Examen écrit

Règles d'association
Majeure Science des données

le 26 novembre 2015, durée 1h30

L'ordre de résolution des sujets n'est pas imposé. Les sujets sont indépendants.

1. Dans le script `arules_titanic.R` fourni sur Campus :

```
library(arules)

df <- as.data.frame(Titanic)

titanic.raw <- NULL
for(i in 1:4) {
 titanic.raw <- cbind(titanic.raw, rep(as.character(df[,i]),
}

titanic.raw <- as.data.frame(titanic.raw)
names(titanic.raw) <- names(df)[1:4]

rules <- apriori(titanic.raw, control = list(verbose=F),
                 parameter = list(minlen=2, supp=0.005, conf=0.8),
                 appearance = list(rhs=c("Survived=No", "Survived=Yes"),
                 default="lhs"))
rules
inspect(rules)
plot(rules)
plot(rules, method="graph")
```

les deux commandes `plot` ne fonctionnent pas. Pourquoi ?

2. Les règles d'association ont une visée **descriptive**, à savoir on arrive à trouver des motifs intéressants (itemsets fréquents ou règles d'association) qui ne sont pas explicits dans les données. Est-ce que les règles d'association, sous leur forme $x \rightarrow y$, peuvent être utilisées dans un but prédictif ? Pourquoi /comment ?
3. Si on fournit un ensemble des transactions sur d items distincts, prouver que le nombre de règles d'association est

$$3^d - 2^{d+1} + 1$$

sans compter les règles avec la partie droite ou la partie gauche nulle.

4. Dans certains cas on est intéressé uniquement par les règles d'association qui font intervenir **tous** les items (attributs) d'un ensemble donné, par exemple on s'intéresse à tout l'ensemble $\{i_1, i_2, \ldots, i_k\}$ quand l'ensemble complet d'items est $\{i_1, i_2, \ldots, i_k, \ldots i_d\}$, avec $d > k$. Proposez un algorithme ou une méthode qui résout ce problème particulier.
5. La base suivante traduit une série de transactions de type "panier".

Id	Transaction
T100	C, A, F, E
T200	C, A, F, E, I, N
T300	C, I, E, L
T400	L, I, A, N, E
T500	L, A, I, T
T600	C, A, T, E

Soient les limites du support ($min\_support$) à 49% et de la confiance ($min\_confiance$) à 80%.

— Appliquant l'algorithme **A priori** calculez tous les itemsets fréquents par rapport à *min_support (Attention : la limite est donnée en pourcentage !)*

— A partir du résultat calculé au point précédent calculez des règles d'association de type $X, Y \rightarrow Z$ qui sont au-delà de $min\_confiance$. Retenez uniquement les règles d'association avec un LIFT convenable.