CSC 4740/6740 **Data mining** **D**ue by: Oct. 6th

Name:

Assignment 3

Q1: The Apriori algorithm makes use of *prior knowledge* of subset support properties. ((10 points))

a) Explain that the (relative) support of any nonempty subset s' of frequent itemset s must be at least as great as the (relative) support of s.

b) Given frequent itemset L and subset s of L, prove that the confidence of the rule "$s' \Rightarrow (L - s')$" cannot be more than the confidence of "$s \Rightarrow (L - s)$", where s' is a subset of s.

Q2: A database has 5 transactions. Let *min sup* = 60% and *min conf* = 80%.

TID	*items bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, K, I, E}

(a) Find all frequent itemsets using Apriori. Please do **NOT** call the functions from existing package or library. Do it step by step by hands or by your own code. Grades are given based on intermediate results from each step. (20 points)

(b) Find all frequent itemsets using FP-growth. Please do **NOT** call the functions from existing package or library. Do it step by step by hands or by your own code. Grades are given based on intermediate results from each step. (35 points)

(c) List all of the *strong* association rules (with support s and confidence c *satisfying the min requirement*) matching the following rule, where X is a variable representing a customer, and *item_i* denotes variables representing items (e.g., *"A", "B","*M" i.e., if any customer buys A and B, then the customer will buy M) (5 points))

$$\forall x \in transaction,\ buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)$$

Q3: Give a short example to show that items in a strong association rule (Ex: min_support=10%, min_confidence=50%) may actually be *negatively correlated*. (10 points)

CSC 4740/6740 **Data mining** **D**ue by: Oct. 6th

Name:

Q4: The following contingency table summarizes supermarket transaction data, where *hot dogs* refers to the transactions containing hot dogs, $\overline{hotdogs}$ refers to the transactions that do NOT contain hot dogs, *hamburgers* refers to the transactions containing hamburgers, and $\overline{hamburgers}$ refers to the transactions that do NOT contain hamburgers.

	hot dog	$\overline{hotdogs}$	Σ_{row}
hamburgers	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
Σ_{col}	3000	2000	5000

(a) Suppose that the association rule "hot dogs ->hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong? (5 points))

(d) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship (positively or negatively) exists between the two? (5 points))

Chi-Square Table

TABLE 24.1 To be significant at level α, a chi-square statistic must be larger than the table entry for α

	Significance Level α						
df	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.01**	**0.001**
1	1.32	1.64	2.07	2.71	3.84	6.63	10.83
2	2.77	3.22	3.79	4.61	5.99	9.21	13.82
3	4.11	4.64	5.32	6.25	7.81	11.34	16.27
4	5.39	5.99	6.74	7.78	9.49	13.28	18.47
5	6.63	7.29	8.12	9.24	11.07	15.09	20.51
6	7.84	8.56	9.45	10.64	12.59	16.81	22.46
7	9.04	9.80	10.75	12.02	14.07	18.48	24.32
8	10.22	11.03	12.03	13.36	15.51	20.09	26.12
9	11.39	12.24	13.29	14.68	16.92	21.67	27.88

(e) Compute and compare the **all confidence, max confidence, Kulczynski**, **cosine measures,** and **lift** on the given data (definition of each metric are in the lecture slides) (10 points)