

## UNIT III

### 1. Define Datamining?

It refers to extracting or “mining” knowledge from large amount of data. Datamining is a process of discovering interesting knowledge from large amounts of data stored either, in database, data warehouse, or other information repositories.

### 2. Give some alternative terms for datamining.

- Knowledge mining
- Knowledge extraction
- Data/pattern analysis.
- Data Archaeology
- Data dredging

### 3. What is KDD?

KDD- Knowledge Discovery in Databases.

### 4. What are the steps involved in KDD process?

- Data cleaning
- Data Mining
- Pattern Evaluation
- Knowledge Presentation
- Data Integration • Data Selection
- Data Transformation

### 5. What is the use of the knowledge base?

Knowledge base is domain knowledge that is used to guide search or evaluate the interestingness of result in pattern. Such knowledge can include concept hierarchies used to organize attribute / attribute values in to different levels of abstraction.

### 6. What is the purpose of Datamining Technique?

It provides a way to use various data mining tasks.

### 7. Define Predictive model?

It is used to predict the values of data by making use of known results from a different set of sample data.

**8. Define descriptive model?**

It is used to determine the patterns and relationships in a sampled data. Data mining tasks that belong to descriptive model:

- Clustering
- Summarization
- Association rules
- Sequence discovery

**9. Define the term summarization?**

The summarization of a large chunk of data contained in a web page or document. Summarization = characterization = generalization

**10. List out the advanced database systems?**

- Extended-relational databases
- Object-oriented databases
- Deductive databases
- Spatial databases
- Temporal databases
- Multimedia databases
- Active databases • Scientific databases
- Knowledge databases

**11. Define cluster analysis?**

Cluster analysis is data objects without consulting a known class label. The class labels are not present in the training data simply because they are not known to begin with.

**12. Describe challenges to data mining regarding data mining methodology and user interaction issues?**

- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages and ad hoc data mining
- Presentation and visualization of data mining results

- Handling noisy or incomplete data
- Pattern evaluation

**13. Describe challenges to data mining regarding performance issues?**

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, and incremental mining algorithms

**14. Describe issues relating to the diversity of database types?**

- Handling of relational and complex types of data
- Mining information from heterogeneous databases and global information systems

**15. What is meant by pattern?**

Pattern represents knowledge if it is easily understood by humans; valid on test data with some degree of certainty; and potentially useful, novel, or validates a hunch about which the user was curious. Measures of pattern interestingness, either objective or subjective, can be used to guide the discovery process.

**16. How is a data warehouse different from a database?**

Data warehouse is a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision-making. Database consists of a collection of interrelated data.

**17. Define Association Rule Mining.**

Association rule mining searches for interesting relationships among items in a given data set.

**18. When can we say the association rules are interesting?**

Association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Users or domain experts can set such thresholds.

**19. Define support and confidence in Association rule mining.**

Support  $S$  is the percentage of transactions in  $D$  that contain  $A \cup B$ .

Confidence  $c$  is the percentage of transactions in  $D$  containing  $A$  that also contain  $B$ . Support (

$$A \Rightarrow B = P(A \cup B)$$

$$\text{Confidence } (A \Rightarrow B) = P(B/A)$$

**20. How are association rules mined from large databases?**

I step: Find all frequent item sets:

II step: Generate strong association rules from frequent item sets

**21. Describe the different classifications of Association rule mining?**

- Based on types of values handled in the Rule
  - ✓ Boolean association rule
  - ✓ Quantitative association rule
- Based on the dimensions of data involved
  - ✓ Single dimensional association rule
  - ✓ Multidimensional association rule
- Based on the levels of abstraction involved
  - ✓ Multilevel association rule
  - ✓ Single level association rule
- Based on various extensions
  - ✓ Correlation analysis
  - ✓ Mining max patterns