Fouille de Données - TP3

M2- 2011-2012

Durant cette séance, vous allez devoir mettre en application ce que vous avez vu durant les séances précédentes afin de tirer profit des données qui vous sont présentées. Les ressources nécessaires sont disponibles à l'adresse : `http://archive.ics.uci.edu/ml/`.

Un compte rendu par (bi | tri)nome devra être renvoyé à `marc.plantevit@univ-lyon1.fr` **avec [M2Pro] TP3-CR en objet avant le 19/12/2011.**

Voici une liste non exhaustive de jeux de données[1] :

Car Evaluation Data Set : `http://archive.ics.uci.edu/ml/datasets/Car+Evaluation`

Communities and Crime Unnormalized Data Set : `http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized`

OpinRank Review Dataset Data Set* **:** `http://archive.ics.uci.edu/ml/datasets/OpinRank+Review+Dataset`

Pittsburgh Bridges Data Set : `http://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges`

Flags Data Set : `http://archive.ics.uci.edu/ml/datasets/Flags`

Student Loan Relational Data Set* **:** `http://archive.ics.uci.edu/ml/datasets/Student+Loan+Relational`

Statlog Project Data Set : `http://archive.ics.uci.edu/ml/datasets/Statlog+Project` (il y a plusieurs jeux de données).

Wine Data Set : `http://archive.ics.uci.edu/ml/datasets/Wine`

Zoo Data Set : `http://archive.ics.uci.edu/ml/datasets/Zoo`

US Census Data Set** **:** `http://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29` ou `http://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29`

Movie Data Set * : `http://archive.ics.uci.edu/ml/databases/movies/movies.data.html`

Plants Data Set : `http://archive.ics.uci.edu/ml/datasets/Plants`

Wine Quality Data Set : `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`

*** :** Peut nécessiter des prétraitements importants.

**** :** Il se peut que le nombre de tuples soit difficile à gérer en passant par l'interface graphique.

Choisissez 3 ou 4 jeux de données. Essayez de les exploiter avec les techniques utilisées précédemment (clustering, règles d'association, classifieur si besoin, etc.). Vous pouvez combiner ces techniques pour, par exemple, décrire chaque cluster à l'aide de règles d'association pour faciliter leur interprétation. Vous pouvez utiliser l'interface graphique ou l'API Java, voire des logiciels plus évolués (e.g. Knime). **Attention**, n'oubliez pas de normalisez vos données dans certains cas. Tous les attributs ne sont pas forcément nécessaires selon la tâche que vous souhaitez effectuer (attribut-clé, etc.). Il est également possible d'ajouter de nouveau attributs qui peuvent être des combinaisons d'autres attributs. Vous pouvez prétraiter vos données pour les mettre en forme ("weka-readable") avec votre langage de script préféré.

1. Vous êtes libre d'étudier un autre jeu de données qui vous tient à cœur.