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, we had to use a much higher minimum support, due to the sheer size of the file.

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:

**Titanic**

* We had a grand total of 30 rules for the titanic dataset, all but two of which had a lift greater than one. At a minsup of .25 and a minconf of .75, it was worth noting that all of the rules dealt with female passengers/crew members. This seemed to imply that rules relating to male passengers/crew members were much less reliable, and likely did not meet the minimum confidence (it’s more likely they met the support requirements, as there were a good number of men on the titanic).
* For obvious reason, Rules encompassing identical itemsets (adult 🡺 female vs. female 🡺 adult) had identical supports, but generally different confidences. If the division between LHS and RHS occurred at the same point, Lift was identical too.
* Many rules display information identical to realizations that occurred while using the decision trees and random forests. For example, (yes, female, third 🡺 adult) because no female children in third class survived. Similarly, (yes, adult, third 🡺 female) because not many men (child or adult) survived in third class (“Women and Children First!”).
* Only one rule had 3 elements in the RHS, which was (third 🡺 female, adult, yes). This had a confidence of .75, so it stands to reason that 75% of the passengers in third class were adult females who survived, if I understand this correctly.
* As confidence cannot increase with items added to the RHS, all subsets of the RHS (female, adult, yes) with identical LHS (third) also are present in the ruleset.

**Mushroom:**

* This time, we had 136 rules if no feature selection was done and both the minsup and minconf were set to .75. (understandably, this meant that many rules had Left Hand Sides that were identical in length to the complete table (i.e. an attribute had a universal value.
* There were 3 rules that had a LHS with 5 elements, the RHS for each of those was a ringNumber of “O” (One ring)
* Similarly, there were 3 rules that had a RHS with 5 elements, which I hypothesize are the same as the LHSs mentioned above, as the LHS was uniformly ringNumber = “O” again.
* Quite unfortunately, no rule had a lift greater than 1 in this set, likely due to our extremely high minsup.
* A grand total of 17 rules had 100% confidence; some of these had as many as 3 different attributes in the LHS and as few as 1.
* One rule with 3 items in the RHS also had 100% confidence.