Problem Solving

Assignment-1

Data Mining (CSE4052)

1 What is Data Mining? What are the different tasks of Data Mining?

Ans: Data mining refers to extracting or mining knowledge from large amounts of data. In other words, Data mining is the science, art, and technology of discovering large and complex bodies of data in order to discover useful patterns.

The following activities are carried out during data mining:

Classification
Clustering
Association Rule Discovery
Sequential Pattern Discovery
Regression
Deviation Detection

2 Discuss the Life cycle of Data Mining projects?

Ans The life cycle of Data mining projects:

Business understanding: Understanding projects objectives from a business perspective, data mining problem definition.
Data understanding: Initial data collection and understand it.
Data preparation: Constructing the final data set from raw data.
Modeling: Select and apply data modeling techniques.
Evaluation: Evaluate model, decide on further deployment.
Deployment: Create a report, carry out actions based on new insights.

3. Name areas of applications of data mining?

Ans Data Mining Applications for Finance
Healthcare
Telecommunication
Energy Industry.
Retail Industry.
E-commerce
Supermarkets
Crime Agencies
Businesses Benefit
Finanacial Analysis
Intrusion Detection.
Spatial Data Mining.

Biological Data Analysis.

5 What are the issues in data mining?

Ans A number of issues that need to be addressed by any serious data mining package

Uncertainty Handling
Dealing with Missing Values
Dealing with Noisy data
Efficiency of algorithms
Incorporating Domain Knowledge
Size and Complexity of Data
Data Selection
Understanding of Discovered Knowledge: Consistency between Data and Discovered Knowledge.

6 What is required, technological drivers in data mining?

Ans Database size: Basically, as for maintaining and processing the huge amount of data, we need powerful systems.

Query Complexity: Generally, to analyze the complex and large number of queries, we need a more powerful system.

7 What is data mining? Is it a simple transformation or application of technology developed from databases, statistics, machine learning, and pattern recognition?

Ans Data mining refers to the process or method that extracts or "mines" interesting knowledge or patterns from large amounts of data. No. Data mining is more than a simple transformation of technology developed from databases, statistics, and machine learning. Instead, data mining involves an integration, rather than a simple transformation, of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis.

8 How is a data warehouse different from a database? How are they similar?

Ans **Differences between a data warehouse and a database:**
A data warehouse is a repository of information collected from multiple sources, over a history of time, stored under a unified schema, and used for data analysis and decision support; whereas a database, is a collection of interrelated data that represents the current status of the stored data. There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another. A database system supports ad-hoc query and on-line transaction processing.

Similarities between a data warehouse and a database:
Both are repositories of information, storing huge amounts of persistent data.

9 Present an example where data mining is crucial to the success of a business.

What data mining functionalities does this business need ?

A department store, for example, can use data mining to assist with its target marketing mail campaign. Using data mining functions such as association, the store can use the mined strong association rules to determine which products bought by one group of customers are likely to lead to the buying of certain other products. With this information, the store can then mail marketing materials only to those kinds of customers who exhibit a high likelihood of purchasing additional products.

Thus to improve the business process, the concept of frequent pattern mining techniques (data mining technique) can be applied to extract the interested pattern for a particular objective.

10 Is discrimination a classification task?

Ans Discriminant analysis is a classification method. It assumes that different classes generate data based on different Gaussian distributions.

11 Write the difference between prediction and classification in machine learning with example

Ans We can think of prediction as predicting the correct treatment for a particular disease for an individual person. Eg. Whereas the grouping of patients based on their medical records can be considered classification.

12 Is noise same as outlier ?

Ans An outlier is a data point which is different from the remaining data [1]. Outliers are also referred to as abnormalities, discordants, deviants and anomalies [2]. Whereas noise can be defined as mislabeled examples (class noise) or errors in the values of attributes (attribute noise), outlier is a broader concept that includes not only errors but also discordant data that may arise from the natural variation within the population or process. As such, outliers often contain interesting and useful information about the underlying system. These particularities have been exploited in fraud control, intrusion detection systems, web robot detection, weather forecasting, law enforcement and medical diagnosis [1], using in general methods of supervised outlier detection

13 What do you understand by Data Purging?

Ans Data Purging is a process that is used in database management systems to maintain relevant data in a database. It is used to clean the junk data by eliminating or deleting the row and columns' unnecessary NULL values. It is essential because whenever we need to load new data in the database, we have to purge the irrelevant data from the database.

14 What are the different problems that "Data Mining" can solve?

Ans Data Mining can solve the following types of problems:

- Data Mining is mainly used to analyze data and make faster business

decisions, increasing revenue with lower costs.

- Data Mining also helps to understand, explore and identify patterns of data.
- Data Mining is used to automate the process of finding predictive information in large databases.
- It is used to identify previously hidden patterns.