CS 412: Intro to Data Mining Exam I

Terms in this set (76)

AGM	-Apriori based graph mining -breadth first search -every iteration grow one vertex at a time
Apriori property of frequent patterns	any subset of a frequent itemset must be frequent
Below is a table of transactions. According to the introduced pattern distance measure, what is the distance between pattern "abc" and pattern "abd"? Transaction Itemset T1 abcde T2 abefg T3 abcdef T4 abcdf T5 abcdeg A.) 0 B.) 0.5 C.) 0.333 D.) 0.2	A.) 0
chi-square	-not null invariant x^2 = E (observed - expected)^2/observed x^2 = 0 indep x^2 > 0 correl
closed patterns	if X is frequent and there exists no super pattern Y where every element of X is a part of Y with the same support

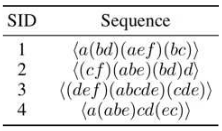
Considering the Apriori algorithm, assume we have 5 items (A to E) in total. In the 1st scan, we find out all frequent items A, B, C, and E. How many size-2 (i.e., containing 2 items, e.g., A, B) itemsets should be considered in the 2nd scan, i.e., have potential to be size-2 frequent itemsets? A.) 25 B.) 10 C.) 4 D.) 6	6
Consider the database containing the transactions T1 : {a1, a2, a3}, T2 : {a2, a3, a4}. Let minsup = 1. What fraction of all frequent patterns are max frequent patterns?	2/11

Consider the database containing the transaction T1 : {a1, a2, a3}, T2 : {a2, a3, a4}, T3 : {a1, a3, a4}. Let mini-support (minsup) = 2. Which of the following frequent patterns is closed?
A.) {a4}
B.) {a1, a3}
C.) {a2}
D.) {a1}
	{a1, a3}

constraint-based pattern mining	-can find a lot of patterns but make sure they're user-interested -user provides constraints on what should be mined -constraint pushing similar to push selection first in DB query processing

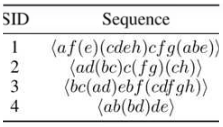
and so do all of its supersets. Which of following constraints are anti-monotone? A.) sum(S.price) > 25 B.) range(S.price) < 10 C.) avg(S.price) > 15 D.) var(S.price) > 20, where var(·) is the variance function	
A constraint is monotone if an itemset S satisfies the constraint and so do all of its supersets. Which of following constraints are monotone? A.) median(S.price) < 20 B.) relative support of S > 0.1 C.) min(S.price) < 15 D.) var(S.price) < 20, where var(·) is the variance function	C.) min(S.price) < 15
A constraint is succinct if the constraint c can be enforced by directly manipulating the data. Which of following constraints are succinct? A.) avg(S.price) > 40 B.) sum(S.price) > 40 C.) max(S.price) < 20 D.) var(S.price) > 10, where var(·) is the variance function	C.) max(S.price) < 20
convertible constraints	c can be converted to monotonic or anti-monotonic if terms can be properly ordered in processing
cosine	-null invariant S(AUB)/sqrt(S(A)xS(B)) [0,1]
criteria to judge quality of phrases	-popularity, concordance, informativeness, completeness
data anti-monotonicity	if a transaction t does not satisfy c then t can be pruned to reduce data processing effort

<table>
<tr><td>For mining text data, which of the following algorithms will not output phrases?
A.) LDA
B.) ToPMine
C.) TurboTopics
D.) SegPhrase</td><td>A.) LDA</td></tr>
<tr><td>FPGrowth</td><td>-a frequent pattern-growth approach
-find frequent single items and partition the database based on each such item
-recursively grow frequent patterns by doing the above for each partitioned database (also called conditional database)</td></tr>
<tr><td>frequent itemsets</td><td>if the support (count) of X is no less than minsup threshold denoted as σ</td></tr>
<tr><td>FSG</td><td>-Frequent sub graphs
-Apriori based
-breadth first search
-every iteration grow one edge at a time
-edge growing is more efficient than vertex growing</td></tr>
<tr><td>Given a sequence database, as shown in table 2, with support threshold minsup = 3, which of the following sequences are frequent?

A.) < abc >
B.) < f(ab) >
C.) < (bd)b >
D.) < (ae)c >
E.) None of the above</td><td>D.) < (ae)c >
<table><tr><th>SID</th><th>Sequence</th></tr><tr><td>1</td><td>$\langle a(bd)(aef)(bc)\rangle$</td></tr><tr><td>2</td><td>$\langle (cf)(abe)(bd)d\rangle$</td></tr><tr><td>3</td><td>$\langle (def)(abcde)(cde)\rangle$</td></tr><tr><td>4</td><td>$\langle a(abe)cd(ec)\rangle$</td></tr></table>Table 2: Sequence database.</td></tr>
</table>

= 1. Which of the following does not belong to the < d >-projected database?

A.) < (_eh)cfg(abe) >
B.) < ebf(cdfgh) >
C.) < (bc)c(fg)(ch) >
D.) < (_b)de >



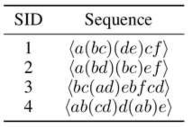
SID	Sequence
1	$\langle af(e)(cdeh)cfg(abe)\rangle$
2	$\langle ad(bc)c(fg)(ch)\rangle$
3	$\langle bc(ad)ebf(cdfgh)\rangle$
4	$\langle ab(bd)de\rangle$

Table 11: Sequence database.

Given a sequence database as shown in the following table, suppose we use the SPADE algorithm to find the frequent sequential patterns. Which of the following sequences (in the format of <SID, EID>) belong to the mapped database of item a?

A.) <1, 2>
B.) <4, 1>
C.) <1, 1>
D.) <3, 2>

B.) <4, 1>
C.) <1, 1>



SID	Sequence
1	$\langle a(bc)(de)cf\rangle$
2	$\langle a(bd)(bc)ef\rangle$
3	$\langle bc(ad)ebfcd\rangle$
4	$\langle ab(cd)d(ab)e\rangle$

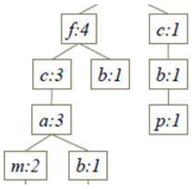
Table 3: Sequence database.

Given a text corpus, which of the following can be used for measuring the colocation strength for a pair of words? Select all that apply.
A.) Likelihood ratio
B.) Edit distance
C.) Mutual information
D.) Chi-squared test

A.) Likelihood ratio
C.) Mutual information
D.) Chi-squared test

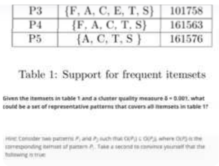
Given the FP-tree as shown in figure 1, what is the support of {c,p}?

3



what could be a set of representative patterns that covers all itemsets in table 1?

A.) {{F, A, C, E, S}, {F, A, C, T, S}}
B.) {{F, A, C, E, S}, {A, C, E, S}}
C.) {{A, C, E, S}, {A, C, T, S}}
D.) {{F, A, C, E, S}, {F, A, C, E, T, S}, {F, A, C, T, S}}
E.) {{F, A, C, E, T, S}}



P3	{F, A, C, E, T, S}	101758
P4	{F, A, C, T, S}	161563
P5	{A, C, T, S }	161576

Table 1: Support for frequent itemsets

Given the itemsets in table 1 and a cluster quality measure δ = 0.001, what could be a set of representative patterns that covers all itemsets in table 1?

Hint: Consider two patterns P_1 and P_2 such that $O(P_1) \subseteq O(P_2)$, where $O(P_i)$ is the corresponding itemset of pattern P_i. Take a second to convince yourself that the following is true:

GSP

Apriori based sequential pattern mining
length 1 => length 2 => go until minsup cannot be met

gSpan

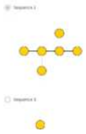
-depth first search
-try to control the order of growth
-grow in order of rightmost path (the path from root to the right most leaf choosing the vertex with the smallest index at each step)

If we know the support of itemset {a, b} is 10, which of the following numbers are the possible supports of itemset {a, c}? Select all that apply.
A.) 10
B.) 11
C.) 9

10,11,9

If we know the support of itemset {a} is 50 and the support of itemset {a, b, c} is 10, which of the following numbers are the possible supports of itemset {a, w}? Select all that apply.
A.) 50
B.) 100
C.) 30
D.) 10
E.) 5

50,30,10,5

sup = 3, which of the following sequences is not a frequent graph pattern?	
itemsets	a set of one or more items
Jaccard	-null invariant S(AUB)/[S(A)+S(B)-S(AUB)] [0,1]
KERT	-phrase construction as a post-processing step to LDA -ranks according to popularity, concordance, informativeness, completeness
lift	-not null invariant lift(B,C) = S(BUC)/(S(B)xS(C)) lift(B,C) = 1 B and C are indep >1 pos correl <1 neg correl
max patterns	if X is frequent and there exists no frequent super pattern Y where every element of X is a part of Y
null-invariance	value does not change with the number of null transactions
pattern anti-monotonicity	if constraint c is violated, its further mining can be terminated
pattern mining with any-monotonic constraints	if an itemset S violates constraint c, so does any of its supersets
pattern mining with monotonic constrains	we do not need to check c again if an itemset S satisfies the constraint c, so does any of its supersets
pattern mining with multiple constraints	-if there exists an order R making both c1 and c2 convertible try to sort items in the order that benefits pruning most
pattern monotonicity	if c is satisfied, no need to check c again

phrasal segmentation	-partition a sequence of words by maximizing likelihood -consider length penalty and filter out phrases with low rectified frequency
PhraseLDA	-viewing each sentence as a time-series of words -the generative parameter (topic) changes periodically -each word is drawn based on previous m words (context) and current phrase topic
Phrase Mining	model a phrase as a sequence labeling problem
PrefixSpan	pattern growth Approach, split into prefix/suffix
SegPhrase	-mining quality phrase with tiny training sets -built off TopMine -integrates phrase mining with phrasal segmentation and classification - the frequent patterns are used as candidate phrases which are later filtered
SPADE	based on vertical data format; building up of positions where items occur together; sequence ID is same, then join by element ID; uses Apriori candidate generation
Spider-Mine	-mining large patters/social networks -mine top-k largest frequent substructure patterns whose diameter is bounded by Dmax with a probability at least 1-ε -large patterns are composed of a number of small components ("spiders") which will eventually connect together after some rounds of patter growth

transactions contained beer, while 5,000 contained frying pans. 600 transactions contained both beer and frying pans. Which of the following is true? A.) For ε = 0.1, {beer, frying pans} is a negative pattern under the null-invariant definition of negatively correlated patterns. B.) More information is needed to determine if {beer, frying pans} is a negative pattern. C.) There does not exist a value for ε such that {beer, frying pans} is a negative pattern by the null-invariant definition of negative patterns. D.) {beer, frying pans} is a negative pattern under the support-based definition of negatively correlated patterns.	patterns.
succinct constraints	if the constraint c can be enforced by directly manipulating the data
support	probability that a transaction contains XUY

hot dogs (HD) vs. hamburgers (HM). We have the following 2×2 contingency table summarizing the statistics. If lift is used to measure the correlation between HD and HM, what is the value for lift(HD, HM)?

	HD	¬HD	Σrow
HM	40	24	64
¬HM	210	126	336
Σcol	250	150	400

A.) 1
B.) -∞
C.) 0
D.) -1

Suppose one needs to frequent patterns at two different levels, with mini-support (minsup) of 5% (higher level) and 3% (lower level), respectively. If using shared multi-level mining, which mini-support (minsup) threshold should be used to generate candidate patterns for the lower level?
A.) 1%
B.) 6%
C.) 5%
D.) 3%

D.) 3%

(CM) and fiction (FC) in the transaction history of a bookstore. We have the following 2×2 contingency table summarizing the transactions. If χ2 is used to measure the correlation between CM and FC, what is the χ2 score?

	CM	¬CM	Σrow
FC	300	700	1000
¬FC	1200	800	2000
Σcol	1500	1500	3000

A.) -240
B.) -80
C.) 80
D.) 240

Suppose we are interested in analyzing the transaction history of several supermarkets with respect to the purchase of apples (A) and bananas (B). We have the following table summarizing the transactions.

	AB	¬AB	A¬B	¬A¬B
S1	100,000	7,000	3,000	300
S2	100,000	7,000	3,000	90,000

Denote χ2i as the χ2 measure and ci as the cosine measure for supermarket Si (i = 1, 2). Which of the following is correct?
A.) χ^21 = χ^22, c1 = c2
B.) χ^21 ≠ χ^22, c1 = c2
C.) χ^21 ≠ χ^22, c1 ≠ c2
D.) χ^21 = χ^22, c1≠ c2

χ^21 ≠ χ^22, c1 = c2

the frequent sequential patterns. After scanning the database once, we find the frequent singleton sequences are: a, b, d. Which of the following could be possible length-2 candidate sequences?
A.) <ab>
B.) <(bd)>
C.) <ac>
D.) <(bc)>

Suppose we use the CloSpan algorithm to find all closed sequential patterns from a sequence database with minimum support 15. During the mining process, we derive the following sequences along with the sizes of their projected DBs: <c>: 50, <ac> 40, <ab> 30, <bc>: 50. Then we use the backward sub-pattern rule and the backward super-pattern rule to prune redundant search space. Which of the projected DBs will remain after the pruning?
A.) <ab>
B.) <c>
C.) <bc>
D.) <ac>

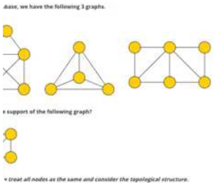
A.) <ab>
C.) <bc>
D.) <ac>

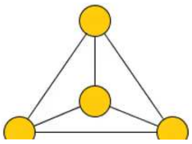
extract phrases. Given the five statements below and a support threshold 3, which of the given phrases can be considered? Select all that apply.
(1) Support vector machine is a classifier.
(2) Neural network performs equally well as support vector machine.
(3) We propose a method that combines support vector machine with kernel method.
(4) Neural network is harder to tune than support vector machine.
(5) Support vector machine is important for regression.
A.) equally well
B.) vector machine
C.) support machine
D.) support vector

T_id Items Bought
10 Beer, Nuts, Diapers
20 Beer, Coffee, Diapers, Nuts 30
Beer, Diapers, Eggs
40 Beer, Nuts, Eggs, Milk
50 Nuts, Coffee, Diapers,
Eggs, Milk

Given the transaction in table 1 and mini-support (minsup) s = 40%, which of the following is a length-3 frequent item set?
A.) Beer, Nuts, Diapers
B.) Beer, Nuts, Eggs
C.) Beer, Coffee, Milk
D.) Coffee, Diapers, Eggs

Beer, Nuts, Diapers

20 Beer, Coffee, Diapers, Nuts 30 Beer, Diapers, Eggs 40 Beer, Nuts, Eggs, Milk 50 Nuts, Coffee, Diapers, Eggs, Milk Given the transactions in table 1, what is the confidence and relative support of the association rule {Diapers} ⇒ {Coffee, Nuts}?	
TNG (topical N-Grams)	-generalization of Bigram topic model -probabilistic generative model that generates words in textual order -create n-grams by concatenating successive bigrams -conditions on previous word and topic when drawing next word
topic-modeling based phase mining	first topic modeling then phrase construction
TOPMine	mining quality phrases without training data
training based phrase mining	-bag of words, n-grams, phrases -use topic modeling (represent documents by multiple topics in different proportions) -tends to overfitting and can be slow
TurboTopic	-phrase construction as a post processing step to LDA -merge adjacent unigrams with the same topic label by a distribution free permutation test on arbitrary length back off model
What is the support of the following graph? A.) 0 B.) 1 C.) 2 D.) 3	D.) 3  base, we have the following 3 graphs. support of the following graph? treat all nodes as the same and consider the topological structure.

A.) [0, 1] B.) [0, +∞) C.) (-∞, +∞) D.) [-1, 1]	
When we use the Apriori-based approach to find the frequent graph pattern for a candidate graph, we need to check all its subgraphs. Suppose only connected subgraphs are considered, given the following graph, how many distinct subgraphs with 3-vertices are there? A.) 1 B.) 2 C.) 3 D.) 4	2 
Which of the following measures has been used for ranking phrases in KERT? Select all that apply. A.) Completeness B.) KL divergence C.) Concordance D.) Popularity	A.) Completeness C.) Concordance D.) Popularity
Which of the following measures is NOT null invariant? A.) Cosine B.) All confidence C.) Kulcyzynski D.) Lift	Lift

A.) We can recover all frequent patterns and their supports from the set of closed frequent patterns. B.) We can recover all frequent patterns and their supports from the set of max frequent patterns. C.) The set of closed frequent patterns is always the same as the set of max frequent patterns. D.) Since both closed and max frequent patterns are a subset of all frequent patterns, we cannot recover all frequent patterns and their supports given just the closed and max frequent patterns. E.) Closed frequent patterns can always be determined from the set of max frequent patterns.	
word colocation	a sequence of words that occur more frequently than expected