HOMEWORK 2 – ASSOCIATION RULES

(Total points 10)

Mining Association Rules from Movie Data

Business Context: Streaming platforms

Assume you work as a data scientist for a streaming platform (for example, Prime Video) and you are asked to analyze the dataset **movies_binary.csv** containing information about movies watched by different individuals. You are asked to mine the dataset to find interesting relationships about the movies that users watch, so to provide the platform with potentially, interesting insights.

Data Description

For the purposes of this assignment (and this assignment only), the dataset provided has been extracted and manipulated from a larger database of movie rating data collected by a recommendation service. It therefore includes only a limited number of movies transactions and results obtained may not be applicable to the entire, complete, dataset.

The data is in binary format, where each column represents one movie. Each row can be thought of as capturing the collection of movies watched by a given individual, where a value of 1 indicates that the individual watched the corresponding movie and a value of 0 indicates that the individual did not watch the corresponding movie.

Each movie's name contains the title (sometimes shortened) and the year of the movie. For example, one of the movies in the dataset is X.Men.2000.

Analysis

Perform the following analysis in R:

a) (0.4) Load the dataset movies_binary.csv, following the steps to convert a binary dataset into a transactions dataset. (Simply report the code used for this question)

b) (0.4) Report the total number and the names of the unique movies existing in the dataset (you can take a screenshot of the output obtained in R Studio and simply paste it).

c) (0.8) Plot the support percentage of the top 10 movies. Your plot should have a proper title, proper y-axis limits and be filled with a color of your choice. Which are the top 2 movies (based on support) and how do you interpret the support % obtained for them?

d) (0.4) Find how many transactions contain both "Up.2009." and "LionKing.The.1994."

e) (0.6) Implement the Apriori algorithm and find all the association rules with minsupp = 0.7 and minconf = 0.7 and containing a minimum of 2 items and also a max of 2 items (HINT: you can set the max using maxlen, which works as minlen). Get a descriptive summary of the rules, using summary(*rules_name*) and comment on the mean values of support, confidence and lift.

f) (0.8) Inspect the rules found by the algorithm in point e) and report them (you can take a screenshot of the output generated in R Studio). How many rules are generated? By simply looking at the names of the movies involved in the different rules (ignore support, confidence, lift), identify one rule that you do not find very surprising (or obvious) and one that, instead, you may find less obvious and explain why.

g) (0.6) Next, order the rules found in point e) in increasing order by confidence (paste in the HW the ordered table generated). Identify the rule (or rules) with the highest and lowest confidence: by only looking at the confidence, how do you interpret the results?

h) (0.6) Order the rules found in point e) in increasing order by lift (paste in the HW the ordered table generated). Identify the rule (or rules) with the highest and lowest lift ratio: by only looking at the lift, how do you interpret the results for these rules?

i) (1.2) Analyze again at the table produced in point g), and look at both confidence and lift for the different rules. Identify at least 2 rules that you would trust and at least 2 rules that, instead, you would not trust as much (note: saying that you trust them all is not a good answer. There are rules that are more trustworthy than others). You need to clearly explain your choices and explain how confidence and lift affect your reasoning.

j) (1.4) Based on (all) the analysis completed so far, you need to form suggestions to provide to the streaming company you (fictionally) work for. What type of suggestions/actions would you recommend to the platform, based on the results of your analysis? Note: Your answer does not need to be a page long; it can be concise (10 lines) but still show a clear understanding of how to interpret results from association rules. In your answer, make sure to use metrics such as confidence and lift discussed in the points above, and give specific recommendations.

Assume that, after presenting the results to your company, you are requested to show which association rules can be considered "bad" enough or "non-trustworthy", and therefore it would be better to not consider them.

k) (0.3) You decide to run again the Apriori algorithm using minsupp = 0.5 and minconf = 0.6 and minlen = 2. Copy and paste the code used and report only how many rules, in total, the algorithm found.

l) (0.5) Next, plot the rules found in point k) using a scatterplot, where your axis capture support and lift, and the color of the points in the plot changes based on the confidence measure.

m) (1.2) Look at the scatterplot produced in point l), and identify which rules (or group of rules) you would consider "bad enough" to exclude from consideration. Use the "Interactive" version of the scatterplot to get more details about the chosen rules. Inspect the chosen group of rules, report the details and clearly explain why you are selecting such rule(s) as candidates for exclusion. Note: while answers may vary to an extent, there are good answers and bad answers for this question. Your reasoning must make sense, and use the metrics learnt in this class to assess and interpret association rules.

n) (0.8) Finally, think about another context (it can be business or related to another field) where you could use association rules mining to find interesting relationships. Specify why association rules would apply to the chosen context and what type of data you would need to perform an analysis similar to the one completed for the movies. Try to be specific as you can – that is, think about the "ideal" data you could get, without thinking about whether it would be hard/feasible to get such data.

NOTE: you cannot use as context anything related to Groceries/Retail Products, nor Health, that are the examples provided in the videos. Movies are excluded as well.