CSci 5523 Homework 3 Due Nov 2, 2011

Only a subset of questions will be graded.

1. Consider the market basket transactions shown in Table 1.

Table 1: Market basket transactions.

Transaction ID	Items Bought
1	{Beer, Diapers}
2	{Milk, Diapers, Bread, Butter}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Milk, Beer, Diapers, Eggs}
6	{Beer, Cookies, Diapers}
7	{Milk, Diapers, Bread, Butter}
8	{Bread, Butter, Diapers}
9	{Bread, Butter, Milk}
10	{Beer, Milk, Cookies}

(a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

Answer: There are seven items in the data set. Therefore the total number of rules is 1932.

(b) What is the maximum size of frequent itemsets that can be extracted (assuming $minsup > 0$)?

Answer: Because the longest transaction contains 4 items, the maximum size of frequent itemset is 4.

(c) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

Answer: $\binom{7}{3} = 35$.

(d) Find an itemset (of size 2 or larger) that has the largest support.

Answer: {Bread, Butter}.

(e) Find a pair of items, a and b, such that the rules $\{a\} \longrightarrow \{b\}$ and $\{b\} \longrightarrow \{a\}$ have the same confidence.

Answer: (Bread, Butter).

2. Consider the following set of frequent 3-itemsets:

$$\{a, b, c\}, \{a, b, d\}, \{a, b, e\}, \{a, c, d\}, \{a, d, e\}, \{a, e, f\}, \{c, d, e\}, \{c, e, f\}, \{d, e, f\}.$$

Assume that there are only six items in the data set.

(a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

Answer:

$\{a, b, c, d\}, \{a, b, c, e\}, \{a, b, c, f\}, \{a, b, d, e\}, \{a, b, d, f\}, \{a, b, e, f\}, \{a, c, d, e\}, \{a, c, d, f\}, \{a, d, e, f\}, \{a, c$

(b) List all candidate 4-itemsets obtained by the candidate generation procedure in *Apriori*.
Answer:

$$\{a, b, c, d\}, \{a, b, c, e\}, \{a, b, d, e\}, \{a, b, e, f\}, \{a, c, d, e\}, \{a, d, e, f\}, \{a, c, e, f\}, \{c, d, e, f\}$$

(c) List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.

Answer: None

(d) Based on the list of frequent 3-itemsets given above, is it possible to generate a frequent 5-itemset? State your reason clearly.

Answer: None, as there are not frequent 4-itemsets

3. Consider the interestingness measure, $M = \frac{P(B|A)-P(B)}{1-P(B)}$, for an association rule $A \rightarrow B$.

(a) What is the range of this measure? When does the measure attain its maximum and minimum values?

Range of the measure: $-\infty < M \leq 1$.
Since $M = \frac{P(B|A)-P(B)}{1-P(B)}$

Maximum value occurs when $P(A, B) = P(A)$:

$$\text{At this value, } M = \frac{1-P(B)}{1-P(B)} = 1$$

Minimum Value occurs when $\frac{P(A,B)}{P(A)} = 0$:

$$\text{At this value } M = \frac{-P(B)}{1-P(B)}$$

This value decreases with increasing $P(B)$.

(b) How does M behave when $P(A, B)$ is increased while $P(A)$ and $P(B)$ remain unchanged?
M increases.

(c) How does M behave when $P(A)$ is increased while $P(A, B)$ and $P(B)$ remain unchanged?
M decreases.

(d) How does M behave when $P(B)$ is increased while $P(A, B)$ and $P(A)$ remain unchanged?
Let $a = \frac{P(A,B)}{P(A)}$ and $y = P(B)$. Since $P(A, B)$ and $P(A)$ remain unchanged then a is a constant. However, the measure M can be shown as a function

$$f(y) = \frac{a-y}{1-y}.$$

The derivative of f is:

$$\frac{df(y)}{dy} = \frac{a-1}{(1-y)^2}$$

Since $a \leq 1$, M decreases while $y = P(B)$ increases.

(e) Is the measure symmetric under the variable permutation?
No.

(f) What is the value of measure when A and B are statistically independent?

$$A \text{ and } B \text{ are statistically independent So:}$$
$$P(A,B) = P(A) \times P(B)$$

Then

$$\frac{P(A,B)}{P(A)} = P(B)$$

Therefore the numerator of M becomes zero, so $M = 0$.

(g) Does the measure remain invariant under the row or column scaling operations?
Example:

	B	$\bar{B}$
A	4	6
$\bar{A}$	6	4

For this table, $M = \frac{0.4-0.5}{1-0.5} = -0.2$

Column scaling operation produces following table:

	B	$\bar{B}$
A	40	6
$\bar{A}$	60	4

For this table, $M = \frac{0.86-0.9}{1-0.9} = -0.4$

So M doesn't remain invariant under column or row scaling operations.

(h) Does the measure remain invariant under the inversion operation?
Example:

	B	$\bar{B}$
A	2	18
$\bar{A}$	28	12

For this table, $M = \frac{0.1-0.5}{1-0.5} = -0.8$

Under the inversion operation: $A = \bar{A}$ and $B = \bar{B}$. Therefore following table is produced:

	B	$\bar{B}$
A	12	28
$\bar{A}$	18	2

For this table, $M = \frac{0.4-0.5}{1-0.5} = -0.2$

So M doesn't remain invariant under the inversion operation.

Table 2: Data set for Question 4.

Age	Play Piano	Enjoy Classical Music
5	Yes	Yes
7	Yes	Yes
11	Yes	No
18	No	No
20	Yes	Yes
22	No	No
24	No	Yes
29	Yes	Yes
34	No	No
35	No	Yes
40	No	No
49	No	Yes

4. Consider the data set shown in Table 2. Suppose we are interested in extracting the following association rule:

$$\{\alpha_1 \leq \text{Age} \leq \alpha_2, \text{Play Piano} = \text{Yes}\} \longrightarrow \{\text{Enjoy Classical Music} = \text{Yes}\}$$

To handle the continuous attribute, we apply the equal-frequency approach with 3, 4, and 6 intervals. Categorical attributes are handled by introducing as many new asymmetric binary attributes as the number of categorical values. Assume that the support threshold is 10% and the confidence threshold is 70%.

(a) Suppose we discretize the Age attribute into 3 equal-frequency intervals. Find a pair of values for α_1 and α_2 that satisfy the minimum support and minimum confidence requirements.

Answer: ($\alpha_1 = 20$, $\alpha_2 = 29$): s = 16.7%, c = 100%

(b) Repeat part (a) by discretizing the Age attribute into 4 equal-frequency intervals. Compare the extracted rules against the ones you had obtained in part (a).

Answer: No rules satisfies the support and confidence intervals.

(c) Repeat part (a) by discretizing the Age attribute into 6 equal-frequency intervals. Compare the extracted rules against the ones you had obtained in part (a).

($\alpha_1 = 5$, $\alpha_2 = 7$): s = 16.7%, c = 100%

(d) From the results in part (a), (b), and (c), discuss how the choice of discretization intervals will affect the rules extracted by association rule mining algorithms.

Answer: If the discretization interval is too narrow, the rule in part (a) will be lost and sometimes spurious rules are being discovered. So, the rules found are sensitive to the choice of the width of the discretization interval.

5. Consider a binary data set representing the movie preference of N Netflix users of K movies. In the tabular representation of this data set shown below, the rows are users, and the columns are movies. The entry corresponding to the i^{th} row (user) and j^{th} column (movie) is a 1 if the user likes the movie and 0 if the user dislikes it.

(a) Which of these properties an interestingness measure should possess to be useful for evaluating if users have similar likes or dislines for movies ? Briefly justify your answer.

Table 3: Example movie preference dataset

	Movie 1	Movie 2	Movie 3
User 1	1	0	1
User 2	0	1	0
User 3	1	0	0

i. Symmetry

ii. Invariant under null addition

iii. Invariant under inversion

Answer: i and iii. Because both like and dislike matter in this case.

(b) Based on your answers above, would you prefer confidence, the cosine measure, or correlation for this task ? Briefly justify your answer.

Answer: Correlation, because it possesses i and iii.

(c) Repeat the above two parts (a and b) if a 1 is assigned in the above table only if a user likes a movie. A zero in the rest of the entries means that the user does not know about the movie

Answer: i and ii, because only ”like” matters. Consine should be used. Note that confidence is not symmetric.

6. Given a data set, let S be the set of patterns with support $\geq a$, and T be the set of hypercliques with support $\geq b$ and h-confidence $\geq c$.

(a) What's the relationship between S and T, if $a = b$, and Why?

Answer: T is a subset of S.

(b) What's the relationship between S and T, if $a < b$, and Why?

Answer: T is still a subset of S, since T is a subset of the set of patterns with support $\geq b$, which is a subset of S.

(c) What's the relationship between S and T, if $a > b$, and Why?

Answer: If we consider size-1 itemset, then S and T is guaranteed to be overlapping (those size-1 itemset with support $\geq a$). If we only consider itemsets of size greater than 1, then, S and T could be either overlapping or disjoint, depending on specific transaction data.

Turn over for other questions...

7. Consider a transaction data set shown in the table below. The corresponding itemset lattice for this data set is shown in the figure below, where the TID-list associated with each itemset is also shown by its side. For example, node $\{A\}$ is present in TIDs 1, 2 and 4. Now, answer the questions below, assuming a minimum absolute support threshold of 2. (You may ignore the null set when listing your answers).

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

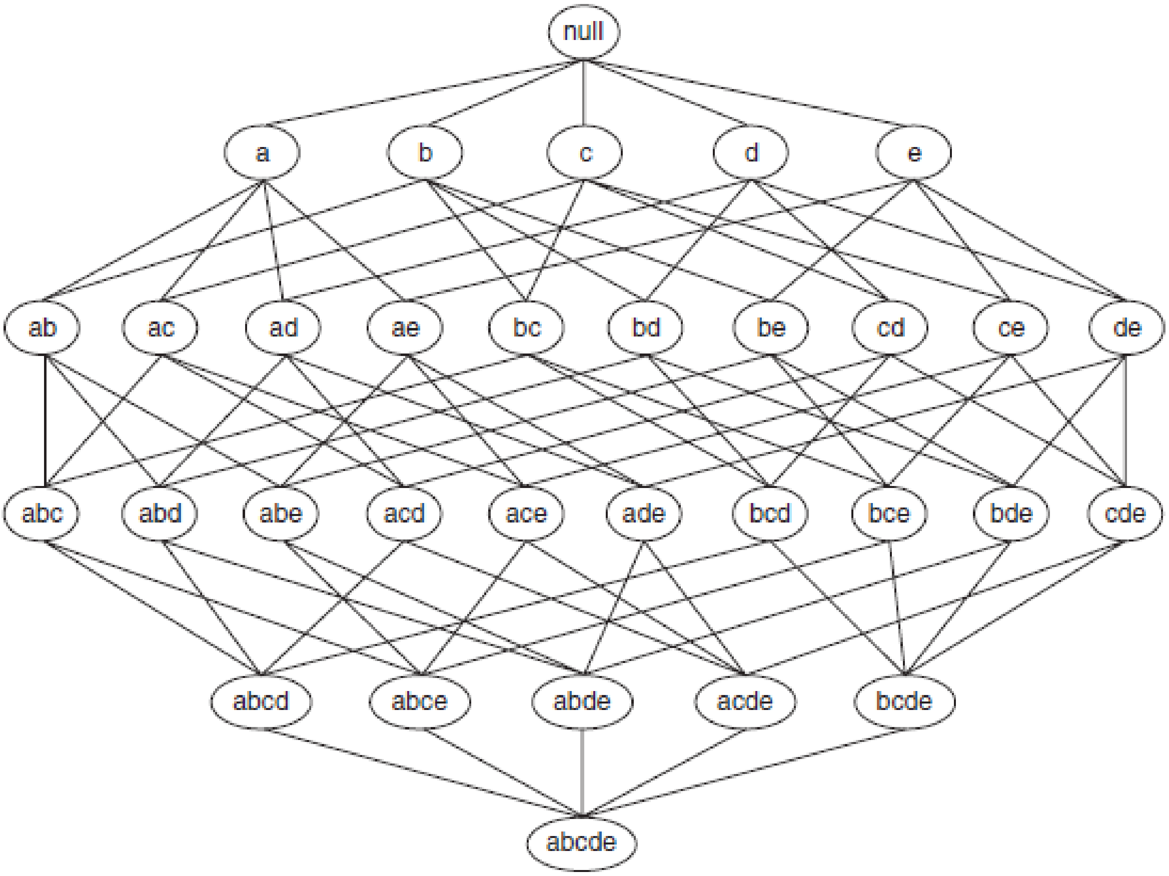


Figure 1: Itemset Lattice

(a) Which size-3 items are maximal
Answer: ABC, ACD.

(b) Which size-3 itemsets with non-zero support are not closed
Answer: ABD, ACD, ACE, ADE, BCD, CDE

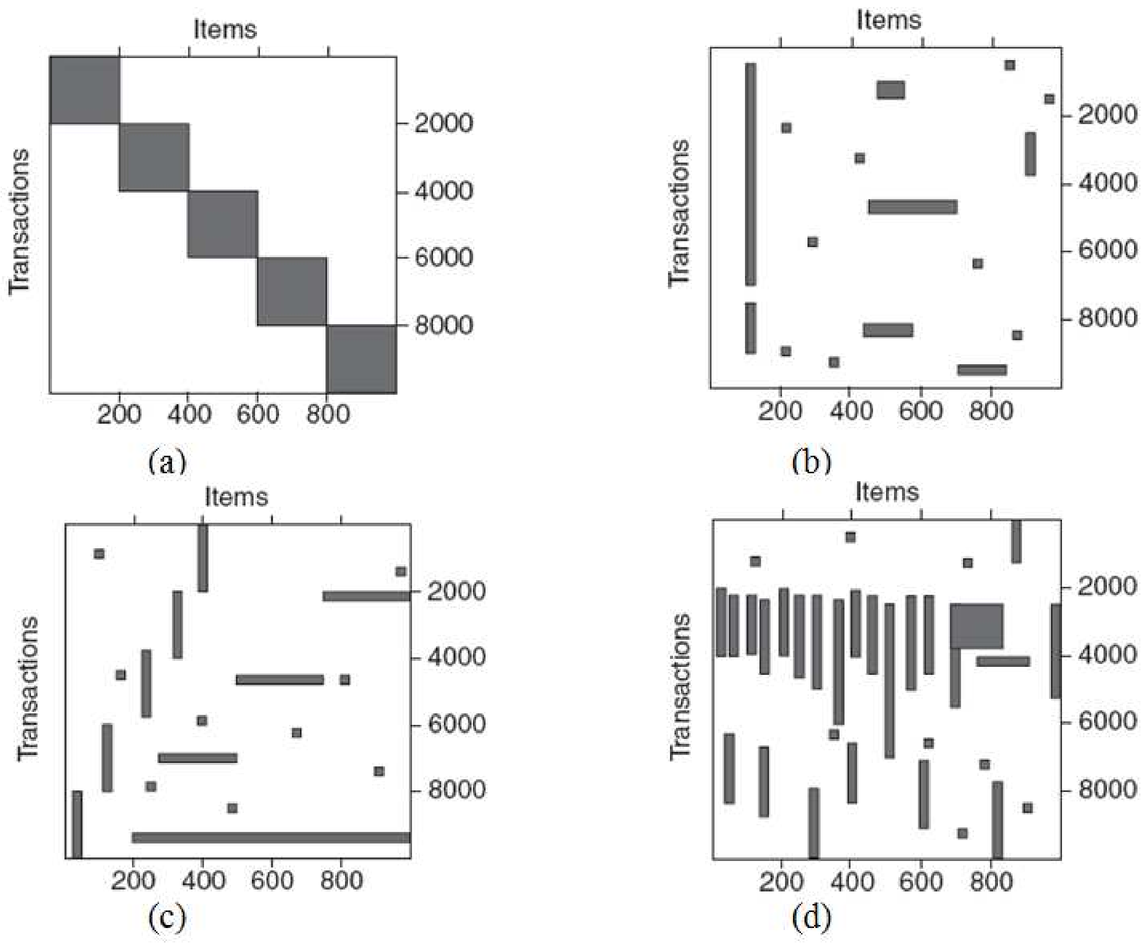
(c) If an itemset is maximal frequent, then its supersets must be infrequent. Answer True of False with a short reason ?

Answer: Yes. a maximal itemset is a frequent itemset that has no frequent superset.

(d) A closed itemset is not necessarily frequent. Answer True or False with a short reason.

Answer: Yes, a closed itemset is an itemset that has no superset with the same support.

8. The figures below depict 4 transaction data sets, each having 1000 items and 10,000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of items. We apply the Apriori algorithm to extract frequent itemsets with minsup=10



(a) (b) (c) (d)

Figure 2: Transaction datasets

Answer: Note that, the minsup is 10, as stated clearly, not 10 percent.:

(a) (i) Which data set will produce the highest number of frequent itemsets?

Answer: C, mainly because the long horizontal bar between transaction 8000-10000 and items 200-1000.

(b) (ii) Which data set will produce the smallest number of closed frequent itemsets?

Answer: A, just five.

(c) (iii) Which data set will produce the longest (size) frequent itemset?
Answer: C, longest size greater than 800.

(d) (iv) Which data set will produce the itemset with the highest support? **Answer**: B, more than 8000.

(e) (v) Which data set will produce the longest (size) closed itemset? **Answer**: C