IFT 501 – Recherche d'information et forage de données
TP1 : Règles d'associations
Date de remise : le mardi 29 octobre 2019

Ce TP porte sur l'extraction des règles d'associations. Il consiste à développer une étude d'une base de données bancaire à l'aide des techniques d'analyse d'association. Le but de l'étude est d'identifier des profiles d'épargnants parmi les clients, identifier des non épargnants qui seraient intéressés d'acheter des produits d'épargne, et détecter des clients atypiques.

Analyse d'une base de données bancaire

Les données du TP :

- "bank-data" fournies en deux formats, .csv et .ARFF (format de Weka).
- Explication des attributs : Tableau 1.

Le travail peut être réalisé par Weka. Vous pouvez aussi le faire avec un langage de votre choix. Par exemple, si vous voulez le faire avec Python, la page suivante vous donne des informations sur l'extraction des règles en utilisant l'algorithme Apriori avec Python.
https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/

Un épargnant est une personne ayant soit un « mortgage » soit un plan « pep ». Les tâches à effectuer pour ce TP sont :

- Extraire des règles significatives (intéressantes) pour décrire les profiles des épargnants. Les seuils comme min-conf et min-lift n'étant pas fournis, vous devez chercher des règles ayant les plus grandes valeurs de *confiance* et de « *lift* ». Le but n'est pas d'extraire TOUTES les règles possibles, une dizaine, s'il y en a, serait suffisant;
- Identifier des clients non-épargnants correspondant aux profiles extraits afin de considérer ces clients comme susceptibles de devenir des épargnants ;
- Identifier (ou détecter) des clients atypiques qui, selon un critère que vous allez définir, sont très différents des autres. Votre critère n'a pas besoin d'être très sophistiquée et vous n'êtes pas obligés d'extraire TOUS les clients atypiques non plus.

Vous pouvez utiliser l'un des deux algorithmes (Apriori et FP) pour l'extraction des "itemsets" fréquents. Votre rapport doit décrire clairement les différentes étapes de traitement incluant les prétraitements effectués, les résultats obtenus, vos commentaires et votre conclusion.
Vous avez beaucoup de liberté pour développer votre propre solution. Vous devez faire preuve d'un peu d'imagination. Il n'y a pas UNE meilleure solution. Toutefois, le bon sens est de mise.

TABLE 1 – Explication des attributs

id	a unique identification number
age	age of customer in years (numeric)
sex	MALE / FEMALE
region	inner_city/rural/suburban/town
income	income of customer (numeric)
married	is the customer married (YES/NO)
children	number of children (numeric)
car	does the customer own a car (YES/NO)
save_acct	does the customer have a saving account (YES/NO)
current_acct	does the customer have a current account (YES/NO)
mortgage	does the customer have a mortgage (YES/NO)
pep	did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO)

Points Importants :

1. Le TP doit être fait tout seul ou en équipe de deux personnes;
2. La date limite de remise officielle est le mardi 29 octobre 23h59. Un bonus de 10 points sera accordé si la remise est faite le jour d'avant l'intra, c-à-d. avant 17 octobre 23h59 (le bonus ne vous permet pas d'avoir plus que 100% cependant);
3. Soignez votre rapport, une pénalité de 10% sera appliquée pour un rapport mal écrit; Votre rapport ne doit pas dépasser 3 pages + la page de couverture (pas indispensable);
4. Les fichiers à soumettre sont le rapport (en Word ou pdf). **Ne pas soumettre les données!**
5. N'oubliez pas de vous identifier. Indiquez votre nom et matricule dans chacun des fichiers que vous soumettez. La remise doit être faite par http://opus.dinf.usherbrooke.ca