UNIVERSIDADE ESTADUAL DE CAMPINAS
Curso de Aperfeiçoamento – Fundamentos de Minera¸c˜ao de Dados INF-1020 Fundamentos de Minera¸c˜ao de Dados
Prof. Julio Cesar dos Reis

LISTA - Regras de Associação

Nome Aluno 1: Diego da Silva Coimbra

RA Aluno 1: 14358389-12

Nome Aluno 2: Nara Miranda Guimarães

RA Aluno 2: 44662603-x

Instrução: Esta lista de exercícios deve ser resolvida em dupla. A nota será no intervalo [0, 10].

Questão	Valor	Nota
1	1,50	
2	1,50	
3	1,50	
4	1,50	
5	2,00	
6	2,00	
Total	10,0	

1. O que são as regras de associação?

As regras de Associação têm como premissa básica encontrar elementos que implicam na presença de outros elementos em uma mesma transação, ou seja,

encontrar relacionamentos ou padrões frequentes entre conjuntos de dados. O termo transação indica quais itens foram consultados em uma determinada operação de consulta. Em outras palavras, são usadas para descobrir elementos que ocorrem em comum dentro de um determinado conjunto de dados.

2. Identifique e descreva um problema de mineração em que a tarefa baseada em Regras de Associação possa ser utilizada.

Analisar os hábitos de compra de clientes por meio da descoberta de associações entre diferentes itens que aparecem nas “cestas de compras”. A descoberta destas associações pode ajudar os varejistas no desenvolvimento de estratégias de marketing já que revelam quais itens são frequentemente comprados juntos pelos clientes.

Um exemplo seria um conjunto de itens disponíveis em uma loja, a cada item podemos associar uma variável booleana que representa a presença ou ausência daquele item em um evento. Assim, cada “compra” (ou transação) pode ser representada por um vetor booleano de valores associados a estas variáveis. Os vetores booleanos, então, podem ser analisados como padrões de compras que refletem itens que são frequentemente associados ou comprados juntos. Esses padrões podem ser representados na forma de regras de associação.

3. Explique a diferença entre as métricas de suporte e de confiança na geração de regras de associação.

Métrica de Suporte: medida que indica a proporção de um conjunto **X** em uma base de dados **D**. Ou seja, a frequência de vezes que uma determinada regra é aplicável ao conjunto de dados.

$$Supp(X) = (\#X \text{ in } D) / (\#D)$$

Métrica de Confiança: definida sobre a regra (X=>Y). Em outras palavras, é a frequência na qual os elementos de Y aparecem no conjunto de dados com transações que possuem X.

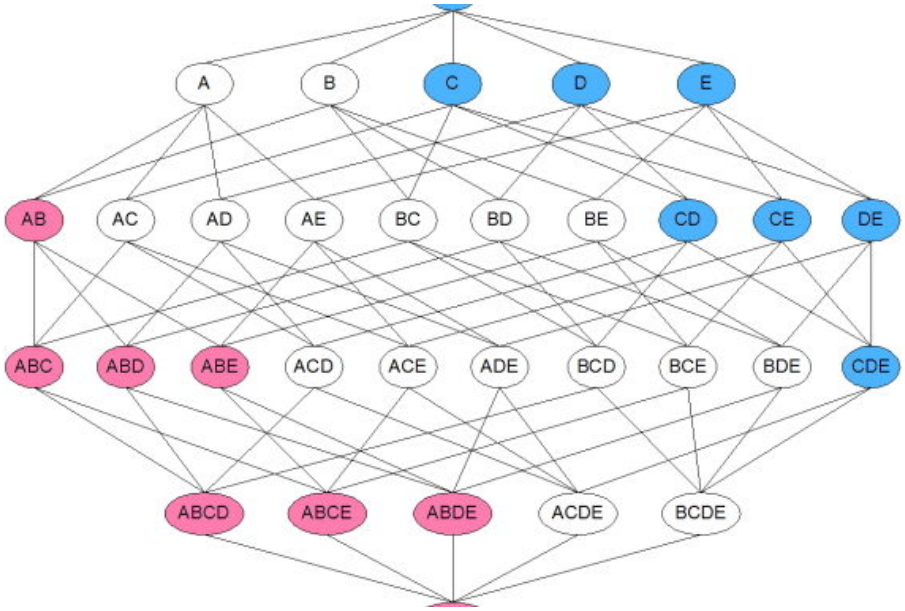
$$Conf(X=>Y) = Supp(X \text{ union } Y) / Supp(X)$$

4. Explique o algoritmo Apriori. Em que etapa da mineração de regras de associa¸c˜ao ele é importante e por quˆe?

O Apriori trabalha com o conceito de Itens frequentes, que são os Itens do seu conjunto I que têm a pontuação do **Suporte** mais que um threshold (hiperparâmetro). Ou seja, precisamos calcular o **Suporte** de todas as combinações de Itens (I) e extrair um subconjunto de Itens frequentes (I_freq). Sabendo disso, os passos do Apriori são:

A. Dentro dos Itens I, extraia um subconjunto (I_freq) dos itens que tem o seu Suporte maior que o threshold.
B. Dentro de I_freq, itere para formar as combinações dos I_freq com I, aplique o threshold e acumule em I_freq.
C. Pare quando ao aplicar o threshold nenhum item sobrar.

Na mineração de regras de associação, o Apriori é utilizado para reduzir o número de candidatos (M).



A cor azul destaca os nós que foram considerados frequentes, enquanto que em rosa estão os nós dos itens que foram desconsiderados para a criação das regras.

5. Considere as seguintes transações:

Cliente ID	Transação ID	Itens Comprados
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

(a) Calcule o suporte para conjuntos de itens *{e}*, *{*b, d*}* e *{*b, d, e*}*, tratando cada ID de transação como uma cesta de compras.

(b) Use os resultados da parte (a) para calcular a confiança para as regras *{b, d}* → *{e}* e *{e}* → *{b, d}*.

	a	b	c	d	e
1	1	0	0	1	1
24	1	1	1	0	1
12	1	1	0	1	1
31	1	0	1	1	1
15	0	1	1	0	1
22	0	1	0	1	1
29	0	0	1	1	0

40	1	1	1	0	0
33	1	0	0	1	1
38	1	1	0	0	1

a) O suporte determina com que frequência uma regra é aplicável a um determinado conjunto de dados
Suporte (s): Fração de transações que contém um itemset.
Support count (σ): Frequência da ocorrência de um itemset.

{e} - 8/10 = 0,8
{b, d} - 2/10 = 0,2
{b, d, e} - 1/10 = 0,1

b) A confiança determina com que frequência os itens em Y aparecem nas transações que contêm X
Confiança (c): Mede a frequência de itens em Y que aparece nas transações que contêm X. $c(X \Rightarrow Y) = s(X \cup Y) / s(X)$

{b, d} → *{e}*
X = *{b, d}*
Y = *{e}*
$c(X \Rightarrow Y) = s(X \cup Y) / s(X) = 0,1/0,2 = 0,5$
50% das vezes que **b e d** são comprados, então **e** é comprado.

{e} → *{b, d}*
X = *{e}*
Y = *{b, d}*
$c(X \Rightarrow Y) = s(X \cup Y) / s(X) = 0,1/0,8 = 0,125$
12,5% das vezes que **e** é comprado, então **b e d** são comprados.

6. Considerando o arquivo de transações abaixo:

id	Itens comprados
1	A,B,C
2	C,D,E,F
3	A,G,H
4	E,H,I,J
5	B,C
6	A,B,C,D,F
7	D,E,F
8	G,H
9	A,C,J,K
10	A,B,C,D

Figura 1: Transações de compras.

(a) Encontre todas as regras de associação com suporte mínimo 40% e confiança mínima 80%.

Para ter 40% de suporte mínimo, o itemset precisa aparecer no mínimo 4 vezes

	A	B	C	D	E	F	G	H	I	J	K
1	1	1	1								
2			1	1	1	1					
3	1						1	1			

4					1			1	1	1	
5		1	1								
6	1	1	1	1		1					
7				1	1	1					
8							1	1			
9	1		1							1	1
10	1	1	1	1							

{A}, {B}, {C}, {D}, {AC}, {BC}

Sendo o valor de suporte para cada é igual a:

{A} = 5/10 = 0,5

{B} = 4/10 = 0,4

{C} = 6/10 = 0,6

{D} = 4/10 = 0,4

{AC} = 4/10 = 0,4

{BC} = 4/10 = 0,4

Cálculo da confiança:

{A} => {C} = 0,4/0,5 = 0,8

{C} => {A} = 0,4/0,6 = 0,67

{B} => {C} = 0,4/0,4 = 1

{C} => {B} - 0,4/0,6 = 0,67

Resposta: apenas as reagras de associação entre {A} => {C} e {B} => {C} possuem suporte maior do que 40% e confiança maior que 80%

(b) Calcule a medida lift para as duas regras encontradas de maior confiança.

lift: Mede o quanto os 2 lados da regra X ⇒ Y são dependentes
lift = 1 indica que X e Y são independentes.
lift(X ⇒ Y) = s(X ∪ Y) / (s(X) * s(Y))

Regras com maior confiança:

{A} => {C} = 0,4/0,5 = 0,8

lift = 0,4/(0,5x0,6) = 1,33

Como o valor resultante é maior do que 1, então A e C são positivamente correlacionados, significando que a ocorrência de um implica na ocorrência de outro.

C é provável de ser comprado quando A for comprado

{B} => {C} = 0,4/0,4 = 1

lift = 0,4/(0,4x0,6) = 1,67

Como o valor resultante é maior do que 1, então B e C são positivamente correlacionados, significando que a ocorrência de um implica na ocorrência de outro.

C é provável de ser comprado quando B for comprado