TP 2 : Fouille d’itemsets fréquents et de règles d’association sous Python

IMT Atlantique – FIL A3

Apprentissage Automatique

Élève : Xavier Aleman

1. Charger et transformer les données de façon à ce qu’elles soient reconnues comme des transactions.

En pratique on construit un tableau de données binaires.

Nous récupérons les données du fichier retail_dataset.csv, seulement ses données ne sont pas exploitables en l'état.

En effet le fichier csv répertorie des transactions allant de 1 à 7 produits achetés. Or lors de la lecture du fichier la librairie pandas va produire une matrice de taille 7 × t où les 7 colones représentent la place du produit dans la transaction et t le nombre de transactions.

Pour ce faire la librairie doit donner une valeur pour chaque colonne d'un transaction et va donc affecter la valeur NaN s'il y a moins de 7 produit dans une transaction.

Nous pouvons pour éviter les erreurs de type (str et float) dans la matrice produite indiquer le paramètre "keep_default_na=False" ce qui produira des chaines de caractère vide à la place des NaN.

Une fois la matrice obtenue il faut la nettoyer des chaines de caractères vides. Pour cela nous allons créer une liste de liste afin d'avoir uniquement les produits de chaque transaction.

Maintenant que nous avons nettoyés les données nous devons les mettre au bon format pour les exploiter. Nous utilisons la libraire mlxtend pour transformer notre liste de liste en une matrice binaire n × t où n est le nombre de produits différents toutes transactions confondues et m le nombre de transactions.

	Bagel	Bread	Cheese	Diaper	Eggs	Meat	Milk	Pencil	Wine
0	False	True	True	True	True	True	False	True	True
1	False	True	True	True	False	True	True	True	True
2	False	False	True	False	True	True	True	False	True
...	...								

De là nous obtenons une base de donnée binaire D ⊆ T × I avec :

Tidsets: $T=\{t_1,t_2,...,t_{315}\},|T|=315$

Itemsets: $I=\{Bagel,Bread,Cheese,Diaper,Eggs,Meat,Milk,Pencil,Wine\},|I|=9$

Ce qui nous donne un espace de recherche de $2^{|I|}=2^9=512$

Nous créons ensuite un DataFrame, à partir de la base binaire pour pouvoir traiter ses données grace à la librairie pandas.

2. Utiliser l’algorithme Apriori pour extraire les itemsets fréquents et les maximaux. Vous choisirez un support minimum de 3 %.

— Que se passe t’il si on fait varier le seuil du support ?
— Tracer une courbe montrant l’évolution du nombre d’itemsets extraits en fonction du support minimum.

Nous cherchons à extraire les itemsets fréquents et maximaux de la base binaire obtenue précède

ment, pour ce faire il existe plusieurs approches. Cependant le plus simple appelé brute force à une

complexité en temps $O(|I|\times|D|)$ soit $O(|I|^2\times|T|)$ où $|I|^2\times|T|=9^2\times315=25515$ ce qui est long.

Pour éviter cela nous utiliserons l'algorithme apriori qui est moins complexe en temps $O(2^{|I|})$ où $2^{|I|}=2^9=512$. Cet algorithme prend en entrée un seuil de support ce qui lui permet de discriminer les résultats inférieurs et lui fait gagner du temps.

En choisissant un support à 3 % on obtient 301 itemsets différents. On remarque qu'en modifiant le seuil de support dans l'algorithme, le nombre d'itemsets augmente quand le support diminue (422 itemsets à 1 %) et diminue si le seuil de support augmente (3 itemsets à 50 %).

3. Nous souhaitons pouvoir filtrer les itemsets selon la présence d’items ou d’un ensemble d’items.

Par exemple, quels sont les itemsets qui contiennent le produit ’Eggs’ ? les produits {’Eggs’,’Meat’} ?

Il peut être intéressant de filtrer les itemsets obtenues si l'on s'interroge sur tel ou tels itemsets. Pour

cela nous utilisons la librairie pandas avec la méthode 'query' pour récupérer les transactions

contenant un ou plusieurs produits en particulier.Cependant les transactions qui en résultent sont des résultats binaires, il faut donc transformer les valeurs 'True' en l'item correspondant et supprimer les valeurs 'False'.
Nous obtenons donc tous les itemsets répondant à la fois au critère de support minimum attendu et à celui de la présence de certains items.

Number of itemsets with eggs : 138

id \\ item	1	2	3	4	5	6	7
0	Bread	Cheese	Diaper	Eggs	Meat	Pencil	Wine
2	Cheese	Eggs	Meat	Milk	Wine		
3	Cheese	Eggs	Meat	Milk	Wine		
...	...						
310	Bread	Cheese	Eggs				
312	Bread	Cheese	Diaper	Eggs	Meat	Pencil	Wine
314	Bagel	Bread	Eggs	Meat	Wine		

4. Utiliser l’algorithme Apriori pour extraire les règles d’association à partir des itemsets fréquents et des itemsets maximaux. Vous choisirez une confianceminimale de 75 %.

Extraire les règles ayant pour conséquents ’Chesse’.

Les règles d'association correspondent au fait que la présence d'un itemset implique la présence d'un autre. Sa fréquence correspondant au nombre de transactions dans lesquelles le couple d'itemset est compris dans l'itemsets.

Nous réutilisons les résultats précédents (apriori avec support minimum à 3 %) pour en extraire les règles d'associations et l'on limitera les résultats à ceux ayant une confiance minimale de 75 %.

Ce premier résultat obtenu nous ne gardons que ceux qui ont pour conséquent l'item 'Cheese'.

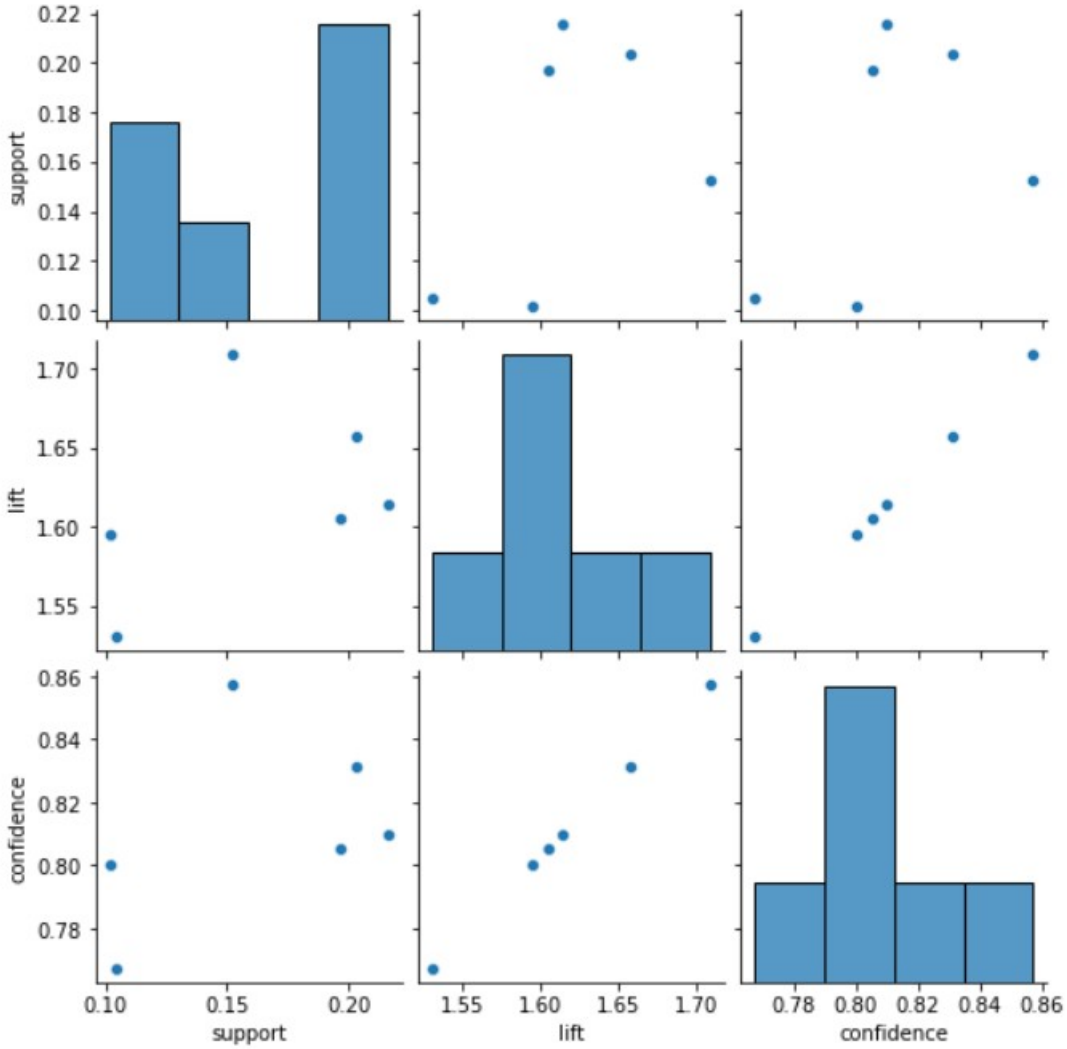
t	antecedents	consequents	support	lift	confidence
1	(Meat, Eggs)	(Cheese)	0.215873	1.613924	0.809524
2	(Milk, Eggs)	(Cheese)	0.196825	1.605293	0.805195
3	(Milk, Meat)	(Cheese)	0.203175	1.657077	0.831169
10	(Milk, Meat, Eggs)	(Cheese)	0.152381	1.708861	0.857143
11	(Milk, Eggs, Wine)	(Cheese)	0.104762	1.530026	0.767442
13	(Milk, Meat, Wine)	(Cheese)	0.101587	1.594937	0.800000
18	(Meat, Eggs, Diaper, Bagel)	(Cheese)	0.038095	1.495253	0.750000
20	(Pencil, Wine, Diaper, Bagel)	(Cheese)	0.031746	1.533593	0.769231
31	(Milk, Meat, Eggs, Wine)	(Cheese)	0.073016	1.581187	0.793103

Ce tableau nous montre que les paniers qui contiennent ces antécédents ont plus 75 % de chances de contenir aussi l'item 'Cheese'.

5. Compléter l’analyse des différentes règles d’association extraites via des graphiques permettant d’étudier la corrélation entre les trois mesures (lift, confiance et support) d’évaluation des règles.

La librairie seaborn nous permet de créer facilement des graphique depuis les DataFrames pandas.

Avec 3 mesures à comparer le graphique le plus intéressant est le 'pairplot' qui nous permet de comparer les nuages de points entre chacune des mesures.



On remarque que les mesures Lift et Confidence sont fortement corrélées.