Homework No. 1 - Introduction & Data
CSCE 5380.401 (5516)

Uday Bhaskar Valapadasu
11696364

Q1. What is data mining? In your answer, address the following:

Ans: According to my understanding, data mining is the process of identifying hidden patterns, relationships, and insights in massive datasets. It analyses and explores data using statistical methodologies, machine learning algorithms, and database management tools to derive meaningful and actionable knowledge. The purpose of data mining is to find previously undiscovered and possibly useful information hidden inside huge volumes of data. This extracted knowledge can subsequently be used to guide and support decision-making processes in a variety of areas and applications.

(a) Is it another discipline?

Ans:

No, Data mining is not a distinct discipline. It is an interdisciplinary area that combines and applies techniques from databases, statistics, machine learning, and artificial intelligence to analyze massive datasets and derive useful insights.

(b) Is it a simple transformation of technology developed from databases, statistics, and machine learning? (3 points**)**

Ans:

No. Data mining is a complex technology that combines databases, statistics, and machine learning. Data mining integrates techniques from various disciplines, including database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis.

(c) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

Ans:
The steps involved in data mining, when regarded as a knowledge discovery process, are as follows:

1. **Data Cleaning:** Remove noise and inconsistencies from the data. Handle missing values and correct errors.
2. **Data Integration:** Combine data from multiple sources into a coherent data store. Ensure consistency and resolve any data conflicts.
3. **Data Selection:** Select relevant data for analysis from the larger dataset. Identify the necessary attributes and dimensions.
4. **Data Transformation:** Transform data into appropriate forms for mining. Normalize, aggregate, or summarize data as needed.
5. **Data Mining:** To extract patterns from altered data, use intelligent techniques such as regression, association rules, clustering, classification, and anomaly detection.
6. **Pattern Evaluation:** Make use of validation techniques and interestingness metrics to find genuinely helpful patterns.
7. **Knowledge Presentation:** Use dashboards, reports, and visualizations to communicate mined knowledge in an intelligible manner.

Q2. For each data set given below, give specific examples of classification,

clustering, association rule mining and anomaly detection tasks that can be performed on the data. For each task, state how the data matrix should be constructed (i.e., specify the rows and columns of the matrix).

(a) Ambulatory Medical Care data, which contains the demographic and medical visit information for each patient (e.g., gender, age, duration of visit, physician's diagnosis, symptoms, medication, etc). (4 points)

Ans:

DM Task: Classification of Patients Question:

What kind of illness or medical condition is the patient most likely to have?

Row: Individual patient visits Column:

Characteristics such as diagnosis, age, gender, medical history, symptoms, etc.

DM Task: Clustering Question:

How many individuals with comparable illnesses or reactions to treatment be grouped together?

Row: Individual patient visits Column:

Characteristics such as age, gender, illness history, diagnosis, course of therapy, etc.

DM Task: Association Rule Mining Question:

Which illnesses or symptoms are usually linked to certain diagnoses or prescription drugs?

Row: Individual patient visits Column:

Binary indicators for presence/absence of symptoms, diagnoses, medications

DM Task: Anomaly Detection Question:

Binary markers indicating whether symptoms, diagnoses, or prescriptions are present or absent

Row: Individual patient visits Column:

Are there any odd or abnormal patient situations that might point to an inaccuracy in the data or an uncommon condition?

(b) Stock market data, which include the prices and volumes of various stocks on different trading days.

Ans:

DM Task: Classification Question:

Can we predict if a stock will be a top gainer, top loser, or remain neutral based on its historical data?

Row: Individual trading days Column:

Features like open, high, low, close prices, trading volume, moving averages, etc.

DM Task: Clustering Question:

How can we group stocks from similar sectors or industries based on their price and volume patterns?

Row: Individual stocks Column:

Features like daily price changes, trading volume, volatility measures, etc.

DM Task: Association Rule Mining Question:

Are there any strong relationships between the price movements of stocks from different sectors or asset classes?

Row: Individual trading days Column:

Binary indicators for price increase/decrease of different stocks or sectors

DM Task: Anomaly Detection Question:

Can we identify any abnormal trading activities or price movements that could indicate insider trading or market manipulation?

Row: Individual stocks Column:

Features like price changes, trading volume, news sentiment, social media mentions, etc.

Q3. Suppose your task as a software engineer at University of North Texas is to design a data mining system to examine their university course database, which contains the following information: the name, address, and status (e.g., undergraduate or graduate) of each student, the courses taken, and the cumulative grade point average (GPA). Describe the architecture you would choose. What is the purpose of each component of this architecture?

Ans:

As a software engineer at the University of North Texas, I would design a data mining system using a layered architecture to efficiently process, analyze, and discover knowledge from the university course database. This database includes student names, addresses, status (e.g., undergraduate or graduate), courses taken, and cumulative GPAs. The architecture consists of the following components:

1. Data Sources:

- Purpose: Provide raw data.

- Description: University course database containing student and course information.

2. Data Warehouse:

- Purpose: Store integrated data.

- Description: Central repository for structured data optimized for querying and analysis.

3. ETL (Extract, Transform, Load) Process:

- Purpose: Ensure data consistency and quality.

- Description: Extracts data, applies necessary transformations, and loads data into the data warehouse.

4. Data Preprocessing:

- Purpose: Prepare data for mining.

- Description: Performs cleaning, integration, selection, and transformation of data.

5. Data Mining Engine:

- Purpose: Discover patterns and insights.

- Description: Applies algorithms and techniques such as classification, clustering, and association rule mining.

6. Pattern Evaluation:

- Purpose: Validate discovered patterns.

- Description: Uses predefined measures and validation techniques to ensure relevance and accuracy.

7. Knowledge Base:

- Purpose: Store validated knowledge.

- Description: Repository for storing patterns and insights for decision-making and further analysis.

8. User Interface and Visualization:

- Purpose: Facilitate user interaction and present insights.

- Description: Provides tools for querying the system, accessing reports, and visualizing patterns and insights.

This architecture ensures that critical information is effectively gleaned from the university course database, promoting informed decision-making and enhancing understanding of student and course-related patterns.

Q4. For each attribute given, classify its type as:

- **discrete or continuous AND**
- **qualitative or quantitative AND**
- **nominal, ordinal, interval, or ratio**

Indicate your reasoning if you think there may be some ambiguity in some cases

Example: Age in years.

Ans:

1. **Average number of hours a user spent on the Internet in a week:**

   Continuous AND Quantitative AND Ratio

- The number of hours is a continuous value that can take on any real number, it's quantitative as it represents a measurable quantity, and it has a true zero point (ratio scale).

2. GPA of a student:

Continuous AND Quantitative AND Interval

- GPA is a continuous value that can take on any real number within a specified range, it's quantitative as it represents a measurable quantity, and it doesn't have a true zero point (interval scale).

3. Credit card number:

Discrete AND Qualitative AND Nominal

- Credit card numbers are discrete as they are distinct values, qualitative as they don't represent a measurable quantity, and nominal as they don't have an inherent order.

4. Salary above the median salary of all employees in an organization:

Continuous AND Quantitative AND Interval

- Salary is a continuous value that can take on any real number, it's quantitative as it represents a measurable quantity, and it doesn't have a true zero point (interval scale). Being above or below the median doesn't change its data type.

5. Number of students enrolled in a class:

Discrete AND Quantitative AND Ratio

- The number of students is a discrete value that can only take on whole numbers, it's quantitative as it represents a measurable quantity, and it has a true zero point (ratio scale).

6. Daily user traffic volume at YouTube.com (number of daily visitors):

Discrete AND Quantitative AND Ratio

- The number of daily visitors is a discrete value that can only take on whole numbers, it's quantitative as it represents a measurable quantity, and it has a true zero point (ratio scale).

7. IP address of a machine:

Discrete AND Qualitative AND Nominal

- IP addresses are discrete as they are distinct values composed of integers, qualitative as they don't represent a measurable quantity, and nominal as they don't have an inherent order.

8. Number of days since Jan 1, 2011:

Discrete AND Quantitative AND Ratio

- The number of days is a discrete value that can only take on whole numbers, it's quantitative as it represents a measurable quantity, and it has a true zero point (ratio scale).

Q5. Null values in data records may refer to missing or inapplicable values. Consider the following table of employees for a hypothetical organization:

Name	**Sales commission**	**Occupation**
John	**5000**	**Sales**
Mary	**1000**	**Sales**
Bob	**null**	**Non-sales**
Lisa	**null**	**Non-sales**

The null values in the table refer to <u>inapplicable values</u> since sales commission are calculated for sales employees only. Suppose we are interested to calculate the <u>similarity</u> between users based on their sales commission.

(a) Explain what is <u>the limitation</u> of the approach to compute similarity if we replace the null values in sales commission by 0. (2 points)

Ans:

Replacing null values with 0 in the sales commission column for similarity calculations can lead to a significant drawback. This approach assumes that employees with missing or irrelevant sales commission values have earned no commission, which may not accurately represent the true nature of the data.

For example, in a dataset containing developers, project managers, and sales representatives, replacing null values with 0 for developers and project managers in the "Sales Bonus" column would incorrectly suggest they have a bonus structure similar to low-performing sales representatives who actually received no bonus. This bias can distort similarity calculations and lead to misleading conclusions. It's essential to handle null values in a way that accurately reflects the data's context, such as excluding the sales commission column for non-sales employees or using a separate metric that considers different roles and compensation structures.

(b) Explain what is <u>the limitation</u> of the approach to compute similarity if we replace the null values in sales commission by the average value of sales commission (i.e., 3000). (2 points)

Ans:

Consider a dataset of real estate agents with a column for "Commission Earned." Replacing null values with the average commission of $10,000 would incorrectly suggest that agents with missing values have earned commissions similar to the average performer. This can distort similarity calculations and make it difficult to identify top and bottom performers accurately.

Using the average value to replace null values in the "Commission Earned" column can lead to inaccurate similarity assessments. It assumes that agents with missing values have earned commissions close to the average, which may not reflect reality. This approach can mask the true variability in commissions and hinder the ability to distinguish between high and low performers effectively.

(c) Propose a method that can handle null values in the sales commission so that employees that have the same occupation are closer to each other than to employees that have different occupations.

Ans:

To manage null values in the sales commission column while ensuring that employees within the same occupation are more closely aligned than those in different ones, we can implement a weighted similarity metric that separately evaluates occupation and sales commission:

1. Occupation Similarity: Apply a binary measure for occupation (1 if the same, 0 if different).

2. Normalized Sales Commission: Adjust sales commission values only for sales roles.

3. Weighted Similarity Score: Integrate the occupation similarity and adjusted sales commission similarity using specific weights.

Detailed Procedures:
- Occupation Similarity: Give a value of 1 for matching occupations and 0 for non-matching occupations.
- Normalised Sales Commission: When converting commission figures for sales personnel, use standard values (e.g., John's 5000 becomes 1.0, Mary's 1000 becomes 0.2).
- For non-sales roles, leave out the sales commission computations.
- Weighted Similarity: The occupation match and the normalised sales commission are added together to determine the overall similarity score.

This approach prioritises occupational alignment in similarity assessments, with sales commission coming in second.

Q6. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the Euclidean distance between the two objects. (3 points)

(b) Compute the Manhattan distance between the two objects.

(c) Compute the Minkowski distance between the two objects, using p = 3.

Ans:

To compute the Euclidean distance, Manhattan distance, and Minkowski distance between the two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8), we can use the following formulas:

(a) Euclidean distance:

Let's denote the two objects as (x1, y1, z1, w1) and (x2, y2, z2, w2). The Euclidean distance is calculated as:

sqrt((x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2 + (w1 - w2)^2)

Given: (x1, y1, z1, w1) = (22, 1, 42, 10) (x2, y2, z2, w2) = (20, 0, 36, 8)

Euclidean distance

= sqrt((22 - 20)^2 + (1 - 0)^2 + (42 - 36)^2 + (10 - 8)^2) = sqrt(4 + 1 + 36 + 4)

= sqrt(45) ≈ 6.7082

(b) Manhattan distance: The Manhattan distance is calculated as:

|x1 - x2| + |y1 - y2| + |z1 - z2| + |w1 - w2|

Manhattan distance = |22 - 20| + |1 - 0| + |42 - 36| + |10 - 8|

= 2 + 1 + 6 + 2

= 11

(c) Minkowski distance (with p = 3):

The Minkowski distance is a generalization of Euclidean and Manhattan distances. It is calculated as:

$(|x1 - x2|^p + |y1 - y2|^p + |z1 - z2|^p + |w1 - w2|^p)^{(1/p)}$

Minkowski distance (p=3)

$= (|22 - 20|^3 + |1 - 0|^3 + |42 - 36|^3 + |10 - 8|^3)^{(1/3)}$

$= (8 + 1 + 216 + 8)^{(1/3)}$

$= 233^{(1/3)} \approx 6.1962$

Therefore, the distances between the two objects are:

(a) Euclidean distance ≈ 6.7082

(b) Manhattan distance = 11

(c) Minkowski distance (with p=3) ≈ 6.1962