Data Mining
Assignment 1

September 12, 2025

Viet Huy Duong
811433146

---

1. Based on your understanding, please explain what data mining is. (10 pts)

Ans: Data mining is the process of discovering interesting patterns and knowledge from large volumes of data, often using techniques like classification, clustering, and association rule mining.

2. Please list 3 real applications for data mining. Explain the data mining problems and solutions in these applications (give citations in reference format, if you use any online resources) (30 pts)

Ans: Data mining applications in retail and marketing: For example, based on shopping carts, we can analyze and segment customers [1]. However, to fully exploit data for these tasks, retailers often have to deal with a variety of data types and real-time data processing. Solution: we can use a lighter model for faster prediction. Retailers can also integrate LLMs to handle varied data types.

Application in finance and banking: Fraud detection and risk analysis[2] [3] . Problem: economic data usually extremely sensitive, especially with the upcoming data, which often very extraordinary. Developers need to be careful in accepting new data, balancing the frequency of updating the model while still ensuring the preservation of previously learned knowledge. In addition, there needs to be a specific method to handle string data, such as stock prices and market supply and demand.

Application in bio informatics: Genetic data analysis and drug discovery [4] [5]. The main difficulty in mining data for these problems is that it requires a large amount of specialized knowledge from leading experts. The data fields in these problems are often relatively complex and difficult to understand for technology engineers. The most effective solution is to collaborate with experts in the field and doctors to provide more knowledge and exploit the meaning of the data fields.

3. Please discuss whether or not the following problems are data mining tasks. Explain why. (30 pts)
   (a) Retrieve students' records from a relational table with grade = "A".
   (b) From the table of students' information, check if attributes last name and address have any correlations.
   (c) Find all the documents from the text database containing keywords "data mining".
   (d) Divide the text database into several groups, each group containing near duplicate or similar documents.
   (e) Based on historical stock data, as well as other attributes (e.g., gold price, gas price, etc.) for the past few days, predict the trend of a stock tomorrow.
   (f) Please provide your own example of data mining.

Ans:

3(a). No, this is not a data mining task.This problem is about database management and querying, there is no operation to mine hidden patterns in data.

3(b). Yes, this is a data mining task. This task will attempt to find hidden relationships between last names and addresses, potentially uncovering more relationships and information than what is initially visible in the dataset.

3(c). No, this is not a data mining task. This task is simply data retrieval.

3(d) Yes, this is a data mining task. Dividing documents into groups based on similarity or near-duplicates may require clustering techniques or higher-level natural language processing techniques.

3(e) Yes, this is a data mining task. This is also known as time series problem, exploiting time series data to predict based on hidden characteristics of past and present relationships.

3(f) The example I chose is intrusion detection system (IDS), specifically detecting network attacks based on bandwidth data and other collected features. Ids problems focus on finding and exploiting hidden features from pre-labeled data patterns and then predicting on newly encountered data patterns.

4. Please choose a specific data mining problem (e.g., classification, clustering, regression, etc.), and discuss the corresponding solutions to this data mining problem. You can search for these problems via Google from the Internet (e.g., Wikipedia, Web pages, research papers, etc.), and explain the problem definition and solutions. (30 pts)

Ans: I choose classification to present in this question. The main task of classification is try to classify data samples based on available labels, unlike regressor, classification does not try to predict a certain value, instead, classification predicts which category this data sample belongs to. Classification is often supervisor learning problems.

There are many solutions for classification problems. Usually, we will decide which algorithm to use based on the factors of the problem, such as the size of the data set, the characteristics of the data set: string, image, audio, video, ... computing resources, memory limits, prediction time, ... Algorithms can be mentioned as:

- Decision Trees: The algorithm is built on the idea of a tree structure, where each node represents a decision based on a feature, and the leaf nodes represent class labels. Decision trees can handle numerical or categorical features well, but are prone to overfit. A more advanced variant of Decision trees is random forests.
- Support Vectore Machines (SVM): SVM finds a hyperplane that separates classes in a high dimensional space. Try to maximizing the margin between support vectors (which represent for the nearest point to the hyperplane
- Multi layer perception (MLP): This can be considered the simplest version of a deep learning neural network, MLP can solve the classification problem well with the number of nodes in the output layer equal to the number of categories. However, training MLP or some deep learning neural networks can be more complicated than the above methods, and requires special techniques to prevent overfitting.

In summary, the algorithms I mentioned above are just some of the algorithms applied to classification problems. In addition, some metrics can be used to evaluate the effectiveness of a classification problem, such as: accuracy, precision, F1 score, Recall,...

References

[1] A. Griva, C. Bardaki, K. Pramatari, D. A. Papakiriakopoulos, Retail business analytics: Customer visit segmentation using market basket data, Expert Syst. Appl. 100 (2018) 1–16. doi:10.1016/j.eswa.2018.01.029.

[2] A. A. Lawal, E. Ezeife, J. O. Akande, A. Olapade, A. O. Olatunji, Data mining for

financial fraud detection: Techniques, case studies and challenges, Asian Journal of Mathematics and Computer Research (2025). doi:10.56557/ajomcor/2025/v32i29127.

[3] S. K. Hashemi, S. L. Mirtaheri, S. Greco, Fraud detection in banking data by machine learning techniques, IEEE Access 11 (2023) 3034–3043. doi:10.1109/ACCESS.2022.3232287.

[4] P. Thareja, R. S. Chhillar, A detailed survey on data mining based optimization schemes for bioinformatics applications, ECS Transactions (2022). doi:10.1149/10701.4689ecst.

[5] K. Lan, D. tong Wang, S. Fong, L. Liu, K. K. L. Wong, N. Dey, A survey of data mining and deep learning in bioinformatics, Journal of Medical Systems 42 (2018) 1–20. doi:10.1007/s10916-018-1003-9.