CS57300: Homework 5

Long Zhen, lzhen@purdue.edu

1. Implement the Apriori association rule algorithm. (25 pts)

   - Consider only itemsets involving feature values of "1". Do not construct itemsets involving the absence of words (i.e., feature values of "0").
   - Consider only itemsets up to size 3 (i.e., sizes 1-3).
   - Construct rules with a single consequent (e.g., `IF good AND service THEN isPositive` *or* `IF she AND said THEN her`).
   - Use a support threshold of **3%** and a confidence threshold of 25%.
   - Compute the association rules that meet the minimum thresholds above.
     Note: you should not change the association rule algorithm to try to construct rules involving the class label features, just treat them in the same way as the word features.

   Analysis of your algorithm:

   (a) Determine the size of the pattern space for itemsets (e.g., the number of elements in the lattice).

   Pattern space for itemset size 1 $\binom{2002}{1} = 2002$

   Pattern space for itemset size 2 $\binom{2002}{2} = 2003001$

   Pattern space for itemset size 3 $\binom{2002}{3} = 1335334000$

   (b) While your algorithm is running, track how many itemsets are:
   (i) considered by the algorithm (i.e., support is counted) and found to be frequent, and
   (ii) considered by the algorithm but found to be infrequent.

Itemset Size	Considered by the algorithm	Frequent	Infrequent
1	2002	376	1626
2	70500	314	70186
3	68	2	66

Table 1: Q1(b) Algorithm Itemsets

   (c) What is the pruning ratio of the Apriori algorithm in this case?
   Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.

   Pruning ratio for itemset size 1 $\frac{0}{2002} = 0$, no pruning

   Pruning ratio for itemset size 2 $\frac{2003001-70500}{2003001} = 96.48\%$

   Pruning ratio for itemset size 3 $\frac{5341336000-68}{5341336000} = 99.99\%$

   (d) What is the false alarm rate of the Apriori algorithm in this case?
   False alarm rate is defined as the percentage of candidate itemsets that are found to be infrequent after explicitly counting support.

False alarm for itemset size 1 $\frac{2002-376}{2002} = 81.22\%$
False alarm for itemset size 2 $\frac{70500-314}{70500} = 99.55\%$
False alarm for itemset size 3 $\frac{68-2}{68} = 97.06\%$

(e) Report the top 30 association rules that are discovered, ordered by confidence. State each rule along with its support and confidence values. Discuss whether the results are interesting or surprising based on your past experience analyzing this dataset.

Rank	Description	Confidence	Support
1	IF ever AND worst THEN isNegative	0.993421052632	0.0302
2	IF worst THEN isNegative	0.992537313433	0.0532
3	IF horrible THEN isNegative	0.983516483516	0.0358
4	IF rude THEN isNegative	0.968	0.0484
5	IF terrible THEN isNegative	0.966666666667	0.0348
6	IF fantastic THEN isPositive	0.926380368098	0.0302
7	IF manager THEN isNegative	0.922033898305	0.0544
8	IF excellent THEN isPositive	0.919642857143	0.0412
9	IF delicious THEN isPositive	0.919444444444	0.0662
10	IF amazing THEN isPositive	0.908256880734	0.0594
11	IF waited THEN isNegative	0.889534883721	0.0306
12	IF madison THEN isPositive	0.876	0.0438
13	IF perfect THEN isPositive	0.867724867725	0.0328
14	IF awesome THEN isPositive	0.860068259386	0.0504
15	IF friendly AND staff THEN isPositive	0.858536585366	0.0352
16	IF phone THEN isNegative	0.857142857143	0.0336
17	IF wonderful THEN isPositive	0.856382978723	0.0322
18	IF money THEN isNegative	0.850318471338	0.0534
19	IF later THEN isNegative	0.846153846154	0.0484
20	IF asked THEN isNegative	0.840517241379	0.078
21	IF friendly THEN isPositive	0.827094474153	0.0928
22	IF favorite THEN isPositive	0.824915824916	0.049
23	IF minutes THEN isNegative	0.816443594646	0.0854
24	IF customers THEN isNegative	0.809734513274	0.0366
25	IF finally THEN isNegative	0.807407407407	0.0436
26	IF 15 THEN isNegative	0.800995024876	0.0322
27	IF customer THEN isNegative	0.796610169492	0.0658
28	IF nothing THEN isNegative	0.790697674419	0.0544
29	IF should THEN isNegative	0.79020979021	0.0678
30	IF point THEN isNegative	0.785714285714	0.0308

Table 2: Q1(e) Top 30 association rules

Discussion This time the result is more accurate than the past experience analyzing this dataset. In the past experience, there are often some results which are opposite with common sense (even within the top score results). But this time the result (at least top 30 in the table) is very accurate to the common knowledge. Based on this accuracy observation, there are some very interesting rules found, such as *IF 15 THEN isNegative.* This is one unexpected or not obvious rule. Detecting it may benefit other experiment.

2. Consider using the χ^2 score instead of confidence as the interestingness measure. (10 pts)

(a) Describe an arbitrary contingency table you would use for calculating χ^2 from a rule in this context—where the rows correspond to whether the antecedent holds (or not) and the columns correspond to whether the consequent holds (or not).

Answer Assume that we are calculating χ^2 for the rule $A \to B$ in this context, where the rows correspond to the antecedent and the columns correspond to the consequent. The contingency table should look like the following. In the table,

- x_{11} is the number of reviews, whose feature A and feature B are both 1
- x_{10} is the number of reviews, whose feature A is 1 and feature B is 0
- x_{01} is the number of reviews, whose feature A is 0 and feature B is 1
- x_{00} is the number of reviews, whose feature A and feature B are both 0

	B	$\bar{B}$
A	x_{11}	x_{10}
$\bar{A}$	x_{01}	x_{00}

Table 3: Q2(a) Contingency Table for $A \to B$

(b) Give the formula for calculating χ^2 from a rule in this context.

Formula

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Explanation Assume when $i = 1$, it is the cell, x_{11}. Use it as an example to explain the formula.

- $o_i = x_{11}$
- e_i is the expected value for this cell. $e_i = [\frac{x_{11}+x_{10}}{x_{11}+x_{10}+x_{01}+x_{00}}] \times [\frac{x_{11}+x_{01}}{x_{11}+x_{10}+x_{01}+x_{00}}] \times [x_{11} + x_{10} + x_{01} + x_{00}]$

(c) For one of your reported rules above, show the associated contingency table and calculated χ^2 score.

Answer Example rule: $worst \to isNegative$, from below data and table, we can easily compute $\chi^2 = 274.788357452$

- $e_{11} = 134$
- $e_{10} = 134$
- $e_{01} = 2366$

	$isNegative$	$isNe\bar{g}ative$
$worst$	266	2
$wo\bar{r}st$	2234	2498

Table 4: Q2(c) Contingency Table for $worst \rightarrow isNegative$

- $e_{00} = 2366$

(d) Describe how the numbers in the contingency table will change when you *generalize* the rule (e.g., by removing terms from the antecedent).

	B	$\bar{B}$
A	increase/no change	increase/no change
$\bar{A}$	decrease/no change	decrease/no change

Table 5: Q2(d) Contingency Table Changes for Generalizing the Rule

Explanation

Case 1 By generalizing the rule, antecedents are not generalized or no useful information is lost/removed from antecedents. Therefore, in this case, no changes will happen to the contingency table.

Case 2 By generalizing the rule, some constraints are removed from antecedents. Because of less constraints, more samples may fall into the first row cells (x_{11}, x_{10}). Both cells in the first row may increase or stay no change. Considering no change to consequents, the sum of each column should stay the same ($x_{11} + x_{01}, x_{10} + x_{00}$). Therefore, cells in the second row (x_{01}, x_{00}), will either decrease or stay no change.

(e) Describe how the numbers in the contingency table will change when you *specialize* the rule (e.g., by adding terms to the antecedent).

	B	$\bar{B}$
A	decrease/no change	decrease/no change
$\bar{A}$	increase/no change	increase/no change

Table 6: Q2(e) Contingency Table Changes for Specializing the Rule

Explanation Similar to Question 2(d), specializing the rule may not bring in new constraints or more useful information. Any cell may stay the same. In good cases, more constraints can result in less samples in the corresponding cells (first row). Because there is no change to the sum of each column, the cells in the second row may increase.

(f) For your example contingency table, report the cell counts for the *best possible* specialization of your reported rule. What is the accuracy of the best possible specialized rule? What is the accuracy of your initial rule (from the contingency table above)?

Answer $Accuracy = \frac{x_{11}+x_{00}}{x_{11}+x_{10}+x_{01}+x_{00}}$
Accuracy of initial rule $\frac{266+2498}{5000} = 55.28\%$
Accuracy of the best possible specialized rule $\frac{266+2500}{5000} = 55.32\%$

	B	$\bar{B}$
A	no change	decrease, best possible 0
$\bar{A}$	no change	increase to maximum possible, 2500

Table 7: Q2(f) Contingency Table Changes for *Best Possible* Specializing the Rule

(g) What can we guarantee about the accuracy of further specializations of the rule? How can this be used to prune the space of rules during search? How does this relate to the Apriori principle?

Explanation For our experiment, the accuracy that can be guaranteed is equal to or less than 55.32%. During the search, this best possible accuracy may be used as a threshold. If the accuracy reached this number, the experiment can be stopped confidently. However, this ideal number should be hard to achieve in the real world scenario. We may use $bestPossible - 0.05\%$ as the guideline to continue/stop search procedure.

3. Modify your association rule algorithm to use different search criteria (PDM p.439). (15 pts)

(a) Use the χ^2 score as the interestingness measure and a test of significance ($\alpha = 0.05$) as the promise criterion.

Answer see code for details.

(b) Rerun your algorithm and report the top 30 newly found rules (ranked inversely by significance). Discuss the results and how they compare to the rules you identified in part 1.

Discussion

First Because χ^2 is mainly designed to detect the dependency between two features, which they are independent or related. In the itemset of size 2, *IF a THEN b* and *IF b THEN a* will have the exactly same χ^2 score. Therefore, we can see many rules (of size 2) are inverse to each other. Knowing this feature, to observe more rules, I also provide the top 30 χ^2 rules without size 2 inverse duplicates.

Second Both top 30 with duplicate or without duplicate, many different rules can be detected. Many different rules make senses but are not meaningful, such as *IF make THEN sure* or *IF better THEN than.* They are just commonly used English phrase but do not provide too much information.

(c) Discuss how association rule algorithms can suffer from multiple comparison problems[1].

[1] Reference: D. Jensen and P. Cohen (2000). "Multiple Comparisons in Induction Algorithms." *Machine Learning* 38: 309-338.

Rank	Description	p-value	χ^2	Support
1	IF isNegative AND ever THEN worst	$2.55998211088e - 191$	870.484581221	0.0302
2	IF isPositive AND staff THEN friendly	$1.87483259718e - 104$	471.069848154	0.0352
3	IF worst THEN ever	$5.14162800447e - 99$	446.080567961	0.0304
4	IF ever THEN worst	$5.14162800447e - 99$	446.080567961	0.0304
5	IF isNegative AND worst THEN ever	$1.80333339361e - 98$	443.576467968	0.0302
6	IF friendly THEN staff	$1.82661137263e - 65$	291.996022388	0.041
7	IF staff THEN friendly	$1.82661137263e - 65$	291.996022388	0.041
8	IF worst THEN isNegative	$2.91180992558e - 61$	272.710570142	0.0532
9	IF isNegative THEN worst	$2.91180992558e - 61$	272.710570142	0.0532
10	IF isPositive THEN delicious	$6.22805630979e - 61$	271.195522031	0.0662
11	IF delicious THEN isPositive	$6.22805630979e - 61$	271.195522031	0.0662
12	IF friendly THEN isPositive	$1.91447810428e - 60$	268.95781557	0.0928
13	IF isPositive THEN friendly	$1.91447810428e - 60$	268.95781557	0.0928
14	IF friendly AND isPositive THEN staff	$1.25976546809e - 59$	265.203637587	0.0352
15	IF make THEN sure	$4.02680397964e - 56$	249.126093482	0.0336
16	IF sure THEN make	$4.02680397964e - 56$	249.126093482	0.0336
17	IF better THEN than	$1.6925322909e - 53$	237.093183946	0.0372
18	IF than THEN better	$1.6925322909e - 53$	237.093183946	0.0372
19	IF ordered THEN order	$1.760013431567e - 53$	237.015181641	0.0342
20	IF order THEN ordered	$1.760013431567e - 53$	237.015181641	0.0342
21	IF isNegative THEN asked	$3.369291394 81e - 53$	235.721982759	0.078
22	IF asked THEN isNegative	$3.36929139481e - 53$	235.721982759	0.078
23	IF isNegative THEN minutes	$1.6599266691e - 52$	232.546121647	0.0854
24	IF minutes THEN isNegative	$1.6599266691e - 52$	232.546121647	0.0854
25	IF amazing THEN isPositive	$2.77790612572e - 52$	231.520655781	0.0594
26	IF isPositive THEN amazing	$2.77790612572e - 52$	231.520655781	0.0594
27	IF rude THEN isNegative	$1.2130456683e - 51$	228.585263158	0.0484
28	IF isNegative THEN rude	$1.2130456683e - 51$	228.585263158	0.0484
29	IF love THEN isPositive	$1.80312385328e - 50$	223.210906302	0.0958
30	IF isPositive THEN love	$1.80312385328e - 50$	223.210906302	0.0958

Table 8: Q3(b) Top 30 newly found rules

Rank	Description	p-value	χ^2	Support
1	IF isNegative AND ever THEN worst	$2.55998211088e-191$	870.484581221	0.0302
2	IF isPositive AND staff THEN friendly	$1.87483259718e-104$	471.069848154	0.0352
3	IF ever THEN worst	$5.14162800447e-99$	446.080567961	0.0304
4	IF isNegative AND worst THEN ever	$1.80333339361e-98$	443.576467968	0.0302
5	IF friendly THEN staff	$1.82661137263e-65$	291.996022388	0.041
6	IF isNegative THEN worst	$2.91180992558e-61$	272.710570142	0.0532
7	IF isPositive THEN delicious	$6.22805630979e-61$	271.195522031	0.0662
8	IF friendly THEN isPositive	$1.91447810428e-60$	268.95781557	0.0928
9	IF friendly AND isPositive THEN staff	$1.25976546809e-59$	265.203637587	0.0352
10	IF sure THEN make	$4.02680397964e-56$	249.126093482	0.0336
11	IF better THEN than	$1.6925322909e-53$	237.093183946	0.0372
12	IF order THEN ordered	$1.76013431567e-53$	237.015181641	0.0342
13	IF isNegative THEN asked	$3.369291394\,81e-53$	235.721982759	0.078
14	IF isNegative THEN minutes	$1.6599266691e-52$	232.546121647	0.0854
15	IF isPositive THEN amazing	$2.77790612572e-52$	231.520655781	0.0594
16	IF isNegative THEN rude	$1.2130456683e-51$	228.585263158	0.0484
17	IF love THEN isPositive	$1.80312385328e-50$	223.210906302	0.0958
18	IF manager THEN isNegative	$4.13116413804e-50$	221.560186603	0.0544
19	IF isPositive THEN again	$2.58493072948e-44$	194.993413809	0.0306
20	IF isNegative THEN again	$2.58493072948e-44$	194.993413809	0.0966
21	IF how THEN asked	$8.05144523785e-43$	188.151293611	0.031
22	IF ordered THEN came	$1.67547273605e-42$	186.693334095	0.0306
23	IF isNegative THEN horrible	$7.22713071175e-40$	174.625602708	0.0358
24	IF before THEN minutes	$3.54836728256e-39$	171.461223186	0.031
25	IF excellent THEN isPositive	$2.0115672328e-37$	163.433424563	0.0412
26	IF isNegative THEN money	$2.53073916258e-37$	162.976993402	0.0534
27	IF isNegative THEN terrible	$7.85686440337e-37$	160.724988474	0.0348
28	IF awesome THEN isPositive	$1.20130963956e-36$	159.880970249	0.0504
29	IF isNegative THEN customer	$4.78425086549e-36$	157.134252976	0.0658
30	IF isNegative THEN should	$5.6003341252e-36$	156.821228797	0.0678

Table 9: Q3(b) Top 30 newly found rules without size 2 duplicate

Answer Since in the association rule algorithm multiple χ^2 comparisons are performed to filter (accept or reject null hypothesis) rules, if the same original threshold is used for all comparisons, incorrectly filtering are more likely to occur. In other words, when one considers the set as a whole, errors are more likely to happen. Therefore, some adjustments are needed to apply to prevent this.

(d) Outline how to use a Bonferroni correction to adjust for multiple comparison problems when using the χ^2 interestingness measure.

Answer According to the Bonferroni correction,

First count the number of candidates, in my example (with size 2 duplicate), 634
Second compute $newThreshold = \frac{oldThreshold}{numOfCandidate} = \frac{0.05}{634}$
Third use this new threshold to compute qualified rules

(e) Rerun the algorithm using Bonferroni correction and discuss any change in the discovered rules.

Answer

I. Since the top 30 are the best rules that are detected, if the size of final rule set is not less than 30, the top should stay the same. As expected, the result of top 30 says the same.

II. However, we do reject more rules. The original size of accepted rule set for size 2 and 3 is $\{546, 6\}$. After Bonferroni correction, the size is $\{418, 6\}$. There are 128 more rules rejected.