Université Sultan Moulay Slimane
Ecole National des Sciences Appliquées
Khouribga

Module : Data Mining & Texte Mining

1ère Année Master Big Data & Aide à la Décision

Semestre 2 / Année 2018/2019 / Feuille de TD N° 3

Exercice 1

Considérons le dataset du tableau suivant :

Customer ID	Transaction ID	Items Bought
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	$\{a, b, e\}$

1. Calculer le support de chacun des itemsets suivants : $\{e\}$, $\{b, d\}$, et $\{b, d, e\}$, en traitant chaque transaction ID comme un panier de ménagère.
2. Utiliser les résultats calculés dans la question (1) pour calculer la confiance de chacune des règles d'association suivantes : $\{b, d\} \rightarrow \{e\}$ et $\{e\} \rightarrow \{b, d\}$. La confiance est-elle une mesure symétrique ou asymétrique ?
3. Calculer le support de chacun des itemsets suivants : $\{e\}$, $\{b, d\}$, et $\{b, d, e\}$, en traitant chaque client ID comme un panier de ménagère (chaque item doit être traité comme une variable binaire qui vaut 1, si l'item apparaît dans au moins une transaction payée par le client, et 0 sinon).
4. Utiliser les résultats calculés dans la question (3) pour calculer la confiance de chacune des règles d'association suivantes : $\{b, d\} \rightarrow \{e\}$ et $\{e\} \rightarrow \{b, d\}$.
5. Supposons que s_1 and c_1 sont respectivement le support et la confiance d'une règle d'association r lorsqu'on traite chaque transaction ID comme un panier de ménagère. De même, soient s_2 and c_2 respectivement le support et la confiance d'une règle d'association r lorsqu'on traite chaque client ID comme un panier de ménagère. Existe-t-elle une relation entre s_1 et s_2 ? ou c_1 et c_2 ?

Exercice 2

1. Quelles sont les confiances des deux règles : $\varnothing \rightarrow A$ et $A \rightarrow \varnothing$?
2. Soient c_1, c_2 et c_3 les valeurs des confiances des règles $\{p\} \rightarrow \{q\}$, $\{p\} \rightarrow \{q, r\}$ et $\{p, r\} \rightarrow \{q\}$ respectivement. Si l'on suppose que c_1, c_2 et c_3 ont des valeurs différentes, quelles sont les relations possibles qui peuvent exister entre c_1, c_2 et c_3? Quelle est la règle ayant la plus petite confiance ?
3. Répondre à la question précédente en supposant que les trios règles ont le même support. Quelle est la règle ayant la plus grande confiance ?
4. Transitivité : supposons que les confiances respectives des deux règles $A \rightarrow B$ et $B \rightarrow C$ sont plus grandes au seuil de confiance ***min_conf***. Est-t-il possible que la règle $A \rightarrow C$ aura une confiance moins de ***min_conf*** ?

Exercice 3

Montrer que le nombre total R de règles d'association extraites à partir d'un dataset contenant d items est :

$$R = 3^d - 2^{d+1} + 1$$

Exercice 4

Considérons les transactions d'un panier de la ménagère illustrées par le tableau suivant :

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

1. Quel est le nombre maximum de règles d'association qui peuvent être extraites à partir de ce dataset ? (y compris les règles de support nul)
2. Quelle est la taille maximale d'un itemset fréquent qui peut être extrait à partir de ce dataset ? (on suppose que ***min_sup*** > 0)
3. Quel est le nombre maximum des 3-itemsets qui peuvent être dérivés à partir de ce dataset ?
4. Trouver un itemset de taille ≥ 2 de support maximal.

5. Trouver une paire d'items a et b, telle que les règles $\{a\} \rightarrow \{b\}$ et $\{b\} \rightarrow \{a\}$ soient de même confiance.

Exercice 5

Dans une étape d'identification d'itemsets fréquents dans une base de données transactionnelle, nous avons trouvés que les 3-itemsets fréquents sont : $\{B, D, E\}$, $\{C, E, F\}$, $\{B, C, D\}$, $\{A, B, E\}$, $\{D, E, F\}$, $\{A, C, F\}$, $\{A, C, E\}$, $\{A, B, C\}$, $\{A, C, D\}$, $\{C, D, E\}$, $\{C, D, F\}$, $\{A, D, E\}$. Lesquels des 4-itemsets suivants peut être probablement fréquent ?

a) $\{A, B, C, D\}$

b) $\{A, B, D, E\}$

c) $\{A, C, E, F\}$

d) $\{C, D, E, F\}$

Exercice 6

Supposons que l'algorithme Apriori est appliqué au dataset du tableau suivant :

Transaction ID	Items Bought
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	$\{b, d\}$

On suppose que le seuil de support est ***min_sup*** = 30%.

1. Dessiner le treillis d'itemsets engendré par l'exécution de l'algorithme Apriori. Etiqueter chaque nœud du treillis avec les lettres suivantes :
   - **N** : si l'itemset est non considéré à être un itemset candidat par l'algorithme Apriori. Il y a deux raisons pour qu'un itemset soit non considéré comme un itemset candidat : (*i*) il est non généré pendant toutes les étapes de génération de candidats, ou bien (*ii*) il est généré pendant une étape de génération de candidats, mais il est supprimé pendant l'étape de l'élagage de candidats car l'un de ses sous-ensembles est non fréquent.
   - **F** : si l'itemset candidat est prouvé fréquent par l'algorithme Apriori.

- **I** : si l'itemset candidat est prouvé non fréquent après le calcul de son support.

2. Quel est le pourcentage des itemsets fréquents ?
3. Quel est le rapport d'élagage de l'algorithme Apriori sur ce dataset ?

Exercice 7

Considérons, pour une règle d'association A → B, la mesure d'intérêt suivant :

$$M = \frac{P(B|A) - P(B)}{1 - P(B)}$$

1. Quel est l'intervalle de valeurs de la mesure M ? Préciser quand-est-ce que M atteint sa valeur maximale et sa valeur minimale.
2. Qu'arrive-t-il à M lorsque :
   a. $P(A, B)$ croît alors que $P(A)$ et $P(B)$ restent inchangées ?
   b. $P(A)$ croît alors que $P(A, B)$ et $P(B)$ restent inchangées ?
   c. $P(B)$ croît tandis que $P(A, B)$ et $P(A)$ restent inchangées ?
3. La mesure M est-elle symétrique sous permutation de variables ?
4. Que vaut M lorsque A et B sont statistiquement indépendants ?

Exercice 8

Supposons que l'on dispose des données d'un panier de ménagère consistant de 100 transactions avec 20 items. Si le support d'un item a est de 25%, le support d'un item b est de 90% et le support de l'itemset $\{a, b\}$ est de 20%. Les seuils de support et de la confiance sont 10% et 60%, respectivement.

1. Calculer la confiance de la règle d'association $\{a\} \rightarrow \{b\}$. Est-elle intéressante selon la mesure de la confiance ?
2. Calculer la mesure de l'intérêt (***lift***) de la règle d'association $\{a\} \rightarrow \{b\}$. Décrire la nature de relation entre l'item a et l'item b en termes de la mesure de l'intérêt.
3. Quelle est la conclusion que l'on peut tirer à partir des résultats précédents ?
4. Montrer que si la valeur de la confiance de la règle $\{a\} \rightarrow \{b\}$ est plus petite que celle du support de l'item b, alors :
   a. ***confiance***$(\{\bar{a}\} \rightarrow \{b\}) >$ ***confiance***$(\{a\} \rightarrow \{b\})$
   b. ***confiance***$(\{\bar{a}\} \rightarrow \{b\}) >$ ***support***$(\{b\})$.