Homework #3: Association Rules

Taylor Graham
CSCI 4502: Data Mining

February 10, 2015

Honor Code Pledge: On my honor, as a University of Colorado at Boulder student, I have neither given nor received unauthorized assistance on this work.

1a) $lift(ski \Rightarrow football)$ is the same as $\frac{supp(ski \Rightarrow football)}{supp(ski)*supp(football)}$. Note:

$$\begin{aligned} supp(ski \Rightarrow football) &= 0.375 \\ supp(ski) &= 0.625 \\ supp(football) &= 0.5 \\ lift(ski \Rightarrow football) = \frac{0.375}{0.625 * 0.5} &= 1.2 \end{aligned}$$

1b) As stated earlier, $supp(ski \Rightarrow football) = 0.375)$. $conf(ski \Rightarrow football)$ is the ratio of the number of people that participate in both skiing and football, to the number of people who ski in general.

$$conf(ski \Rightarrow football) = \frac{\|ski \bigcap football\|}{\|ski\|} = \frac{1500}{2500} = 0.6$$

Using $min\_supp = 0.25$ and $min\_conf = 0.5$, we can conclude:

$$supp(ski \Rightarrow football) > min\_supp$$

$$conf(ski \Rightarrow football) > min\_conf$$

Therefore the association rule can be considered strong.

2a) In this problem, for an itemset to be considered frequent, it must be in at least 3 out of 5 transactions. I know that there are only 6 frequent 1-item itemsets: B, E, G, I, N, Z. Because of this, I conclude that there are $2^6 - 2 = 62$ possible frequent itemsets. The reason for the -2 is because you ignore the empty set, and the set of all 6 letters, as 5 is the maximum set size.

2b) There are 11 frequent itemsets: $B, E, G, I, N, Z, BI, BN, GZ, IN, BIN$. Here are the steps I followed to get my results:

(a) Find all the frequent 1-item itemsets: B, E, G, I, N, Z. Any superset without any of these letters gets pruned.

(b) Find all of the 2-item itemsets that are frequent: BI, BN, GZ, IN. Again, any superset without one of these itemsets is then pruned.

(c) From the remaining 3-item itemsets, evaluate which are considered frequent sets. The only one I found is BIN. Every superset without BIN gets pruned.

(d) There are only three 3-item itemsets remaining: $BEIN, BGIN, BINZ$. Unfortunately, none of these are frequent. All remaining 5-item itemsets get pruned, and were done!

2c) There were four rounds of DB scan done, with 31 total candidates tested.

2d) The 1st approach runs in $O(b)$ time, as each transaction will have b items to check. The 2nd approach runs in $O(m)$ time, because with each transactions, it checks m potential candidates. Eventually in the Apriori algorithm, the m value will be less than the b value, and at that point, the 2nd approach will run faster. The correlation between m and b depends mainly on how large b is.

3a) I found that the largest k value for my frequent k-valued itemsets is $k = 3$. This appears in both the $Milk, Pie, Bread$ itemset as well as the $Milk, Cheese, Break$ itemset. One example of this is
$buys(Milk) \bigcap buys(Pie) \Rightarrow buys(Bread)$ $[0.75, 1]$ or
$buys(Bread) \bigcap buys(Cheese) \Rightarrow buys(Milk)$ $[0.75, 1]$
I'm actually a bit curious about this result. Since Milk and Bread were included in every item set, it seems like they could be skewing the confidence value a bit. I can't say whether or not buying pie or buying cheese implies more that you're going to buy milk. Could we treat items in every set just like items in none of the sets when calculating association rules?

3b) These new restrictions just increase the number of values that the data can take, which in general just reduces the frequency of a particular data value. Because of this, the Apriori pruning happens very quickly. I found that the frequent-k itemset with the largest k value is actually $(WonderBread, SweetPie)$ with a k value of $k = 2$