Compsci 361 Assignment 3

Hasnain Cheena
190411106

Parts A & B

The dataset contains information related to grocery shopping. It contains 124 unique items and 4627 unequal length transactions. The longest length transaction in the dataset is 49 items.
The tool used in this report is the *mlxtend* library in Python (Python 3.8.2). The algorithm used within the library is *fpgrowth.* Pre-processing of the dataset (using the .csv format) was done prior to analysis. This pre-processing chain involved:

1. Using pandas to read in the csv data as a dataframe (need to specify that the number of columns in the dataframe is equal to the longest length transaction)
2. Remove all NaN values present in the dataframe and then convert the structure to a list of lists. In this nested list structure, each transaction is stored as a list which is then stored in a bigger list containing all the transactions.
3. Transform the nested list structure using a Transactional Encoder. The output of the Transaction Encoder is ready to be passed into the *fpgrowth* algorithm for association rule mining.

Part C

The final parameters selected were *minsup* of 0.1, *minconf* of 0.5 and a *minlift* of 2.2.

The minimum support parameter was selected systematically by iterating through a range of support levels of 0.1 to 0.7 at intervals of 0.1. 0.1 was chosen as the lower bound as the compute resources required at lower support levels was infeasible. 0.7 was chosen as the upper limit as the highest support of a single item is around 0.72. Furthermore, from Table 1 below you can see that as the minimum support value increases the run time decreases. Therefore, I selected the *minsup* parameter value to be 0.1. This is because the runtime is acceptable, and a wide number of rules are generated that will be filtered by using interestingness measures.

The minimum lift parameter was selected by iterating through lift levels of 1 to 2.6 at steps of 0.2 (keeping *minsup* at 0.1). 1 was chosen as the lower bound to filter out all negatively correlated rules. 2.6 was selected as the upper bound as at/after 2.6 all possible rules have been filtered out. From Table 2 you can see that as the minimum lift value increases the number of rules generated decreases. Therefore, the *minlift* parameter value was selected to be 2.2. This value is high enough to generate high quality rules and low enough to produce a sufficient set of rules.

Then the rules were filtered by confidence to only return rules with confidence of greater than 0.5. This was to ensure only reliable and reasonably reoccurring rules were assessed.

To calculate timing characteristics of the *fpgrowth* algorithm 100 runs of the algorithm were performed using the *timeit* functionality in Python. The results were then averaged. Using parameters *minsup* of 0.1, *minlift* of 2.2 and *minconf* of 0.5 the runtime of the algorithm is around 2.4 to 2.5 seconds.

Table 1: Minimum Support

Minimum Support	Run time
0.1	2.5 s
0.2	465 ms
0.3	187 ms
0.4	104 ms
0.5	53 ms
0.6	42 ms
0.7	33 ms

Table 2: Minimum lift

Minimum lift	Number of rules generated
1	181658
1.2	162012
1.4	92112
1.6	40258
1.8	18616
2	6162
2.2	882
2.4	66
2.6	0

Part D

Filtering using interestingness measures of lift and confidence (at the values specified in part C) results in a rule set of 420 rules within it. Lift is used to extract rules with high positive correlation and confidence is used to extract reliable rules. A few interesting rules are shown below:

Antecedents	**Consequents**	**Support**	**Confidence**	**Lift**
bread and cake, fruit, party snack foods, frozen foods	vegetables, total = high	0.118651	0.571875	2.083516
biscuits, prepared meals	frozen foods, total = high	0.102442	0.563615	2.048582
party snack foods, margarine, sauces-gravy-pkle	baking needs, total = high	0.101578	0.601023	2.172604

Rule 1 states that when a customer has fruit, bread and cake, party snacks and frozen foods they are also likely to purchase vegetables and have a high total cost.
Rule 2 states that when a customer has biscuits and prepared meals, they are likely to purchase frozen foods and have a high total cost.
Rule 3 states that when a customer has party snacks, margarine and sauce gravy they are likely to purchase baking goods and have a high total cost.