Computer Science and Networking, School of Computing and Data Science, Wentworth Institute of Technology

Data Mining

Dr. Salem Othman

Summer 2023

Assignment 1: Data Mining

In our recent lecture, we explored the fascinating field of data mining and its various applications. To assess your understanding of the key concepts we covered, I have prepared an assessment for you. Please read each question carefully and provide your responses based on the knowledge you have gained.

1. Define data mining in your own words and explain why it is important in today's world.
2. What are the main stages of the data mining process? Briefly explain each stage.
3. In the context of data mining, what is the difference between predictive and descriptive tasks? Provide examples of each.
4. Describe the concept of clustering and its significance in data mining. Give at least two real-world examples where clustering is applied.
5. Choose one of the applications discussed in the slides, either market segmentation or fraud detection, and explain the approach and goal of that application. Include the steps involved and the benefits it brings to businesses or organizations.
6. Explain the concept of regression in data mining and provide examples of its applications in different domains.
7. Explain the process of clustering and how it is applied in market segmentation.
8. Discuss the concept of document clustering. What approach is used to cluster similar documents together?
9. Define Association Rule Discovery in your own words and provide an example of its application in any industry of your choice.
10. A lung cancer dataset revealed a significant Subspace Differential Coexpression Pattern. This pattern was related to the TNF/NFB signaling pathway with a P-value of 1.4*10^-5. Explain what this means in the context of association analysis.
11. Define anomaly detection. How is it used in credit card fraud detection and network intrusion detection?
12. Discuss the following challenges in data mining: scalability, high dimensionality, heterogeneous and complex data, data ownership and distribution, non-traditional analysis. Provide a potential solution or strategy for each.
13. Based on the topics discussed, why is data mining important in today's data-driven world? Include specific examples from the course material in your answer.
14. Discuss whether or not each of the following activities is a data mining task.
    a. Dividing the customers of a company according to their gender.
    b. Dividing the customers of a company according to their profitability.
    c. Computing the total sales of a company.
    d. Sorting a student database based on student identification numbers.
    e. Predicting the outcomes of tossing a (fair) pair of dice.
    f. Predicting the future stock price of a company using historical records.
    g. Monitoring the heart rate of a patient for abnormalities.
    h. Monitoring seismic waves for earthquake activities.

i. Extracting the frequencies of a sound wave.

15. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.
16. For each of the following data sets, explain whether or not data privacy is an important issue.
    a. Census data collected from 1900–1950.
    b. IP addresses and visit times of web users who visit your website.
    c. Images from Earth-orbiting satellites.
    d. Names and addresses of people from the telephone book.
    e. Names and email addresses collected from the Web.

Please provide concise and accurate answers for each question. Take your time and ensure you demonstrate your understanding of the concepts we covered. Good luck!

Deliverables:

You should upload your file to Brightspace.