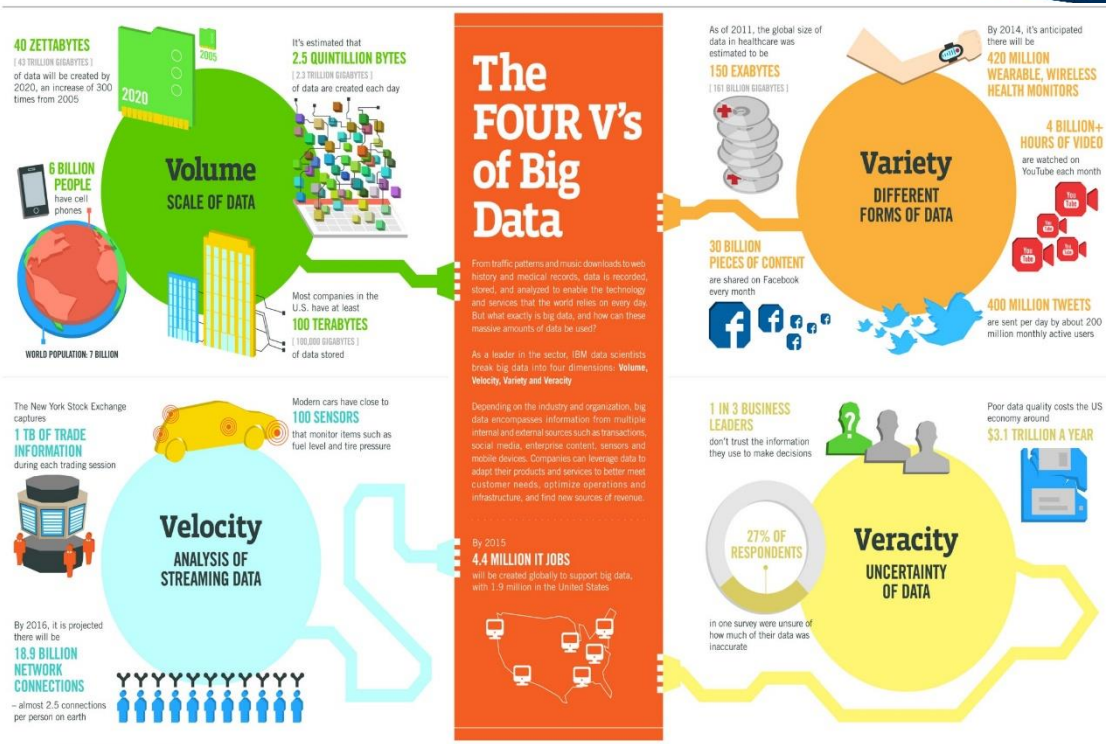
HEC MONTRÉAL

30-650-17
Introduction à l'analytique d'affaires

Thème 2 :
La valorisation des données : les règles d'association

Introduction


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
Volume
SCALE OF DATA
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1:9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Velocity
ANALYSIS OF STREAMING DATA
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

Le « Big Data »

Les données forment le cœur des processus de base (opérations) dans la plupart des entreprises ou organisations:

- Relation avec les clients:
  - commandes, envois, facturation, commentaires sur les réseaux sociaux, etc.
- Inventaire.
- Transactions bancaires.
- Relation avec les employés:
  - paye, avantages sociaux, maladies, promotions, etc.

Les données à analyser peuvent provenir de différentes sources et différents formats

Structurées	Non structurées
• Bases de données relationnelles avec champs bien définis. • Fichiers « plats » (*flat files*) délimités.	• Courriels, textos, « tweets ». • Commentaires, plaintes. • Documents. • Images, vidéos. • Bandes sonores.

Utilisation des données

- Beaucoup d'entreprises ont mis en place des stratégies d'entrepôts de données et ont commencé à s'interroger sur leur exploitation.
- Les données sur les clients et les employés sont qualifiées « **d'or noir** » de l'entreprise.
- L'archivage des données crée la mémoire de l'entreprise mais l'exploitation des données (data mining) crée l'intelligence de l'entreprise.

Plan

1. Retour sur les tableaux croisés
2. Définitions:
   - Probabilité conjointe
   - Probabilité conditionnelle
3. Notion d'indépendance
4. Application de ces notions dans le contexte du « Big data » : les règles d'association

Tableaux croisés

- Les tableaux croisés permettent d'analyser la relation entre deux variables ou plus.
- Le croisement des variables département, sexe des candidats et décision d'admission dans l'exemple des admissions aux études supérieures à Berkeley vu à la séance 1 a permis de constater qu'il n'y a pas eu de lien entre le sexe des candidats et la décision d'accepter ou de rejeter une demande d'admission.

Département	Hommes		Femmes	
	Applicants	Admis	Applicants	Admis
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Probabilité conjointe

Une probabilité **conjointe** est une probabilité qui **fait intervenir deux ou plusieurs variables.**

Exemple: supposons qu'un échantillon aléatoire de 100 consommateurs de boissons énergisantes ont participé à un test de goût de trois marques de ce type de boisson.

Chacun des participants a goûté aux trois boissons à l'aveugle (ordre aléatoire) et indiqué sa préférence (voir fichier ***30-650-boisson_énergisante-DATA.xlsx***).

Probabilité conjointe

- La **distribution conjointe** obtenue **selon le sexe et la préférence** des participants au test de goût des boissons énergisantes est:

	A	B	C
1	**Sondage boissons énergisantes**		
2			
3	**Répondant**	**Sexe**	**Marque préférée**
4	1	Homme	boisson 3
5	2	Femme	boisson 3
6	3	Homme	boisson 3
7	4	Homme	boisson 1
8	5	Homme	boisson 1
9	6	Femme	boisson 2

Nombre de Rép	Étiquettes de			
Étiquettes de	**boisson 1**	**boisson 2**	**boisson 3**	**Total général**
Femme	9	6	22	37
Homme	25	17	21	63
Total général	**34**	**23**	**43**	**100**

- Dans cet exemple on s'intéresse simultanément aux deux variables suivantes: **le sexe** et **la marque préférée** de boisson énergisante.

- La **probabilité conjointe** qu'un participant choisit au hasard **soit une femme et qu'elle préfère la boisson 3** est :

$$P(marque = 3 \; et \; sexe = F) = \cdots$$

Probabilité conditionnelle

- Une probabilité **conditionnelle** est **la probabilité de l'occurrence d'un évènement A sachant qu'un événement B est vrai ou connu.**

- On note généralement la probabilité conditionnelle de l'événement **A sachant B: P(A | B).**

- Dans l'exemple précédent sur les boissons énergisantes, **la probabilité que la marque 3 soit la préférée sachant que la personne qui a répondu est une femme est:**

$$P(\,marque = 3 \mid sexe \;=\; femme) \;= \cdots$$

Probabilité conditionnelle

De façon générale

$$\boldsymbol{P(A\,|B) = \frac{P(A\ et\ B)}{P(B)}}$$

où **P(A <u>et</u> B)** est la probabilité **<u>conjointe</u>** des événements A et B,

et **P(B)** est la probabilité de l'événement B.

$$P(marque = 3|sexe = F) = \frac{P(marque = 3\ et\ sexe = F)}{P(sexe = F)} = \cdots$$

Probabilité conditionnelle

Voici les probabilités conditionnelles pour la marque préférée **sachant le sexe du répondant :**

Nombre de Répon	Étiquettes de co			
Étiquettes de lig	boisson 1	boisson 2	boisson 3	Total général
Femme	9	6	22	37
Homme	25	17	21	63
Total général	**34**	**23**	**43**	**100**

Tableau des probabilités de la marque préférée conditionnellement au sexe du répondant

Nombre de Répon	Étiquettes de co			
Étiquettes de lig	boisson 1	boisson 2	boisson 3	Total général
Femme	24,3%	16,2%	59,5%	100,0%
Homme	39,7%	27,0%	33,3%	100,0%
Total général	**34,0%**	**23,0%**	**43,0%**	**100,0%**

Ces informations sont intéressantes au niveau marketing. Parmi les consommateurs de boissons énergisantes, nous remarquons que **les hommes préfèrent davantage la marque 1** et **les femmes la marque 3**.

Pour augmenter le nombre d'adeptes de boissons énergisantes, les efforts marketing devraient cibler les femmes avec la marque 3 et les hommes avec la marque 1.

Notion d'indépendance

- Deux événements, A et B, sont **indépendants** si l'occurrence de A n'a aucune incidence sur l'occurrence de B et vice versa.

- En d'autres termes, si la probabilité de l'occurrence de A ne dépend pas de l'occurrence ou la non occurrence de B et vice versa.

  Autrement dit,

$$\boldsymbol{P(A \mid B) = P(A)} \text{ et } \boldsymbol{P(B \mid A) = P(B)}$$

Notion d'indépendance

- Par exemple, soit A le résultat d'un premier lancer d'une pièce de monnaie et B le résultat d'un deuxième lancer d'une pièce monnaie.

- Quelle est la probabilité que B soit pile (ou face) sachant que le résultat de A est face (ou pile)?

Notion d'indépendance

Supposons qu'un échantillon aléatoire de 500 clients d'un supermarché dans quatre régions différentes ont participé à un sondage sur la satisfaction du service à la clientèle.

	Satisfaction		
	oui	non	**total**
Région 1	98	42	140
Région 2	84	36	120
Région 3	112	48	160
Région 4	56	24	80
Total	350	150	500

Est-ce que la probabilité qu'un client soit satisfait dépend de la région ou est-ce que la région et la satisfaction sont indépendantes?

Notion d'indépendance

Nous avons vu que la **probabilité conditionnelle** est définie comme :

$$P(A \mid B) = P(A \text{ et } B) / P(B)$$

ou

$$P(B \mid A) = P(A \text{ et } B) / P(A).$$

Ainsi nous avons pour la **probabilité conjointe de A et B** :

$$P(A \text{ et } B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A).$$

Par conséquent, si **A et B sont indépendants** nous obtenons **la règle de multiplication** suivante:

$$\boldsymbol{P(A \text{ et } B) = P(A)\, P(B) = P(B)\, P(A).}$$

Les règles d'association

Vin

Fromage

Objectif de l'analyse des associations

- **Identifier des profils**, associations ou structures entre **les items** ou objets qui sont fréquents à partir des informations dans les bases de données transactionnelles, relationnelles, ou dans les entrepôts de données.

- Autrement dit, il s'agit **d'identifier les items qui apparaissent souvent ensemble** lors d'un événement (ex. les items ou produits achetés lors d'une visite dans un magasin ou consultés sur un site web).

Quelques exemples d'applications

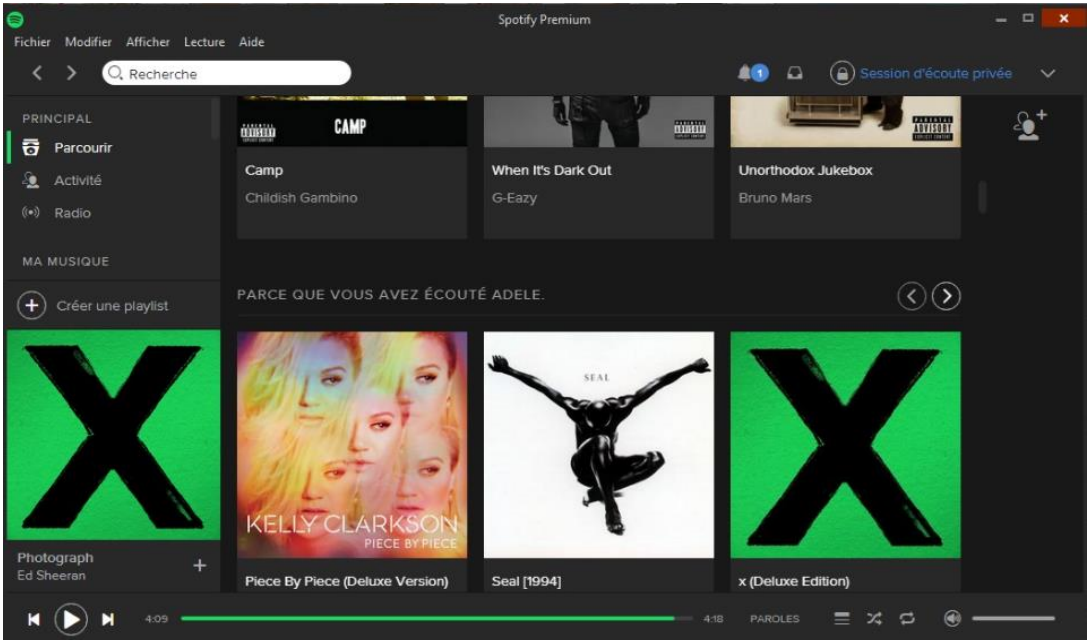
- Offres promotionnelles personnalisées (coupons rabais).

- Design de circulaires, catalogues et sites web.

- Disposition des produits sur les étagères.

Ex: Amazon, Métro et moi, Netflix, Itunes, Spotify, etc.

Exemple : suggestions sur Spotify



Qu'est-ce qu'une règle d'association?

- Une règle d'association décrit **les items qui se retrouvent souvent dans le même panier** et qui sont dépendants (liés).

- Par exemple, un site de vente de musique en ligne pourrait vouloir étudier l'association entre deux artistes. Par exemple, on peut vouloir **mesurer l'effet sur la popularité de Ed Sheeran sachant que l'utilisateur a acheté la musique d'Adele** :

  La règle: **Adèle => Ed Sheeran**

Terminologie pour les règles d'association

Dans une règle d'association, **l'antécédent** est ce qui est à gauche de la règle, c'est celui qui vient avant :

- Adele dans notre exemple

Le conséquent est ce qui vient à droite de la règle, c'est celui qui vient après :

- Ed Sheeran dans notre exemple

Antécédent => conséquent

Cas « Boulangerie »

- Une petite chaîne de boulangeries a **20 succursales dans cinq états** de la côte ouest des États-Unis.
- Chaque succursale offre le même menu d'environ **40 articles de pâtisseries et de 10 boissons.**
- Le fichier **« 30-650-boulangerie-DATA.xlsx »** contient un échantillon de **4523 reçus de caisse.**
- Chaque ligne du fichier correspond à un reçu (une transaction) et comprend les informations suivantes dans les colonnes:
    - Numéro du reçu, jour de fin de semaine (1=oui, 0=jour de semaine), numéro de la succursale, ville, état, et 28 des 50 items offerts sur le menu.
    - Un **1** dans la colonne d'un item indique que l'item a été acheté par le client et un **0** indique que l'item n'a pas été acheté.

Source: https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery

Cas « Boulangerie »

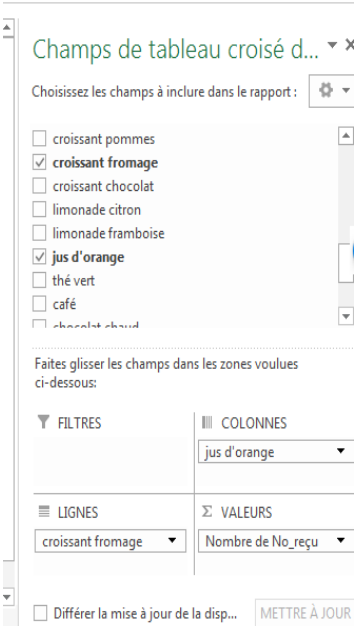
- À l'aide de l'outil « tableau croisé dynamique de Excel » nous étudierons les deux règles liées au couple de produits Croissant au fromage **(CF)** et Jus d'orange **(JO)**.

- La règle $(\boldsymbol{CF} \Rightarrow \boldsymbol{JO})$: **mesure l'effet** sur la probabilité que le client achète du **jus d'orange** **sachant** qu'il a acheté au moins un **croissant au fromage**.

- La règle $(\boldsymbol{JO} \Rightarrow \boldsymbol{CF})$: **mesure l'effet** sur la probabilité que le client achète au moins **un croissant au fromage** **sachant** qu'il a acheté du **jus d'orange**.

Construction du tableau croisé dynamique

Nombre de No_reçu	Étiquettes de colonnes		
Étiquettes de lignes	0	1	Total général
0	3892	246	4138
1	170	215	385
Total général	**4062**	**461**	**4523**



Indicateurs pour mesurer l'intérêt potentiel d'une règle d'association

Dans un premier temps, on commencera par mesurer **la popularité** a priori des deux items faisant partie de la règle.

On s'intéresse donc à la probabilité qu'un client quelconque veuille acheter un item, i.e., **la proportion de fois où l'item** est présent dans la base de données.

On notera par la suite cette probabilité : **le support individuel d'un item**.

- **Fréquence d'un item (X)** est le nombre de transactions dans la base de données qui contiennent l'item X.

- **Support de l'item (X)** $= \boldsymbol{P(X)} = \dfrac{\textbf{Fréquence de l'item (X)}}{\boldsymbol{n}}$

  Où n = nombre de transactions

Exemple

Nombre de No_reçu	Étiquettes de JO		
Étiquettes de CF	0	1	Total général
0	3892	246	4138
1	170	215	385
Total général	**4062**	**461**	**4523**

$support\ (CF) = \cdots$

$support\ (JO) = \cdots$

Interprétation : ……% des transactions contiennent du croissant au fromage.

Support d'une règle

- **Fréquence de la règle** $(X => Y)$ est le nombre de transactions dans la base de données qui **contiennent tous les items de X et Y.**
- **Support de la règle** $(X => Y) = \frac{\text{Fréquence de la règle } (X \Rightarrow Y)}{n}$
- ***Le support de (X=>Y) est la probabilité conjointe*** $P(X \text{ et } Y)$**.**
- Le support est utilisé pour identifier les règles qui sont fréquentes.
- **Le support** est **symétrique**, i.e., que

$$Support\ (X => Y) = Support\ (Y => X) = P(X \text{ et } Y)$$

Exemple

Nombre de No_reçu	Étiquettes de JO		
Étiquettes de CF	0	1	Total général
0	3892	246	4138
1	170	215	385
Total général	**4062**	**461**	**4523**

$support\ (CF) = \cdots$

$support\ (JO) = \cdots$

$support\ (CF \Rightarrow JO) = \cdots$

$support\ (JO \Rightarrow CF) = \cdots$

Interprétation: ……. des transactions contiennent le croissant au fromage et le jus d'orange.

Confiance d'une règle

- **Confiance** $(X => Y)$ est la probabilité que Y soit présent dans la transaction sachant que X est présent.
- La confiance de $(X => Y)$ est la ***probabilité conditionnelle*** $P(Y|X)$
- $Confiance\ (X => Y) = P(Y|X) = \frac{fréquence\ (X \Longrightarrow Y)}{fréquence\ (X)} = \frac{P(X\ et\ Y)}{P\ (X)}$

$$Confiance\ (X => Y) = \frac{\text{nombre de transactions contenant X et Y}}{\text{nombre de transactions contenant X}}$$

$$Confiance\ (X => Y) = \frac{support(X \Longrightarrow Y)}{support\ (X)}$$

- La confiance n'est pas symétrique car $P(Y|X)$ n'est pas nécessairement égale à $P(X|Y)$.
- On comparera la **confiance** $(X => Y)$ avec **le support individuel de Y** pour mesurer si le fait de savoir que le client a acheté l'item X augmente de beaucoup la probabilité (ou confiance) qu'il achètera Y, donc comparer $P(Y|X)$ avec $P(Y)$**!**

Exemple

Nombre de No_reçu	Étiquettes de JO		
Étiquettes de CF	0	1	Total général
0	3892	246	4138
1	170	215	385
Total général	**4062**	**461**	**4523**

On a déjà calculé :

$$support\ (CF) = 8{,}51\%$$
$$support\ (JO) = 10{,}19\%$$
$$support\ (CF => JO) = support\ (JO => CF) = 4{,}75\%$$

Calculons :

$$confiance\ (CF => JO) = P(JO|CF) = \cdots$$

$$confiance\ (JO => CF) = P(CF|JO) = \cdots$$

Lift d'une règle

$$\boldsymbol{lift}\ (\boldsymbol{X}\ =>\ \boldsymbol{Y}) = \frac{confiance\ (X \Longrightarrow Y)}{support\ (Y)}$$

$$\boldsymbol{lift}\ (\boldsymbol{X}\ =>\ \boldsymbol{Y})\ = \frac{P(Y|X)}{P(Y)} = \frac{support\ de\ (X \Longrightarrow Y)}{support\ (X) * support\ (Y)} = \frac{P(XetY)}{P(X) * P(Y)}$$

Lift d'une règle

- Si le **lift = 1** alors X et Y sont **indépendants** car $P(Y|X)$ est égale à $P(Y)$.

- Si le **lift > 1** alors il existe **une association positive**, i.e., la probabilité d'acheter Y a été augmentée (ou « liftée ») par le fait de savoir que le client a acheté X.

  En fait, X et Y se retrouvent plus souvent ensemble que s'ils étaient des événements indépendants. On pourrait avoir ce type de résultats pour des items complémentaires.

- Si le **lift < 1** alors il existe **une association négative** car la probabilité d'acheter Y a été diminuée par le fait de savoir que le client a acheté X.
  En fait, X et Y se retrouvent moins souvent ensemble que s'ils étaient des événements indépendants. On pourrait avoir ce type de résultats pour des items substituables.

- Le lift est symétrique.

Exemple

On a déjà calculé :

$$support\ (CF) = 8{,}51\%$$
$$support\ (JO) = 10{,}19\%$$
$$support\ (CF \Rightarrow JO) = support\ (JO \Rightarrow CF) = 4{,}75\%$$
$$confiance\ (CF \Rightarrow JO) = P(JO|CF) = 55{,}84\%$$
$$confiance\ (JO \Rightarrow CF) = P(CF|JO) = 46{,}64\%$$

Calculons :

$$lift\ (CF \Rightarrow JO) = \cdots$$

$$lift\ (JO \Rightarrow CF) = \cdots$$

Gabarit Excel « Association.xlsx »

- La capsule-vidéo disponible sur Zone Cours explique son fonctionnement.
- Ce gabarit calcule automatiquement les indicateurs de toutes les règles à 2 items pour une base de données contenant jusqu'à **5000 transactions** avec **au plus 30 items.**
- Coller dans la feuille « Données » du gabarit, les données de vos transactions.
- Les feuilles **« Support »**, **« Confiance »** et **« Lift »** donnent les résultats de ces trois indicateurs respectivement pour toutes les règles à 2 items.
- Les supports individuels se trouvent dans la diagonale de la matrice des supports.
- Pour la matrice **« confiance »,** elle se lit dans le sens :

Confiance (item ligne => item colonne)

Commentaires sur l'analyse des règles

- En pratique, le nombre de règles augmente rapidement. On pourrait alors s'intéresser à limiter le nombre de règles à analyser. Un critère parfois utilisé en pratique est de se limiter aux règles qui ont des « lift » élevés, généralement **plus grand que 3 ou 4**.

- Il faut quand même regarder ensuite les autres indicateurs pour voir si la règle est vraiment intéressante et s'il est intéressant de proposer l'item « conséquent » au client. Même si la probabilité d'acheter cet item a été augmentée, il se peut que cet item demeure tout de même moins populaire que d'autres items pour ce client.

- En résumé, **il faut idéalement analyser l'ensemble des indicateurs numériques** (lift, confiance, support conjoint, support individuel) pour en arriver à prendre une bonne décision.

- Au final c'est le gestionnaire qui sait si les règles retenues peuvent être utiles et mener à des actions.

Exercice sur le cas « Boulangerie »

1. Quelles sont les trois items les plus populaires?

2. Si on commençait par étudier les règles avec un haut niveau de lift, quelles seraient les 3 paires avec le plus haut potentiel? Quelles sont les 6 règles associées à ces trois paires d'items?

3. Sachant que le client a acheté l'item «limonade au citron » quels seraient les trois items les plus susceptibles de l'intéresser? Et les trois items les moins susceptibles de l'intéresser?