Web Mining – Règles d'association

A. Analyse des clients d'une banque

Nous travaillons sous le logiciel « Association Rule » (SIPINA) et Excel dans ce premier exercice.

Nous traitons le fichier « **clients banque.xlsx** ». Il décrit les caractéristiques et le comportement des clients d'une banque quelconque. Les variables sont subdivisées en 3 groupes :

a. Fond vert : les caractéristiques du client et son comportement au sein de l'établissement (âge, profession, domiciliation de l'épargne, moyenne des encours, etc.) ;
b. Fond orange : les décisions de la banque sur des services clés (autorisation de découvert, interdiction de chéquier) ;
c. Fond bleu : l'appréciation de la banque vis-à-vis du client (type du client : appréciation du chargé de clientèle ; note : score attribué par le système informatique).

ANALYSER LES ASSOCIATIONS ENTRE LES DIFFERENTS ITEMS DE LA BASE CLIENTELE.

1. Démarrer Excel et chargez le fichier de données.
2. Installez l'add-in SIPINA dans Excel (https://www.youtube.com/watch?v=upqdj6a48Bw).
   a. Remarque 1 : Attention, pour ceux qui travaillent sur les machines de l'université, le logiciel SIPINA est déjà présent dans notre salle informatique, il n'est pas nécessaire d'aller le chercher sur le web et de l'installer sur la machine. Seule l'intégration de l'add-in dans Excel est nécessaire.
   b. Remarque 2 : Dans notre salle machine – tout dépend du niveau de protection qui a été mis en place - l'add-in ne fonctionne pas parfois. Dans ce cas, le plus simple est d'exporter le fichier Excel en fichier « **Texte (séparateur : tabulation) (*.txt)** ». Fermez Excel ensuite parce qu'il a la fâcheuse tendance à verrouiller les fichiers. Puis, après avoir lancé « Association Rule » via la zone de recherche dans la barre des tâches Windows, importez le fichier avec FILE / OPEN [Text file format (*.txt)].
3. Sélectionnez la plage de données (à l'exception de la colonne « note attribuée ») et lancez le logiciel ARS « Association Rule Software » (https://www.youtube.com/watch?v=fpbdjEaGcLw).
4. Sélectionnez l'ensemble des variables dans ARS.
5. Lancez la construction des règles avec les paramètres suivants : support minimum = 0.2, confiance minimum = 0.8, longueur maximum des règles = 4 items, nombre maximum d'items dans le conséquent = 1. Combien de règles obtenez-vous ? (199 règles)
6. Copiez la base de règles dans Excel, convertissez-la en tableau (INSERTION / TABLEAU).

7. Triez les règles selon le LIFT de manière décroissante (cf. « Mesures d'intérêt des règles dans A PRIORI MR » -- http://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#9047606199815096212 pour les formules sous-jacentes aux mesures d'intérêt des règles). Quelle est la règle la plus intéressante au sens du LIFT ? (LIFT = 1.6756)
8. Sous Excel, calculez les critères SURPRISE et IMPORTANCE de cette règle (Remarque : RECALL correspond à la sensibilité de la règle). (SURPRISE = 0.3766 ; IMPORTANCE = 0.8284)
9. Isolez les règles comportant « type de client » dans le conséquent (2 règles).
10. Quelles sont les règles qui contiennent « autorisation de découvert » ET « interdiction de chéquier » dans l'antécédent (Attention, les accents ont été supprimés de la base de données) (6 règles)

ON SOUHAITE METTRE LE FOCUS SUR LES PERSONNES « TYPE DE CLIENT = MAUVAIS CLIENT ».

11. Comment modifierez-vous les paramètres de l'algorithme pour que ce conséquent apparaisse parmi les règles générées ? (Abaisser le support min, par ex. 0.02). Lancez la méthode avec ces nouveaux paramètres. Filtrez les règles de manière à n'afficher que celles avec le conséquent « Type de client = mauvais client ».
12. Est-ce que l'interdiction de chéquier ou l'interdiction de découvert sont associables à la qualification « mauvais client » ? (Utiliser un filtre personnalisé sur l'antécédent sous Excel)
13. (Pas forcément via les règles d'association) Comparez les proportions de bons et mauvais clients parmi les personnes interdits de chéquier. Constat ? (Les mauvais clients sont sur-représentés)
14. (Pas forcément via les règles d'association) Comparez les notes moyennes des clients selon qu'ils sont interdits de chéquier ou non. (Note moyenne basse ➔ chéquier interdit ; et inversement…)

B. Règles d'association sous Python

Nous travaillons sous Python avec le package « mlxtend » (https://rasbt.github.io/mlxtend/) pour cet exercice. Nous traitons toujours le fichier « **clients banque.xlsx** ».

[**TUTO 1**] https://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#7596337565513783851

1. Vérifiez la version de « mlxtend » (0.23.4 pour moi -- https://pypi.org/project/mlxtend/).
2. Chargez le fichier de données dans un data.frame. Vérifiez les informations et affichez les premières lignes.
3. Créez un second data frame sans la colonne « Note attribuée ».
4. Transformez-le en tableau de booléens en y faisant figurer toutes les modalités (voir les options de pandas.get_dummies). Utilisez un séparateur adéquat pour qu'il soit facile de distinguer les noms de variables des désignations des modalités. Affichez les informations et les premières lignes de ce nouveau tableau.

5. A titre de vérification : calculez dans le data frame initial et dans le tableau de booléens le nombre de personnes « Age_du_client = de_23_a_40_ans ». Les valeurs coïncident-elles ? (oui, heureusement, **150**).

On s'interesse aux itemsets frequents.

6. Modifiez les options de pandas pour dépasser la limitation du nombre de colonnes à afficher [`pandas.set_option('display.max_colwidth',None)`]
7. A l'aide de la procédure apriori, affichez les itemsets de (card ≤ 2) [max_len = 2] avec un support supérieur ou égal à 0.5 ([**TUTO 1**], section 3) (12 itemsets).
8. Quel est le type d'objet que nous obtenons ? (un data frame Pandas, on sait manipuler)
9. Vous remarquerez qu'il y a des itemsets de (card = 1) dans les résultats. Ils ne nous intéressent pas. Filtrez les résultats de manière à ce que seuls ceux avec (card = 2) apparaissent (cf. « Example 2 » sur https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/ pour voir comment filtrer les itemsets) (il y en a 6)
10. On s'intéresse à l'itemset « Autorisation de découvert = découvert interdit ET Interdiction de chéquier = chéquier autorisé »). Sans passer par MLXTEND, vérifiez que le support annoncé par la procédure apriori correspond bien à ce que l'on observe dans les données (248 / 468 est bien égal à 0.5299)
11. Parmi les itemsets de (card = 2), affichez ceux pour lesquels le terme « Interdiction_de_chequier » apparaît (4 itemsets, *nous avons un objet non conventionnel dans la colonne « itemsets », il faut savoir le transformer pour recherche la chaîne de caractère qui nous intéresse*).
12. Sans passer par MLXTEND, calculez la proportion de bons et de mauvais clients chez les différentes configurations de « Autorisation de découvert » ET « Interdiction de chéquier ». Quelle est la configuration où le pourcentage de bons clients est le plus élevé ? Quelle est la configuration où le pourcentage de mauvais clients est le plus élevé ? (le plus favorable : découvert interdit et chéquier autorisé (62.09%) ; le plus défavorable : découvert interdit et chéquier interdit (0.00%))
13. Sans passer par MLXTEND, pour les mêmes configurations, calculez les notes moyennes. Quelle est la configuration la mieux (la moins bien) notée par la banque ? Les résultats sont-ils cohérents avec ceux de la question précédente ? (oui, on s'y attend mais c'est mieux lorsque les données les confirment)

On s'interesse aux regles d'association.

14. Générez des règles d'association avec un support minimum de 0.2, une confiance min de 0.8, et un cardinal d'itemsets maximum = 4 ([**TUTO 1**], section 4 ; attention, MLXTEND nous impose de

travailler en 2 temps, générer d'abord les itemsets fréquents avec les paramètres idoines [245 itemsets fréquents], en déduire les règles en appliquant les paramètres dédiés [220 règles]). Quel est le type de l'objet obtenu (data frame) ? Combien de règles observez-vous ? (220 règles donc)

15. Affichez la liste des colonnes. Affichez les premières règles (head). Que constatez-vous ? (plusieurs mesures d'évaluation des règles sont proposées).
16. Simplifiez la représentation de manière à ne conserver des règles : l'antécédent, le conséquent, le support, la confiance, le lift.
17. L'outil ne propose pas de paramètre pour limiter le cardinal des itemsets composant le conséquent. Y a-t-il des règles qui ont un conséquent avec (card ≥ 2), combien (22), lesquelles ?
18. Affichez les 2 règles présentant le lift le plus élevé (quelle que soit la composition des antécédents et conséquents). Que remarquons-nous ? (concernent le type de client dans le conséquent)
19. Affichez les règles contenant le terme « mauvais client », que ce soit dans l'antécédent ou le conséquent. Combien de règles correspondent à cette spécification ? (14 règles)

C. Analyse des préférences de films

Nous travaillons sous R (RStudio) pour cet exercice.

[**TUTO 2**] https://tutoriels-data-science.blogspot.com/p/tutoriels-en-francais.html#1911930121280889267

(en particulier à partir de la page 8)

Notation des films – Associations des films

1. Le fichier « **users_ratings.txt** » recense les notes attribuées par 943 cinéphiles à 1664 films. En voici les premières lignes :

```
user   item   rating
1      Toy Story (1995)    5
1      GoldenEye (1995)    3
1      Four Rooms (1995)   4
1      Get Shorty (1995)   3
Etc…
```

2. Chargez le fichier de données, affichez le nombre d'évaluation disponibles dans le fichier, les numéros d'utilisateurs distincts, les noms des films.
3. Calculez la note moyenne attribuée par chaque utilisateur (1 : 3.60, 2 : 3.70, 3 : 2.76, etc.)
4. On considère que l'utilisateur a noté un film parce qu'il l'a vu. A partir de ces données, on cherche à extraire des règles permettant de proposer des recommandations du type : un utilisateur qui a regardé tel film a également regardé tel autre film. Produisez des règles d'association [**TUTO 2**, page 8] à partir de ces données avec : un support minimum de 0.3, une confiance min de 0.8, une

longueur minimum de 2, une longueur maximum de 2 et **target = 'rules'**. Combien de règles obtenez-vous ? (29) (pour ma part, j'ai créé un tableau de booléens [943 lignes x 1664 colonnes] permettant d'identifier les films visionnés par les personnes, que j'ai soumis ensuite aux fonctions du package « arules »).

5. Quels sont les films associés à 'Mission: Impossible' ? (grep) (2 films)
6. Triez les règles de manière décroissante selon la confiance. Quelle est la règle présentant la confiance la plus élevée ? ({Return of the Jedi} ➔ {Star Wars}, confiance = 0.9467456).

GENRE DES FILMS

7. Le fichier « **movies_types.txt** » associe aux films leur genre (ex. policier, mélo, etc.). Attention, un film peut appartenir à plusieurs genres.

title	unknown	Action	Adventure	Animation	Children	Comedy	Crime	Documentary	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFiction	Thriller	War	Western
Toy Story (1995)	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
GoldenEye (1995)	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Four Rooms (1995)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Get Shorty (1995)	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
Copycat (1995)	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0

8. Affichez la liste des genres (unknown, action, adventure, etc.)
9. Affichez les 5 genres les plus représentés parmi les films (drama : 716, comedy : 502, etc.).
10. Affichez les associations de genres les plus fréquentes dans les films (support min = 0.05) ({comédie, romance}, {drame, romance}, {action, thriller}, {comédie, drame}).

GOUTS DES CINEPHILES SELON LE GENRE

Pour élaborer un système de recommandation efficace, nous souhaitons établir les préférences des personnes en matière d'association de genres de films. L'idée est de pouvoir établir des règles du type : si on aime le western alors on aime l'action également ; etc.

11. Calculez la note moyenne attribuée par chaque utilisateur pour chaque genre. Prenons un exemple simple : l'utilisateur n°1 a vu 13 films que l'on peut classer parmi les films d'horreur, il leur a attribué les notes suivantes

user	item	rating	Horror
1	From Dusk Till Dawn (1996)	3	1
1	Robert A. Heinlein s The Puppet Masters (1994)	4	1
1	Heavy Metal (1981)	2	1
1	Frighteners, The (1996)	4	1
1	Alien (1979)	5	1
1	Army of Darkness (1993)	4	1
1	Psycho (1960)	4	1
1	Shining, The (1980)	3	1
1	Evil Dead II (1987)	3	1
1	Young Frankenstein (1974)	5	1
1	Bram Stoker s Dracula (1992)	3	1
1	Nightmare on Elm Street, A (1984)	1	1
1	Jaws (1975)	4	1

La note moyenne qu'il a attribué aux films d'horreur sera donc égale à (3 + 4 + 2 + …) / 13 = 3.46

Vous devez donc construire un tableau dont les premières lignes seraient les suivantes :

user	unknown	Action	Adventure	Animation	Children	Comedy	Crime	Documentary	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFiction	Thriller	War	Western
1	4.0	3.3	2.9	3.3	2.2	3.5	3.4	4.8	3.9	3.5	5.0	3.5	2.9	3.6	3.9	4.0	3.6	3.7	3.7
2	0.0	3.8	4.3	4.0	3.0	3.8	3.8	0.0	3.8	3.0	4.5	3.0	3.0	3.5	4.1	3.8	3.6	3.7	0.0
3	0.0	2.8	3.5	0.0	0.0	2.6	2.9	5.0	2.8	0.0	2.5	2.4	2.0	3.2	3.5	2.8	2.5	2.8	0.0

Nous mettons la note 0 lorsque l'utilisateur n'a vu aucun film du genre étudié.

Remarque : Regardez du côté des requêtes SQL avec le package « sqldf » ; peut-être aussi la fonction merge() de R qui permet d'effectuer des jointures entre tables ; enfin, le package ''dplyr'' serait également une piste avec les requêtes avec jointures. La solution n'est pas triviale quoiqu'il en soit.

12. A partir de ce nouveau tableau, calculer les notes moyennes (***non pondérées***) attribuée à chaque genre. Quels sont les genres les plus appréciés ? (Drama : 3.72, War : 3.70, Romance : 3.656, etc.)
13. Transformez cette matrice en un tableau de préférences booléen qui prend la valeur TRUE lorsque la note moyenne est >= 4, FALSE sinon (le tableau possède 943 lignes et 19 colonnes normalement)
14. Construisez les règles d'association de cardinal max = 2, avec un support min de 0.1 et une confiance min de 0.7. Triez les règles de manière décroissante selon le lift. Quelles sont les 5 règles les plus intéressantes au regard du LIFT ? ({Action} ➔ {Adventure}, {Comedy} ➔ {Romance}, etc.).