Data Mining – Unit 1 Viva Notes

Introduction to Data Mining

Data Mining (DM) means finding useful knowledge from large amounts of data. We are 'data rich but information poor' — data mining helps convert data ➜ knowledge ➜ action ➜ goal.

Motivation for Data Mining

Data is growing very fast, and manual analysis is impossible. Example: Netflix collects your ratings ➜ understands your taste ➜ recommends shows ➜ keeps you using Netflix.

What is Data Mining (Definition)

Process of automatically discovering useful information from large data repositories. Also called Knowledge Discovery from Databases (KDD).

KDD Process (Knowledge Discovery in Databases)

1. Selection – Choose relevant data from databases
2. Preprocessing – Remove noise or errors
3. Transformation – Convert data into suitable format
4. Data Mining – Apply algorithms to find patterns
5. Pattern Evaluation – Select only interesting patterns
6. Knowledge Presentation – Show patterns using graphs/charts

Data Mining — On What Kind of Data?

Relational databases (tables), Data warehouses (cleaned combined data), Transactional databases (shopping data), and Other data (web, maps, multimedia).

What Kinds of Patterns Can Be Mined?

Descriptive – describe general properties (trends, clusters). Predictive – predict future values (sales prediction).

Tasks: Characterization & Discrimination, Frequent patterns, Association rules, Correlation, Classification, Regression, Clustering, Outlier detection.

Are All Patterns Interesting?

Not all patterns are useful. Objective measures use maths (support, confidence). Subjective measures depend on user/domain knowledge.

Technologies Used in Data Mining

Statistics (models), Machine Learning (supervised, unsupervised), Databases (large data), and Information Retrieval (search systems).

Applications of Data Mining

Business Intelligence – customer behavior, market analysis, competitor study. Web Search Engines – fast search results using mining on huge data.

Data Mining Issues

1. Mining Methodology – handle noise, multi-dimensional data
2. User Interaction – easy interfaces, use user’s knowledge
3. Efficiency & Scalability – must work fast on big data

4 Diversity of Databases – handle text, images, networks
5 Data Mining & Society – privacy, misuse, invisible mining

Attributes and Types

Attribute = property of an object (like name, age).

Quantitative (measurable): Discrete (countable), Continuous (real values).

Qualitative (descriptive): Nominal (names), Ordinal (ranked), Binary (yes/no - Symmetric/Asymmetric).

Extra: Interval (no true zero, like temperature), Ratio (true zero, like age).

Measures of Central Tendency and Spread

Mean – average; Median – middle value; Mode – most frequent; Range – max-min; Standard Deviation – how spread out values are.

Symmetric vs Skewed Data

Symmetric: Mean ≈ Median ≈ Mode. Positively Skewed: Mean > Median > Mode. Negatively Skewed: Mean < Median < Mode.

Quantiles and Five-Number Summary

Q1 (25%), Q2 (50%/Median), Q3 (75%). IQR = Q3 - Q1. Five-number summary = Min, Q1, Median, Q3, Max. Shown with Boxplot.

Data Matrix vs Dissimilarity Matrix

Data Matrix: Rows=objects, Columns=attributes. Dissimilarity Matrix: shows distance between pairs of objects.

Dissimilarity of Numeric Data

Ways: Euclidean distance, Manhattan distance, Minkowski distance, Supremum distance.