Your name:

B565 HW1 (Fall 2023)

Submission instructions

Submit a PDF file to canvas. Your answer should be typed up and you are welcome to use word, latex or anything that works for you. The PDF file includes your answers to Questions 1-5, results of your exploratory analysis of the given dataset (Q6), and paper summary (Q7). In addition, submit your code or notebook for Q6 to github.iu under HW1 folder in your B565 repository. To create B565 repository,

- Go to github.iu.edu.
- Find the organization named B565-Fall2023.
- Create a repository called B565 followed by your email ID (e.g., `B565yye`, where `yye` is Prof. Ye's email ID) under the organization. Make sure that you select "Private" and click on "Initialize the repository with a README."
- Make sure that you submit your code/notebook for HW1 to your B565 repository under HW1 folder; otherwise we won't be able to find your submission.

Questions

1. Transaction data (with each transaction represented as a set of items) can also be represented as a matrix (transaction-item matrix, similar to the document-term matrix we showed in the class). List one advantage and one disadvantage of using such a matrix. Also discuss why a transaction-item matrix is an example of a data set that has asymmetric discrete features. [10 pts]

2. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there many be some ambiguity. [10 pts]

   (a) Brightness as measured by a light meter.

   (b) Brightness as measured by people's judgments.

   (c) Number of customers in a grocery store.

   (d) Letter grades (A, B, C, D and F).

   (e) Distance from the Monroe County Courthouse.

3. Billy's advisor had a new algorithm for solving a problem (his goal is to solve the problem faster), and he asked Billy to test it out. Billy implemented the algorithm and ran it with 5 datasets on a computer. Comparing to a reported baseline performance, Billy's new program achieved the following speedups, 2.0 (i.e., 2 times

faster on dataset 1), 0.5 (2 times slower on dataset 2), 2.0, 1.0, and 10. Compute the average speedup of Billy's program over the five 5 datasets using three different means: arithmetic mean, geometric mean, and harmonic mean. How to calculate the different means? Given n numbers, the arithmetic mean is the sum of the numbers divided by n, the geometric mean is the nth root of the product of the numbers, and the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the given numbers. Show your calculation and the results. Which mean do you think makes the most sense for this case? And why? [10 pts]

4. Distinguish between noise and outliers. Be sure to consider the following questions. [10 pts]

   - Is noise ever interesting or desirable? Outliers?
   - Can noise objects be outliers?
   - Are noise objects always outliers?
   - Are outliers always noise objects?
   - Can noise make a typical value into an unusual one, or vice versa?

5. Learn about ChatGPT using Google Trends. Write a brief summary including four highlights about what you have learned. [20 pts]

6. EDA practice. You will be using this data set on Kaggle: TikTok popular songs 2019. What to check? Summary of the data set, attributes (types and distributions). Are there missing values? Are there correlations between the attributes? [20 pts]

7. Write a summary for this paper: Reducing Exposure to Harmful Content via Graph Rewiring (KDD 2023). It is a math-heavy research paper and it is perfectly fine if you don't understand some (or many) parts of the paper. Focus on the problem it tries to solve. [20 pts]

   - The length of your summary should be about one page.
   - State the main ideas: the problem that the paper tries to address, what data/example was used, and what was the method that was applied/developed to solve the problem.
   - Add your personal opinion. Do you like the paper or not? Why? How do you think about the paper?
   - Write the summary in your own words; don't copy and paste.