UNIVERSIDAD NACIONAL DE COLOMBIA

MINERIA DE DATOS

TALLER: Asociación

PROFESORA ELIZABETH LEON GUZMAN

1. Por cada un de las siguientes preguntas, proveer un ejemplo de una regla de asociación del dominio de "market basket" que satisface las siguientes condiciones. También, describir si las reglas son interesantes (subjetivamente).

   (a) Una regla que que tiene alto soporte y alta confianza

   (b) Una regla que tiene razonablemente alto soporte pero baja confianza

   (c) Una regla que tiene bajo soporte y baja confianza

   (d) Una regla que tiene bajo soporte y alta confianza

2. ¿Por qué el proceso de descubrimiento de reglas de asociación es relativamente simple comparado con la generación de grandes conjuntos de ítems en bases de datos transaccionales?

3. Considere el siguiente conjunto de datos:

TID	Ítems
T01	milk, beer, diapers
T02	bread, butter, milk
T03	milk, diapers, cookies
T04	bread, butter, cookies
T05	beer, cookies, diapers
T06	milk, diapers, bread, butter
T07	bread, butter, diapers
T08	beer, diapers
T09	milk, diapers, bread, butter
T10	beer, cookies

   (a) ¿cuál es el número máximo de reglas de asociación que se pueden generar? (incluyendo reglas con soporte 0)

   (b) ¿cuál es el tamaño máximo de los conjuntos de items frecuentes que se pueden extraer (asumir el umbral de minsoporte > 0)?

   (c) Escribir una regla que contenga 3 ítems que se genere deeste conjunto de datos.

   (d) Encontrat un conjunto de items (de tamaño mayor a 2) con el valor de soporte máximo.

   (e) Encontrar un par de ítems (a y b) tal que las reglas $a \rightarrow b$ y $b \rightarrow a$ tengan la misma confianza.

4. Dado la siguiente base de datos X:

TID	Ítems
T01	A, B, C, D
T02	A, C, D, F
T03	C, D, E, G, A
T04	A, D, F, B
T05	B, C, G
T06	D, F, G
T07	A, B, G
T08	C, D, F, G

Usando valores de umbral de soporte = 25 % y confianza = 60 %, encuentre:

a) Todos los conjuntos de ítems en la base de datos X

b) Reglas de asociación fuertes para la base de datos X

c) Analice las asociaciones engañosas para el conjunto de reglas obtenido en el numeral anterior.

5. El algoritmo *Apriori* usa estrategías de generación y conteo para derivar conjuntos de *items* frecuentes. Conjuntos de item de tamaño $k+1$ son creados de conjuntos de items de tamaño k. Un conjunto candidato es eliminado si uno de sus subconjuntos no es frecuente en la fase de poda. Supongamos que el algoritmo *Apriori* es aplicado a los datos de la siguiente tabla con un soporte mínimo de 30% (ejercicio realizado en clase).

id	items
1	a,b,d,e
2	b,c,d
3	a,b,d,e
4	a,c,d,e
5	b,c,d,e
6	b,d,e
7	c,d
8	a,b,c
9	a,d,e
10	b,d

   (a) Dibujar el lattice de los conjuntos de items y etiquetar cada nodo con las siguientes letras:
      i. **N**: si el conjunto no es considerado candidato
      ii. **F**: si el conjunto candidato es frecuente
      iii. **I**: si el conjunto candidato no es frecuente

   (b) ¿Cuál es el porcentaje de conjuntos de items frecuente?

   (c) ¿Cuál es el radio de poda en este conjunto de datos? (El radio de poda es definido como el porcentaje de conjuntos de items no considerados a ser candidatos, ya sea por que no son generados durante la etapa de generación de candidatos, o por que son podados en la etapa de poda).

   (d) ¿Cuál es la rata de falsa alarma? (porcentaje de los conjuntos de items candidatos que son encontrados NO frecuentes después de calcular el soporte).

6. Dada la siguiente base de datos transaccional Y:

TID	Ítems
1	A,B,C
2	A,C,D,E
3	A,B,D
4	A,C,F
5	A,B
6	A,E,F
7	A,B,D,E,F
8	A,F
9	B,D,E
10	B,D,E,F
11	B,C,D,E
12	C,D,E

   Usar *Apriori* y *FP-Growth* para encontrar los conjuntos de ítems frecuentes con mínimo soporte de 2. Tratar con varios valores de confianza. Ordenar los conjuntos resultados y analizar resultados de cada algoritmo. Comparar eficiencia de los dos procesos de mineria. Repetir con soporte de 3 y comparar los resultados.

7. Usando el conjunto de datos del punto 5,

   (a) realizar la tabla de contingencia para las siguientes reglas: $b \rightarrow$c , $a \rightarrow$d , b$\rightarrow$d , e$\rightarrow$c , c$\rightarrow$a

(b) usar las tablas de contigencia del punto anterior para computar y "realizar un ranking" de las reglas usando:

i. soporte

ii. confianza

iii. lift

8. Importar el conjunto de datos marketBasquet.csv. Extraer reglas de asociación utilizando diferentes valores mínimos de confíanza. Analice los resultados. Reporte los itemsets frecuentes con sus respectivos valores de soporte. Revise los valores de soporte reportados por el operador de itemsets frecuentes. Verifíquelos manualmente. ¿Qué problema se presenta? Extraiga reglas de asociación utilizando diferentes valores mínimos de confíanza. Analice los resultados. Reporte los itemsets frecuentes con sus respectivos valores de soporte. Analice los resultados y discuta las diferencias con el esquema que no posee un filtro de atributos.

9. Importe el conjunto de datos credit-german.csv (repositorio de machinelearning). Discretice los atributos numéricos en máximo 5 bins de igual tamaño. Aplique el algoritmo de reglas de asociación a este conjunto. Interprete las reglas producidas. Varíe los valores de soporte y de confianza ¿Qué sucede? Interprete las reglas producidas y escoja las que en su concepto son las más interesantes para el problema (justifíque).