Apprentissage non-supervisé.

François-Xavier Dupé

IFD Masters BDA/FSI

Introduction

Comment « faire parler les données » ?

- relation entre les données,
- existence de « profils types »,
- quels sont les différentes « composantes ».

Contextes :

- grand ensemble de données (plusieurs Tera octets),
- profils clients pour un commerce,
- recherche de liens pour une maladie...

Recherche des informations dans les données elles-mêmes.

Introduction

Relation entre les données $\Rightarrow$ Règles d'associations

Assembler selon la similarité $\Rightarrow$ Clustering

Trouver les liens entre les éléments $\Rightarrow$ Séparation en composantes

Plan du cours

Plan du cours

Exemple classique

Analyse des achats dans un supermarché :

- Entrée :
  - une liste des produits à vendre,
  - l'ensemble des tickets de caisse sur une période.
- Sortie : les liens entre les achats.

Exemple classique

Analyse des achats dans un supermarché :

- Entrée :
  - une liste des produits à vendre,
  - l'ensemble des tickets de caisse sur une période.
- Sortie : les liens entre les achats.

Exemple :

- 90% des personnes qui achètent un tire-bouchon, achètent aussi du vin,
- 50% des acheteurs d'une maison pensent faire des travaux...

Problématique

- Entrée :
  - une large base d'éléments,
  - une base de transactions liant ces éléments.
- Sortie : les règles qui corrèlent les éléments entre eux.

Problématique

- Entrée :
  - une large base d'éléments,
  - une base de transactions liant ces éléments.
- Sortie : les règles qui corrèlent les éléments entre eux.

$\Rightarrow$ Recherche tous types d'associations entre les éléments.

Principe

Recherche de règles d'associations de la forme,

$$Corps \quad \Rightarrow \quad Tête\ [support,\ confiance]\ .$$

Corps : propriété d'un objet (transaction, personne...),
Tête : propriété *probable* impliquée par le Corps,
support, confiance : mesures de validité de la règle.

Règles d'assocation uni/multi-dimensionnelles

- Règle à une dimension,

$$\text{achète(X, ”pain”)} \Rightarrow \text{achète(X, ”beurre”)} \ .$$

Règles d'assocation uni/multi-dimensionnelles

- Règle à une dimension,

$$\text{achète(X, "pain")} \Rightarrow \text{achète(X, "beurre")} \ .$$

- Règle à plusieurs dimensions (donc plusieurs conditions),

$$\text{age(Y,"19-25")} \wedge \text{achète(Y,"trousse")} \Rightarrow \text{achète(Y,"stylos rouge")} \ .$$

Règles d'assocation uni/multi-dimensionnelles

- Règle à une dimension,

$$\text{achète(X, ”pain”)} \Rightarrow \text{achète(X, ”beurre”)} \ .$$

- Règle à plusieurs dimensions (donc plusieurs conditions),

$$\text{age(Y,”19-25”)} \wedge \text{achète(Y,”trousse”)} \Rightarrow \text{achète(Y,”stylos rouge”)} \ .$$

- Notation simplifié,

$$\{\text{pain}\} \Rightarrow \{\text{beurre}\},$$
$$\{[\text{age, ”19-25”}], [\text{achète, ”trousse”}]\} \Rightarrow \{[\text{achète, ”stylos rouge”}]\} \ .$$

Support et confiance

Support : probabilité que le Corps et la Tête soient impliqués dans une transaction.
Confiance : probabilité que si le Corps apparaît, la Tête apparaisse aussi.

Support et confiance

Support : probabilité que le Corps et la Tête soient impliqués dans une transaction.
Confiance : probabilité que si le Corps apparaît, la Tête apparaisse aussi.

Exemple :

achète(X,"pain") ⇒ achète(X,fromage) [75%, 50%]

Transaction	Achats
10	beurre, lait, fromage
20	lait, fromage, yaourt, pain
30	fromage, pain
40	beurre, fromage, pain

Les règles d'associations

Terminologie et notation

Soit un ensemble d'items (éléments) I, un sous-ensemble de I est un *itemset*.
Une transaction est notée Transaction(tid, T), avec tid son identifiant et $T \subseteq I$.
Soit D l'ensemble des transactions (donc une base de données).

Les règles d'associations (2)

Définition

Une règle est notée $A \Rightarrow B\ [s, c]$.

- A, B sont des itemsets ($A, B \subseteq I$),
- $A \cap B = \emptyset$.

Les règles d'associations (2)

Définition

Une règle est notée $A \Rightarrow B\ [s, c]$.

- A, B sont des itemsets ($A, B \subseteq I$),
- $A \cap B = \emptyset$.

Nous avons donc,

- le support $s = \mathrm{P}(A \cup B)$ la probabilité qu'une transaction contienne $A \cup B$,
- la confiance $c = \mathrm{P}(B|A)$ la probabilité qu'une transaction impliquant A implique aussi B.

L'algorithme A Priori

1. Initialisation : $k = 1$, L_k ensemble des items fréquents de D.

L'algorithme A Priori

1. Initialisation : $k = 1$, L_k ensemble des items fréquents de D.
2. Boucle : tant que L_k n'est pas vide.

L'algorithme A Priori

1. Initialisation : $k = 1$, L_k ensemble des items fréquents de D.
2. Boucle : tant que L_k n'est pas vide.
    1. Calcul de nouveaux itemsets par un jeu de fusion/suppression avec les itemsets de L_k, ces nouveaux itemsets forment C_k.

L'algorithme A Priori

1. Initialisation : $k = 1$, L_k ensemble des items fréquents de D.
2. Boucle : tant que L_k n'est pas vide.
   1. Calcul de nouveaux itemsets par un jeu de fusion/suppression avec les itemsets de L_k, ces nouveaux itemsets forment C_k.
   2. Pour chaque transition t dans D, chercher les itemsets de C_k impliqué et mettre à jour le support.

L'algorithme A Priori

1. Initialisation : $k = 1$, L_k ensemble des items fréquents de D.
2. Boucle : tant que L_k n'est pas vide.
   1. Calcul de nouveaux itemsets par un jeu de fusion/suppression avec les itemsets de L_k, ces nouveaux itemsets forment C_k.
   2. Pour chaque transition t dans D, chercher les itemsets de C_k impliqué et mettre à jour le support.
   3. Former L_{k+1} avec les itemsets de C_k ayant un support superieur à un seuil donné.

L'algorithme A Priori

1. Initialisation : $k = 1$, L_k ensemble des items fréquents de D.
2. Boucle : tant que L_k n'est pas vide.
   1. Calcul de nouveaux itemsets par un jeu de fusion/suppression avec les itemsets de L_k, ces nouveaux itemsets forment C_k.
   2. Pour chaque transition t dans D, chercher les itemsets de C_k impliqué et mettre à jour le support.
   3. Former L_{k+1} avec les itemsets de C_k ayant un support superieur à un seuil donné.
   4. $k = k + 1$.

L'algorithme A Priori

1. Initialisation : $k = 1$, L_k ensemble des items fréquents de D.
2. Boucle : tant que L_k n'est pas vide.
    1. Calcul de nouveaux itemsets par un jeu de fusion/suppression avec les itemsets de L_k, ces nouveaux itemsets forment C_k.
    2. Pour chaque transition t dans D, chercher les itemsets de C_k impliqué et mettre à jour le support.
    3. Former L_{k+1} avec les itemsets de C_k ayant un support superieur à un seuil donné.
    4. $k = k + 1$.
3. Sortie : $\cup L_k$ l'ensemble des itemsets fréquents.

Générer les règles d'associations

Rappel : nous avons L, l'ensemble des itemsets fréquents.

1. Initialisation : pour chaque itemset de L, $I \in L$, générer tous les sous-ensembles non-vides de I formant $S(I)$.

Générer les règles d'associations

Rappel : nous avons L, l'ensemble des itemsets fréquents.

1. Initialisation : pour chaque itemset de L, $I \in L$, générer tous les sous-ensembles non-vides de I formant $S(I)$.
2. Pour chaque sous-ensemble d'un itemset, $S \subseteq S(I),\ I \in L$, générer une règle $S \Rightarrow I \setminus S$ si,

$$\frac{\text{support}(I)}{\text{support}(S)} \geq \text{confiance minimale .}$$

Générer les règles d'associations

Rappel : nous avons L, l'ensemble des itemsets fréquents.

1. Initialisation : pour chaque itemset de L, $I \in L$, générer tous les sous-ensembles non-vides de I formant $S(I)$.
2. Pour chaque sous-ensemble d'un itemset, $S \subseteq S(I),\ I \in L$, générer une règle $S \Rightarrow I \setminus S$ si,

$$\frac{\text{support}(I)}{\text{support}(S)} \geq \text{confiance minimale} .$$

En effet,

$$\text{confiance}(A \Rightarrow B) = \text{P}(B|A) = \frac{\text{support}(A \cup B)}{\text{support}(A)} .$$

Conclusion sur les règles d'associations

Les règles d'associations sont un outils efficace pour mettre en avant les liens existants dans les données, et il existe des améliorations de l'algorithme présenté ici. Il convient toutefois de faire attention sur certain points :

Conclusion sur les règles d'associations

Les règles d'associations sont un outils efficace pour mettre en avant les liens existants dans les données, et il existe des améliorations de l'algorithme présenté ici. Il convient toutefois de faire attention sur certain points :

- la *qualité* des règles dépend évidemment de le taille de la base de données.

Conclusion sur les règles d'associations

Les règles d'associations sont un outils efficace pour mettre en avant les liens existants dans les données, et il existe des améliorations de l'algorithme présenté ici. Il convient toutefois de faire attention sur certain points :

- la *qualité* des règles dépend évidemment de le taille de la base de données.
- seul les liens les plus évidents peuvent être déduits, les plus complexes demandent une lecture attentive des résultats.

Conclusion sur les règles d'associations

Les règles d'associations sont un outils efficace pour mettre en avant les liens existants dans les données, et il existe des améliorations de l'algorithme présenté ici. Il convient toutefois de faire attention sur certain points :

- la *qualité* des règles dépend évidemment de le taille de la base de données.
- seul les liens les plus évidents peuvent être déduits, les plus complexes demandent une lecture attentive des résultats.
- ne pas oublier l'environnement lors de l'acquisition des données.