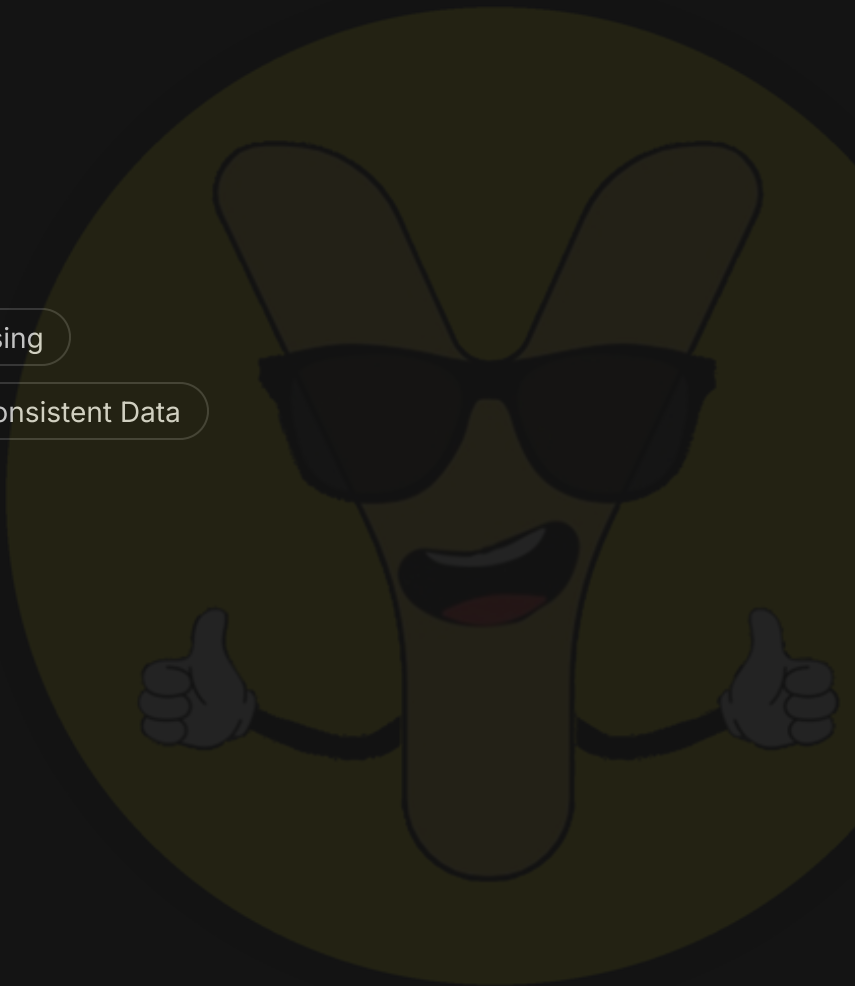
AKTU · 3RD YEAR · SEMESTER 5 · BCS058

Data Warehouse & Data Mining

Unit 3 — Data Mining and Pre-processing

- Data Mining: Overview, Motivation, Definition & Functionalities
- Data Processing
- Form of Data Pre-processing
- Data Cleaning: Missing Values, Noisy Data (Binning, Clustering, Regression, Computer and Human Inspection), Inconsistent Data
- Data Integration and Transformation
- Data Reduction: Data Cube Aggregation, Dimensionality Reduction, Data Compression, Numerosity Reduction



DECODED BY

Chirag Jain

Founder, Ycotes · Gold Medalist in Artificial Intelligence, IIT Madras · taught 10,000+ students

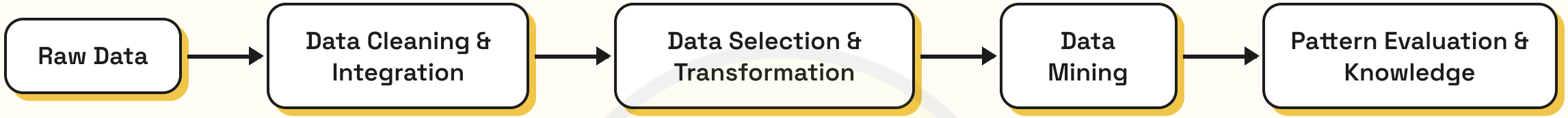
Data Mining: Overview & Definition

DATA MINING

Data mining is the process of discovering interesting patterns, non-trivial knowledge, anomalies, and structure from large amounts of data stored in databases, data warehouses, or other repositories.

- Motivated by the explosive growth of data and the urgent need to convert data into actionable knowledge.
- Forms the core analytical step in Knowledge Discovery in Databases (KDD).
- Enables predictive decision-making, fraud detection, and customer behavior analysis.

The KDD Process Pipeline



Sequential stages of turning raw operational data into useful domain knowledge.

Data Mining Functionalities

Data mining tasks are broadly categorized into descriptive (summarizing properties) and predictive (performing inference).

- Concept Description
  - Summarizes data using Data Characterization and Data Discrimination.
- Mining Frequent Patterns
  - Discovers association rules and frequent itemsets like $A \rightarrow B$.
- Classification & Regression
  - Constructs models to predict categorical labels or continuous values.
- Cluster Analysis
  - Groups data into clusters based on maximum intra-class similarity without labels.
- Outlier Analysis
  - Identifies objects that deviate significantly from general data behavior.

Forms of Data Pre-processing

- **Data Cleaning** (Clean missing, noisy, and inconsistent data)
- **Data Integration** (Combine multiple data sources into one store)
- **Data Transformation** (Normalize, aggregate, and generalize data)
- **Data Reduction** (Obtain reduced data representation preserving integrity)

Four key data pre-processing phases required prior to data mining.

Data Cleaning: Handling Missing Values

Real-world data often lacks attribute values due to collection errors or system failures.

- Ignore the Tuple
  - Used when class label is missing; ineffective if many attributes are missing.
- Fill Manually
  - Time-consuming and infeasible for very large datasets.
- Use Global Constant
  - Replaces missing value with a label like 'Unknown' or ∞.
- Use Attribute Mean or Median
  - Imputes average value across all tuples or tuples of the same class.
- Most Probable Value
  - Uses decision trees, Bayesian inference, or linear regression to predict values.

Data Cleaning: Handling Noisy Data

Noise is a random error or variance in a measured variable, requiring smoothing techniques.

- Binning Methods
  - Sorts data and partitions into bins, smoothing by bin means, medians, or boundaries.
- Regression Analysis
  - Conforms data to a linear or non-linear mathematical function to smooth values.
- Clustering
  - Detects and removes outliers by grouping similar values together into clusters.
- Computer & Human Inspection
  - Combines automated rule checking with visual inspection to catch inconsistencies.

Data Smoothing by Binning Methods

Given sorted data: 4, 8, 15, 21, 21, 24, 25, 28, 34. Smooth the data using equal-frequency binning with bin size 3.

1. Step 1: Partition data into 3 equal-depth bins.
2. Bin 1: [4, 8, 15], Bin 2: [21, 21, 24], Bin 3: [25, 28, 34]
3. Step 2: Smooth by Bin Means.
4. Mean 1 = 9; Mean 2 = 22; Mean 3 = 29.
5. Result: Bin 1 = [9, 9, 9], Bin 2 = [22, 22, 22], Bin 3 = [29, 29, 29]
6. Step 3: Smooth by Bin Boundaries (closest boundary).
7. Result: Bin 1 = [4, 4, 15], Bin 2 = [21, 21, 24], Bin 3 = [25, 25, 34]

Data smoothed by bin means gives [9,9,9, 22,22,22, 29,29,29] and by boundaries gives [4,4,15, 21,21,24, 25,25,34].

Data Integration & Inconsistent Data

Integration combines data from multiple heterogeneous sources into a unified structure.

- Entity Identification Problem
  - Matches real-world entities across databases (e.g., Cust_ID vs Customer_Num).
- Redundancy Handling
  - Redundant attributes cause inconsistency; detected using correlation analysis.
- Chi-Square Test (χ^2)
  - Evaluates independence for categorical data: $\chi^2 = \sum \frac{(O-E)^2}{E}$.
- Pearson Correlation
  - Measures linear relationship $r_{A,B}$ between two continuous numerical attributes.
- Tuple Duplication
  - Identifies duplicated records resulting from inconsistent data entry.

Data Transformation Strategies

Converts data into forms appropriate for mining tasks through specific strategies.

- Smoothing
  - Removes noise using binning, regression, or clustering techniques.
- Aggregation
  - Applies summarization operations to data (e.g., daily sales to monthly totals).
- Generalization
  - Replaces low-level raw data with higher-level concepts (e.g., age to youth/adult).
- Normalization
  - Scales attribute values to fall within a specified range like $[0.0, 1.0]$.
- Attribute Construction
  - Creates new attributes from existing ones to aid mining efficiency.

Data Normalization Worked Example

Given attribute income with range [$12,000, $98,000]. Normalize value $73,600 to range [0.0, 1.0] using Min-Max Normalization.

1. Min-Max Normalization Formula: $v' = \frac{v - \min_A}{\max_A - \min_A}(\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$
2. Given parameters: $v = 73600$, $\min_A = 12000$, $\max_A = 98000$
3. Target boundaries: $\text{new_min}_A = 0.0$, $\text{new_max}_A = 1.0$
4. Calculation: $v' = \frac{73600 - 12000}{98000 - 12000}(1.0 - 0.0) + 0.0$
5. Result computation: $v' = \frac{61600}{86000} = 0.7163$

The normalized value of $73,600 using Min-Max normalization in the range [0.0, 1.0] is 0.7163.

Data Reduction: Data Cube Aggregation

Obtains a reduced dataset representation while preserving analytical completeness.

- Data Cube Concept
  - Stores multidimensional aggregated information at multiple levels of abstraction.
- Cuboid Levels
  - Base cuboid contains lowest-level details; apex cuboid contains highest summary.
- Aggregation Operations
  - Applies rollup operations to aggregate data, reducing total data volume.
- Query Efficiency
  - Precomputes frequent query results, accelerating pattern extraction.

Data Reduction: Dimensionality Reduction

Reduces dataset size by removing redundant or irrelevant attributes.

- Wavelet Transforms
  - Transforms data vectors into sparse numeric coefficients in frequency domain.
- Principal Component Analysis (PCA)
  - Finds k orthogonal vectors (principal components) explaining maximum variance.
- Feature Subset Selection
  - Removes redundant features using forward selection or backward elimination.
- Feature Construction
  - Combines existing features into compact, higher-level attributes.

Numerosity Reduction & Data Compression

Replaces original raw data with smaller alternative representations.

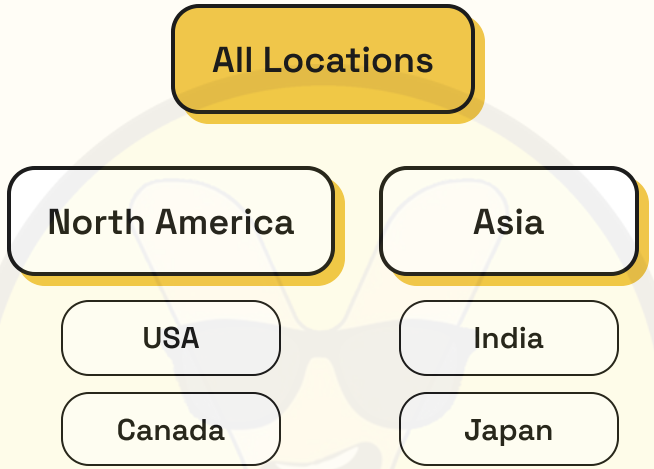
- Parametric Methods
  - Fits data to a model and stores parameters instead of raw data (e.g., Regression).
- Histograms
  - Partitions attributes into sub-intervals using equal-width or equal-frequency.
- Clustering & Sampling
  - Replaces data with cluster centroids or representative sample subsets.
- Lossless vs Lossy Compression
  - Lossless recovers original data exactly; lossy yields approximate data with high compression.

Discretization & Concept Hierarchy Generation

Reduces continuous numeric values by dividing their range into discrete intervals.

- Discretization Concept
  - Transforms continuous attributes into categorical attributes with interval labels.
- Binning & Histograms
  - Unsupervised methods used to automatically create numerical intervals.
- Cluster Analysis
  - Partitions continuous data into clusters, replacing values with cluster IDs.
- Hierarchy Generation
  - Maps low-level concepts to higher-level concepts automatically or manually.
- Categorical Hierarchies
  - Organizes discrete values into trees (e.g., Street → City → State → Country).

Concept Hierarchy Schema



Hierarchical mapping from low-level attribute values to high-level generalized concepts.

Decision Tree Induction: Basics

DECISION TREE

A Decision Tree is a flowchart-like structure where each internal node tests an attribute, each branch represents a test outcome, and each leaf node holds a class label.

- Supervised learning algorithm used for classification and regression tasks.
- Constructed in a top-down recursive divide-and-conquer manner.
- Easily translated into readable IF-THEN classification rules.

Attribute Selection Measures Comparison

Measure	Algorithm	Formula / Concept	Bias Property
Information Gain	ID3	$Gain(A) = Info(D) - Info_A(D)$	Biased towards attributes with many distinct values
Gain Ratio	C4.5	$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$	Overcomes Information Gain bias by normalizing split info
Gini Index	CART	$1 - \sum_{i=1}^{m} p_i^2$	Measures impurity; biased towards multi-valued attributes

Entropy & Information Gain Calculation

Dataset D has 14 tuples (9 Yes, 5 No). Attribute A splits D into D1 (4 Yes, 0 No) and D2 (5 Yes, 5 No). Calculate Entropy(D) and Info_A(D).

1. Step 1: Calculate Total Entropy $Info(D)$.
2. $Info(D) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$ bits.
3. Step 2: Calculate expected info needed after split: $Info_A(D)$.
4. $|D_1| = 4, |D_2| = 10$. $Info(D_1) = 0$ (pure).
5. $Info(D_2) = -\frac{5}{10}\log_2(\frac{5}{10}) - \frac{5}{10}\log_2(\frac{5}{10}) = 1.000$.
6. $Info_A(D) = \frac{4}{14}(0) + \frac{10}{14}(1.000) = 0.714$ bits.
7. Step 3: Calculate Information Gain: $Gain(A) = Info(D) - Info_A(D) = 0.940 - 0.714 = 0.226$ bits.

The Entropy of D is 0.940 bits and Information Gain for attribute A is 0.226 bits.

AKTU Semester Exam Questions

- **Q1** Explain the Knowledge Discovery in Databases (KDD) process with a neat block diagram.
- **Q2** What are the major data pre-processing steps? Why is data pre-processing necessary?
- **Q3** Explain equal-width and equal-frequency binning techniques with numerical examples.
- **Q4** Differentiate between Min-Max normalization and Z-score normalization with suitable mathematical equations.
- **Q5** Discuss dimensionality reduction techniques focusing on Principal Component Analysis (PCA).
- **Q6** Explain Decision Tree induction using ID3 algorithm. How is Information Gain computed?
- **Q7** Compare Information Gain, Gain Ratio, and Gini Index as attribute selection measures.

Unit 3 done. You are exam-ready.

Revise the Exam corner once more, then open Unit 4 on theycotes.com. Every AKTU subject, unit by unit, free. Notes by Chirag Jain, Ycotes.

