

Examen final

UP3 - clustering, classification, règles d'association

le 5 janvier 2020, durée 1h

Les documents papier sont autorisés. L'usage des téléphones portables et autres dispositifs électroniques à l'exception de la calculatrice est prohibé.

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

1. Un data scientist travaille sur un ensemble de transactions `DonneesPaniers` en R. L'ensemble est correctement construit, toutefois lors de l'exécution de l'algorithme **Apriori** le résultat ne contient aucune règle d'association :

```
rules.all <- apriori(DonneesPaniers)
inspect(rules.all)
```

(l'ensemble `rules.all` est vide.)
Que faut-il faire ?

2. Soit le petit ensemble de transactions :

items	transactionID
[1] {Burger, Buns, Ketchup}	T1
[2] {Burger, Buns}	T2
[3] {Burger, Coke, Chips}	T3
[4] {Chips, Coke}	T4
[5] {Chips, Ketchup}	T5
[6] {Burger, Coke, Chips}	T6

- (a) à l'aide de l'algorithme Apriori ou de l'algorithme ECLAT calculez tous les itemsets fréquents pour un $min_support = 0,334$ (supérieur donc à $1/3$).
- (b) déterminez toutes les règles d'association ayant une confiance plus grande que $8/10$.

3. Outlier.

Un outlier (donnée aberrante) est une observation qui se trouve à grande distance de toutes les autres données.

- (a) parmi les algorithmes de clustering lesquels vous semblent plus adaptés pour détecter les outliers ?
- (b) quels algorithmes de clustering sont affectés par la présence des outliers ?
- (c) On vous fournit des données séparées en 3 classes pour faire de la classification et l'algorithme k-NN semble bien fonctionner. Toutefois parmi les données à traiter (à classifier) par la suite il est possible que des outliers se glissent et qu'ils n'appartiennent vraiment à aucune des 3 classes. Comment les identifier ?