

Projet SINF2275
« Data mining and decision making »
Projet clustering de données

Année académique 2006-2007

Professeurs : Marco Saerens
Adresse : Université catholique de Louvain
Information Systems Research Unit (ISYS)
Institut d'Administration et de Gestion
Place des Doyens 1
B-1348 Louvain-la-Neuve
Belgique
Téléphone : 010 47.92.46.
Fax : 010 47.83.24.
Courriel : saerens@ucLouvain.be

Objectif

L'objectif de ce travail est de vous familiariser avec les principales méthodes de clustering de données ainsi que les méthodes d'évaluation.

Modalités pratiques

La réalisation du projet se fera par groupes de deux au maximum. Un programme de génération de données artificielles vous sera fourni à la première séance. Il vous sera demandé de d'implémenter sous Matlab, Octave, R ou S-Plus trois méthodes de clustering :

- La méthode itérative k-means
- La méthode de clustering hiérarchique
- Une méthode de « spectral clustering » au choix

Il s'agira de créer plusieurs bases de données pour tester les méthodes de clustering implémentées. Affichez le résultat sous forme graphique, et évaluez le taux de bonne classification en calculant la matrice de confusion et l'adjusted RAND index. N'hésitez pas à générer des bases de données différentes pour mettre en évidence les avantages et désavantages de chaque méthode. Il est aussi conseillé de modifier la valeur des paramètres des algorithmes pour observer l'influence que cela produit sur le résultat.

Dans le cas de Matlab et R, les deux premières méthodes de clustering, ainsi que diverses opérations matricielles et arithmétiques sont déjà implémentés et peuvent être utilisés pour ce projet. L'évaluation du projet est basée principalement sur la compréhension de la matière, et non sur la complexité du programme. Cependant, ceux qui le désirent peuvent également implémenter les méthodes par eux même. Toute amélioration des algorithmes tel que l'intégration d'une technique de descente de gradient ou de recherche automatique de nombre de clusters seront considérées comme un plus et seront valorisées dans la cotation.

Outils logiciels

Nous demandons l'utilisation d'un outil logiciel spécialisé d'analyse des données.

Deux outils (*Matlab*, *R*) sont **conseillés** dans le cadre de ce travail, et sont disponibles en salle informatique. R est un logiciel libre, et peut être téléchargé à partir du site : <http://www.R-project.org>. L'utilisation de R ou Matlab est conseillée pour les projets de ce cours.

Notons que la version étudiant de S-Plus, l'équivalent commercial de R, est gratuite et peut être téléchargée à partir de www.insightful.com. Octave est un logiciel libre similaire à Matlab, bien que moins évolué au niveau de l'interface utilisateur.

Notez aussi que vous pouvez également utiliser un package Python appelé « orange », qui est en cours de développement tout en étant prometteur : <http://www.aialab.si/orange>.

Programme de génération de données

Un programme tournant sous Matlab vous sera fourni pour la génération de données artificielles à partir des points dessinés directement à l'écran ou à partir d'une image simple échantillonnée. Le programme fournira en sortie les coordonnées des points dans un espace à deux dimensions ainsi que la classe de chacun d'eux sous format .mat pour matlab (donnees.mat), et .txt pour les autres logiciels (donnees.txt).

Organisation du travail en groupe

Les étudiants travailleront par groupes de deux personnes au maximum. Veuillez à constituer les groupes le plus rapidement possible en envoyant un mail à luh.yen@uclouvain.be et en précisant le nom des membres de votre groupe.

Evaluation du projet

L'évaluation se fera durant la **quatrième semaine**, soit durant la deuxième heure de la troisième séance de TP, sur base d'une démonstration du code en salle Siemens; il n'y a donc pas de rapport à remettre. Toutefois, vous devrez être capable d'expliquer votre code, ainsi que les différentes décisions d'implémentation que vous avez choisies et les problèmes éventuels de votre solution. Vous devrez aussi être capable de répondre à des questions de compréhension sur la matière.

L'ordre de passage sera précisé une fois que tous les groupes auront été constitués.
