range of values) from different sources to integrate correctly
- Ensure attribute functional dependencies and referential constraints match in source and target

Redundancy and Correlation Analysis:

- Redundant data occurs often when integrating multiple databases
- An attribute may have different names in different databases (object identification problem)
- One attribute may be a "derived" attribute in another table (e.g., annual revenue derived from monthly sales)
- Redundant attributes can be detected using **correlation analysis** and **covariance analysis**

3. Detecting Redundancy

- **χ^2 Test (Chi-Square Test):** Used for nominal/categorical/qualitative data
  - Test whether two attributes are independent of each other
  - Null hypothesis H_0: attributes are independent
  - If calculated χ^2 > critical value from table → reject H_0 → attributes are correlated
- **Correlation Coefficient and Covariance:** Used for numeric/quantitative data

4. Correlation Analysis

- Correlation: a mutual relationship or association between quantities
- **Positive Correlation:** As one attribute increases, the other also increases (e.g., study hours and test scores, work hours and earnings)
- **Negative Correlation:** As one increases, the other decreases (e.g., TV time and exam marks, loan payment and remaining debt)
- Pearson's Correlation Coefficient:
  - Greater than 0 → positively correlated
  - Equal to 0 → no correlation (independent)
  - Less than 0 → negatively correlated
  - Higher value = stronger correlation = possible redundancy

5. Covariance vs. Pearson's Coefficient

Covariance	Pearson's Coefficient
Can be any real number from $-\infty$ to $+\infty$	Always lies between -1 and $+1$
Magnitude depends on units	Unitless
Shows direction of relationship	Shows both direction and strength

6. Tuple Duplication

- Same record may exist multiple times across different databases
- Must be detected and removed to maintain data integrity

7. Data Value Conflict Detection and Resolution

- Same real-world entity may have different attribute values from different sources
- Example: room prices in different cities may use different currencies
- Attributes may differ on abstraction level — "total sales" in one database may mean one branch, while in another it means all stores in a region
- Data values must be converted to a consistent form

Data Transformation

Q13: Explain Data Transformation techniques — Normalization, Aggregation, Generalization, and Discretization.

Answer:

1. What is Data Transformation?

- Process of converting data into appropriate forms suitable for mining
- Improves accuracy and efficiency of mining algorithms

2. Types of Data Transformation

- **Smoothing:** Remove noise from data (using binning, clustering, regression)
- **Aggregation:** Summarization or data cube construction — e.g., daily sales aggregated to monthly
- **Generalization:** Concept hierarchy climbing — replacing low-level concepts with higher-level ones
- **Normalization:** Scale data to fall within a small, specified range
- **Attribute/Feature Construction:** New attributes constructed from existing ones (e.g., area = height × width)

3. Normalization — Min-Max Normalization

- Performs a linear transformation on original data
- Maps a value to the range [new_min, new_max]
- Formula: v' = (v − min_A) / (max_A − min_A) × (new_max_A − new_min_A) + new_min_A
- Example: income range is \$12,000 to \$98,000; value of \$73,600 normalizes to **0.716** in range [0.0, 1.0]
- Preserves the relationships among original data values

4. Normalization — Z-Score Normalization

- Also called zero-mean normalization
- Values for an attribute A are normalized based on the **mean and standard deviation** of A
- Formula: v' = (v − mean_A) / σ_A
- Example: mean income = \$54,000, SD = \$16,000; a value of \$73,600 normalizes to approximately 1.225
- Useful when minimum and maximum of the attribute are unknown

5. Normalization — Decimal Scaling

- Normalizes by moving the decimal point of values
- Formula: v' = v / 10^j — where j is the smallest integer such that Max(|v'|) < 1
- Example: values range from −986 to 917; maximum absolute value = 986; divide by 1000 → −986 becomes −0.986, 917 becomes 0.917

6. Concept Hierarchy Climbing (Generalization)

- Replaces low-level data with higher-level, more general concepts
- Example: specific street addresses → city → state → country
- Reduces the data to a more manageable level

Data Reduction

Q14: Explain Data Reduction strategies — Dimensionality Reduction, Numerosity Reduction, and Data Compression.

Answer:

1. What is Data Reduction?

- Obtaining a reduced representation of the data set that is much smaller in volume but produces the same (or almost the same) analytical results
- A database or data warehouse may store terabytes of data — complex analysis may take very long on the full dataset
- Goal: reduce data size while maintaining integrity of results

2. Data Reduction Strategies

- Dimensionality reduction
- Data cube aggregation
- Data compression
- Numerosity reduction — parametric and non-parametric methods
- Discretization and concept hierarchy generation

3. Dimensionality Reduction — Attribute Subset Selection

- Removes irrelevant or redundant attributes (dimensions) from the dataset
- Goal: find a minimum set of attributes such that the probability distribution of classes is as close as possible to the original
- Reduces the number of attributes → patterns are easier to understand
- **Heuristic (Greedy) Methods:**
    - **Stepwise Forward Selection:** Start with empty set; add the best attribute one at a time in each iteration
    - **Stepwise Backward Elimination:** Start with all attributes; remove the worst one in each iteration
    - **Combination:** Combine forward selection and backward elimination — add best and remove worst at each step
    - **Decision Tree Induction:** Build a decision tree; attributes NOT part of the tree are considered irrelevant and discarded

4. Numerosity Reduction

- **Parametric Methods:**

- Assume data fits some model
- Estimate model parameters and store only the parameters, discard the data
- Example: Linear regression — $Y = \alpha + \beta X$
- Multiple regression: $Y = b_0 + b_1X_1 + b_2X_2$

- **Non-Parametric Methods:**
  - Do not assume any models
  - Major families: histograms, clustering, sampling

5. Histograms

- At least a century old and widely used
- Graphical method for summarizing the distribution of a given attribute X
- Range of values for X is partitioned into disjoint consecutive subranges called **buckets or bins**
- Height of each bar indicates the frequency (count) of values in that range
- Typically, buckets are of equal width (equal-width partitioning)
- **Equal-frequency (equal-depth) partitioning:** Each bucket has approximately the same number of data points
- Example: price attribute with range $1–$200 partitioned into $1–$20, $21–$40, $41–$60, etc.

6. Clustering

- Partition the dataset into clusters; store only the cluster representation (not all data points)
- Quality measured by: diameter (max distance between any two objects in cluster) or centroid distance
- Very effective if data is naturally clustered
- Not effective if data is "smeared" across the space

7. Sampling

- Obtaining a small sample s to represent the whole data set N
- Allows a mining algorithm to run in complexity that is sub-linear to the size of the data
- **Types of Sampling:**
  - **Simple Random Sampling:** Equal probability of selecting any item; may perform poorly with skewed data
  - **Sampling without Replacement:** Once an object is selected, it is removed from the population
  - **Sampling with Replacement:** Selected object is not removed; can be selected again
  - **Stratified Sampling:** Divide data into subgroups (strata) based on shared characteristics; randomly sample from each stratum; used with skewed data

8. Data Compression

- **String Compression:** Extensive theories and algorithms; typically lossless; limited manipulation without expansion
- **Audio/Video/Image Compression:** Typically lossy — with progressive refinement; small fragments can sometimes be reconstructed
- **Time Series Compression:** Typically short and vary slowly with time

---

Data Discretization

Q15: Explain Data Discretization and Concept Hierarchies in detail.

Answer:

1. What is Data Discretization?

- Process of converting continuous numerical data into a finite set of discrete categories or "bins"
- Simplifies large datasets for easier analysis
- Improves model performance
- Handles noisy data by grouping values
- Example: turning ages 1–100 into "child," "teen," "adult," "senior"
- Example: turning temperature values into "low," "medium," "high"

2. Three Types of Attributes

- **Nominal:** Values from an unordered set (e.g., color, profession)
- **Ordinal:** Values from an ordered set (e.g., military or academic rank)
- **Continuous (Numeric):** Real numbers (e.g., integers or real-valued numbers)

3. Why Discretization?

- Some classification algorithms only accept categorical attributes
- Reduce data size by discretization
- Interval labels can replace actual data values — simpler representation
- Prepare data for further analysis (e.g., classification)
- Can be supervised or unsupervised; split (top-down) or merge (bottom-up)

4. Discretization Methods

- **Binning:** Top-down split; unsupervised — divide data into bins and replace values
- **Histogram Analysis:** Top-down split; unsupervised — use frequency histograms to determine breakpoints

- **Clustering Analysis:** Can be top-down or bottom-up; unsupervised — cluster data and use clusters as intervals
- **Decision Tree Analysis:** Supervised; top-down split — use entropy-based measures to find best split points
- **Correlation Analysis (χ^2):** Unsupervised; bottom-up merge — merge adjacent intervals with similar distributions

5. Concept Hierarchies — Definition

- A concept hierarchy organizes data into a tree-like structure
- Each level of the hierarchy represents a concept more general than the level below it
- Used to organize and classify data in a way that makes it more understandable and easier to analyze
- The same data can have different levels of granularity (detail)

6. Example of Concept Hierarchy

- **Numerical:** Age → Youth, Adult, Senior (replacing exact ages with conceptual labels)
- **Nominal (Location):** Street → City → State → Country
    - E.g., Root: Location → USA, India → New York, Illinois, Gujarat, UP → specific streets

7. Benefits of Concept Hierarchies

- **Improved Data Analysis:** Groups similar concepts together — helps identify patterns and trends
- **Improved Data Visualization and Exploration:** Tree-like structure makes large, complex datasets easier to navigate
- **Improved Algorithm Performance:** Algorithms can more easily process hierarchically organized data
- **Data Cleaning and Pre-processing:** Identify and remove outliers and noise
- **Domain Knowledge:** Represents domain knowledge in a structured, understandable way

8. Applications of Concept Hierarchies

- **Data Warehousing:** Organize data from multiple sources into a consistent structure
- **Business Intelligence:** Analyze customer data to identify patterns for product or service development
- **Online Retail:** Organize products into categories, subcategories, and sub-subcategories
- **Healthcare:** Group patients by diagnosis or treatment plan
- **Natural Language Processing:** Identify topics and themes in text data
- **Fraud Detection:** Identify patterns in financial data that indicate fraudulent activity

---

ADDITIONAL PRACTICE QUESTIONS

Q16: Explain the Binning method for handling noisy data with a detailed example.

Answer:

1. What is Binning?

- A method to smooth a sorted data value by consulting its "neighborhood" (values around it)
- Sorted values are distributed into a number of "buckets" or bins
- Used to handle noisy data in the data cleaning stage

2. Steps in Binning

- Sort the data
- Partition into bins (each bin has equal number of values in equi-depth partitioning)
- Apply a smoothing method on each bin

3. Smoothing by Bin Means

- Replace every value in a bin with the **mean (average)** of that bin
- All values in the bin become the same — the bin's mean

4. Smoothing by Bin Boundaries

- Replace every value in a bin with the **nearest boundary** of that bin (minimum or maximum)
- Each value is compared to both boundaries; it is replaced by whichever is closer

5. Example

- Sorted data: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215
- **Equi-depth partitioning** into 3 bins (4 values each):
  - Bin 1: {5, 10, 11, 13}
  - Bin 2: {15, 35, 50, 55}
  - Bin 3: {72, 92, 204, 215}
- **Smoothing by Bin Means:**
  - Bin 1 mean = (5+10+11+13)/4 = 9.75 → all values become 9.75
  - Bin 2 mean = (15+35+50+55)/4 = 38.75 → all values become 38.75
  - Bin 3 mean = (72+92+204+215)/4 = 145.75 → all values become 145.75
- **Smoothing by Bin Boundaries:**

- Bin 1 {5, 10, 11, 13}: Boundaries 5 and 13 → 5→5, 10→13, 11→13, 13→13 → Smoothed: **5, 13, 13, 13**
- Bin 2 {15, 35, 50, 55}: Boundaries 15 and 55 → 15→15, 35→15, 50→55, 55→55 → Smoothed: **15, 15, 55, 55**
- Bin 3 {72, 92, 204, 215}: Boundaries 72 and 215 → 72→72, 92→72, 204→215, 215→215 → Smoothed: **72, 72, 215, 215**

Q17: Compare and contrast the different types of Normalization techniques with examples.

Answer:

1. Why Normalization?

- Data from different sources may have very different scales
- Normalization scales data to a common range so that no single attribute dominates due to its larger values
- Improves accuracy and efficiency of mining algorithms that involve distance measurements

2. Min-Max Normalization

- Transforms values linearly to a new range [new_min, new_max]
- Formula: v' = (v − min_A) / (max_A − min_A) × (new_max_A − new_min_A) + new_min_A
- Typically maps to [0.0, 1.0]
- **Advantage:** Preserves original relationships between data values
- **Limitation:** If a new value falls outside the original [min, max] range, it causes an out-of-bounds error
- Example: income range $12,000–$98,000; map to [0, 1] → $73,600 → 0.716

3. Z-Score Normalization

- Normalizes using the mean and standard deviation
- Formula: v' = (v − mean_A) / σ_A
- Result is in terms of how many standard deviations away from the mean
- **Advantage:** Useful when min/max are unknown or there are outliers
- **Limitation:** Does not bound values to a specific range
- Example: mean income = $54,000, SD = $16,000 → $73,600 normalizes to ≈ 1.225

4. Decimal Scaling Normalization

- Normalizes by moving the decimal point
- Formula: v' = v / 10^j (where j is the smallest integer such that Max(|v'|) < 1)
- Example: values range from −986 to 917; divide by 1000 → −986 becomes −0.986

5. Comparison

Normalization Type	Formula Used	Range After	Best When
Min-Max	Linear transformation	[new_min, new_max] e.g. [0, 1]	Known min/max, no extreme outliers
Z-Score	Mean and SD	No fixed range	Unknown min/max, or outliers present
Decimal Scaling	Divide by 10^j	(−1, 1) approximately	Simple scaling is needed

---

Q18: What is Data Mining? Explain its Applications and the difference between Data, Information, and Knowledge.

Answer:

1. Data Mining — Definition

- Extraction of interesting, non-trivial, implicit, previously unknown, and potentially useful patterns or knowledge from a huge amount of data
- Also called Knowledge Discovery in Databases (KDD)
- Other names: knowledge extraction, data archeology, information harvesting, business intelligence

2. Why Data Mining?

- Explosive growth of data: from terabytes to petabytes to zettabytes
- Data and storage predictions for 2025: 90ZB of data created on IoT devices; 49% stored in public cloud
- "We are drowning in data, but starving for knowledge!"
- Data mining automates the analysis of massive datasets
- Provides automated prediction of trends and behaviors
- Discovers previously unknown patterns

3. Applications of Data Mining

- Direct marketing
- Health industry
- E-commerce
- Customer Relationship Management (CRM)
- Telecommunication industry and financial sector

- **Specialized forms:** Text mining, web mining, audio/video data mining, pictorial data mining, social networks data mining, relational database mining

4. Data vs. Information vs. Knowledge

	Data	Information	Knowledge
What it is	Raw facts and figures	Processed and organized data	Insights and understanding derived from information
Example	On Monday, total sales were ₹5,400; product P103 had highest quantity sold (5 units); sales dropped to zero for P104	On Monday, product P103 sold the most and P104 had zero sales	We should stock more of P103 because it's popular; investigate why P104 didn't sell — maybe its price, display position, or demand isn't good

5. Example of Discovered Pattern (Association Rule)

- "80% of customers who buy cheese and milk also buy bread, and 5% of customers buy all three together"
- Written as: Cheese, Milk → Bread

Summary

This question bank covers all major topics in Unit I with:

- 18 Long Answer Questions (5 marks each)
- Complete coverage of all Unit I concepts including:
  - Data Engineering definition and ecosystem
  - Evolution of database technology and KDD
  - KDD process with all steps
  - Types of data: Structured, Semi-structured, Unstructured
  - Data types: Nominal, Ordinal, Binary, Interval, Ratio
  - Discrete vs. Continuous data
  - Time Series and Geographical data
  - Measures of Central Tendency: Mean, Median, Mode
  - Measures of Dispersion: Range, Variance, Standard Deviation
  - Data Preprocessing overview and importance
  - Data Cleaning: Missing values, Noisy data, Outliers, Binning
  - Data Integration: Entity identification, Redundancy, Correlation analysis, Chi-square test
  - Data Transformation: Normalization (Min-Max, Z-Score, Decimal Scaling), Generalization

    - Data Reduction: Dimensionality reduction, Attribute subset selection, Histograms, Sampling, Clustering, Compression
    - Data Discretization and Concept Hierarchies
- Each answer is structured for a 5-mark exam response
- Simple, student-friendly language throughout
- Practical examples included wherever possible