CS 6220 Data Mining — Assignment 4

Association Rules

To complete this assignment you will need to download the following resources:

1. You may find a notebook with potential helper functions for implementing the Apriori https://goo.gl/1HtnFQ. Add your portion of the assignment solution to the end of this file and submit it.

2. The dataset you'll be using grocery.csv can be found at https://www.dropbox.com/s/mz6j5glhnd3skfs/grocery.csv?d Download the file and place it in the same directory as your notebook.

Objectives:

1. Define association rules and state their usefulness

2. Programmatically apply association rules to the given dataset and analyze the results

Submission:

Through the assignment submission portal on Canvas, submit your ipynb with a pdf of your assignment solution; no need to zip the files. This is a norm for almost all assignments

Grading Criteria:
Follow the instructions in the pdf, and complete each task. You will be graded on the application of the modules' topics, the completeness of your answers to the questions in the assignment notebook, and the clarity of your writing and code.

What You Need to Do

Part 1 - Apriori [40 Points]:

The dataset above contains nearly 10 thousand transactions recorded from a grocery story. Each row in the dataset refers to a given transaction, where the items purchased are separated by commas. For example, on the second row we have a transaction with three items: tropical fruit, yogurt, and coffee. The attached notebook file (first download link above) contains a helper function that allows you to quickly load that file into a format that can be easily processed in Python.

Task 1: Your task here is to make use of the provided functions to generate candidate itemsets, select those that are frequent using Apriori, and subsequently list association rules derived from these.

[Note that because we have thousands of transactions, it may be hard to find itemsets with high supports (e.g., 20%), so in order to see interesting results, make sure you experiment with lower min support parameters. Make sure to document your code and leave some commentary on the results you obtained, which you will further discuss on the Collaborative Activity for this lesson.]

Task 2: We can find a relationship between the confidence level and number of rules found for a certain support value. For this, plot the number of rules found on y-axis and confidence levels on x-axis for different support values. Use 10%, 20%, 30%, 40%, 50% confidence levels for each of 2%, 3%, 4%, 5% support levels in the same figure. Plot a separate line for each support level.

Part 2 - FPgrowth [30 Points]:

Repeat the above process but this time use FP-growth. You may use the code provided at https://goo.gl/Rv8KAa, or some other Python implementation that you might find online (just be sure to cite your sources).

Part 3 - Interest Factor [30 Points]:

Use either Apriori or FPgrowth algorithm with 2% support and 30% confidence to generate the rules. Now, calculate interest factor for all the rules.

Prepare three sets of rules sorted in descending order by - support, confidence, and interest factor, respectively. Select and print the top-5 rules in each list. Compare and mention if any rules are common in those.