CSE 494: Artificial Intelligence for Cyber Security: Spring 2019

Homework 3: (20 points)

These question tests your understanding on using Apriori principle and association rule mining. We will be using a cybersecurity dataset, which is the DefCon CTF data traffic.

Your first task is to download the data from blackboard, which is the cvs file.

Answer the following questions.

1. The CVE file has 2 columns; one is the payload that was used in the attack and the other column is the instruction and their count (histogram) in this payload.
    a. The first step is to extract each instruction and it's count from each row. Think of it as a dictionary where you have a key (the instruction) and a key_value(the count). (3 points)
    b. Next, find out the highest 50 instruction that were used. (1 point).
2. We saw in the lab how we arranged market basket items into columns:
    a. In this step, arrange each instruction to be a column by itself. (5 points)
    b. Now for each payload, map each instruction to this payload. In other words, look for each instruction in each payload, if it exist in this payload, insert 1 in the cell of this payload and the instruction. (5 points).
3. Use mlxtend python package to call apriori library.
    a. Now you should have your data organized same way we did in the lab. Use apriori library to get the frequent item sets for support =0.5. (2 points).
        i. Note: since we have quite huge dataset, you might need to reduce the size of the dataframe. You may try to work with 100 rows, and 1000 column. If you still have an issue, minimize the number of rows and columns to be all rows (500) and the first 50 and last 50 columns.
    b. Generate association rule that have confidence of at least 40%. (2 points).
    c. Now increase the support threshold to 70% and report which frequent itemsets you got and the association rules for confidence =0.5. What do you observe about the results? (2 points)

BONUS QUESTION: (5 points).

What is the complexity of Apriori algorithm? Can you research and say what is the best complexity science achieved for association rule mining?