Grant Ikehara, Cameron Healy, Dominic Soares

CPSC 310

HW7

Descriptions of Processes and Results

The first thing we did was write functions to get the support, confidence, and lift for itemsets in a table. We tested the functions with single left hand side and right hand side itemsets to ensure they were accurate. We made the apriori algorithm to search through all possible rules and then test each of the resulting rules is against the parameters for minimum support and confidences that are explicitly made in the function call to determine if the rule meets the specified minimum confidence or support. The resulting set of rules can be manually confirmed to be valid and accepted rules by ensuring that all the parent, 2 itemset rules it determined are valid, and that each rule with more than 2 items in the itemset are children of valid rules and have valid support and confidence metrics themselves. Using this approach, all length itemsets that are supported are found in both the titanic and mushroom datasets.

The rules we found for the titanic data set made sense, as we found many different rules for the 2 and 3 itemset rules but comparatively not many for the 4 itemset rules, since there are only 4 attributes to select from, it is rare to have all four in an itemset. The values for confidence was predictably lower for the 2 itemset rules than the others, as there are many outcomes for only a one itemset left hand side, and the odds of it being the given one itemset right hand side is lower than if there were multiple items on the right hand side, as there is a lower change of multiple right hand sides. A similarity can be made to the rules in homework 4 and 5, since in those homeworks, the resulting prediction can be shown in an itemset similar to the result of the apriori algorithm.

When the minimum confidence or support was elevated to .5 for the titanic dataset, the amount of rules that were found was cut down drastically, as the initial 2 itemset rules that were below the threshold could not contribute to more rules. For the mushroom data set…

To modify for the mushroom data set, we

To format the information, we used tabulate for python, showing a unique id for the rule, left hand side, right hand side, support, confidence, and lift. For the mushroom data set we added all 23 characteristics to the tabulate for that data set and set of rules. Each characteristic is shown visually with the letter that it corresponds to.

When we did feature selection for the mushroom dataset, we found that using x features would make the rules … while using y features made the rules… Using all the features.

Test Results: