UNIVERSITÉ DE SHERBROOKE

DÉPARTEMENT D'INFORMATIQUE

IFT 599/799 : Sciences de données

TP #3 — Automne 2019

Analyse des données par le clustering

À remettre le mardi 29 novembre 2019

Ce travail consiste à développer une application de clustering pour analyser l'ensemble de données appelé "Adult" extrait du dépositoire publique des données pour l'apprentissage artificiel (https ://archive.ics.uci.edu/ml/datasets/Adult). L'ensemble des données se trouve aussi dans le répertoire TP3/donnees du répertoire publique du cours. Les principales caractéristiques de cet ensemble sont les suivantes :

- Il contient des variables ("features") numériques et catégoriques, totalisant 14 ;
- Il contient une variable de cible dont la valeur est soit ">50K" soit "<=50K", définissant deux classes d'adult ;
- Certaines données sont manquantes, notamment dans les variables "workclass", "occupation", "capital-gain", "capital-loss", "native-country" ;
- L'ensemble contient un fichier de données d'apprentissage et un fichier de données de test ;
- Les noms des variables sont expliqués dans une fichier .Name à part ;
- Tous les fichiers de données sont du format text.

Ce TP est à remettre le vendredi 29 novembre 2019. Votre remise doit comprendre un rapport et les programmes que vous aurez développés pour ce TP, de même que des données modifiées ou nouvelles. La remise doit se faire par **turnin** sur opus.dinf.usherbrooke.ca .

Tâches à effectuer :

Le but de l'analyse de clustering pour ce TP est d'obtenir des clusters dont chacun représente le mieux possible l'une des deux classes. Un cluster représente bien une classe si les deux conditions suivantes sont satisfaites : (1) la très grande majorité des "personnes adulte" de l'ensemble d'apprentissage, si ce n'est pas la totalité, du cluster appartiennent à la classe ; (2) la très grande majorité des "personnes adulte" de l'ensemble de test, si ce n'est pas la totalité, appartenant au cluster appartiennent à la classe également. Les principales tâches du TP sont listées ci-dessous. **Pour les étudiants du cours IFT799, ce TP contient une tâche supplémentaire**.

— **Tâche 1** : Pour l'analyse de l'ensemble de donndonnées d'apprentissage, effectuer le clustering des données catégoriques. Vous êtes libres de choisir l'approche par partition, l'approche hierarchique, ou une autre approche jugée appropriée ;
— **Tâche 2** : La Tâche 1 implique que vous devez convertir les variables numériques en des variables catégoriques. Cette conversion peut se faire par le clustering ou par une approche plus heuristique par observation de la distribution de chaque variable numérique. Il est à noter que la conversion des donneées en celles catégoriques offre une façon de traiter le problème des données manquantes ;
— **Tâche 3** : À la place de déterminer un nombre optimale de clusters, vous testez le tout pour 10 (ou maximum 20) valeurs de K variant de 2 à 75. Pour chaque valeur de K, faire le clustering sur les données d'apprentissage pour obtenir le taux d'erreur (d'apprentissage), puis tester le résultat de clustering sur les données de test pour obtenir de le taux d'erreur de test. Ce dernier consiste à assigner, selon la mesure de similitude de votre algorithme, chaque personne de test à un des clusters obtenus, puis calculer les taux d'erreurs de ces classifications ;
— **Tâche 4** : La qualité des clusters résultant de vos analyses se mesure par le taux d'erreur d'apprentissage et le taux d'erreur de test. Le taux d'erreur d'apprentissage globale n'est pas une simple moyenne des taux d'erreur individuelle par cluster :
— **Tâche 5** : Incorporer la sélection des variables supervisée dans votre solution. La méthode glouton est un bon choix. Il s'agit de choisir d'abord la meilleure variable individuelle, puis la meilleure couple obtenue par la combinaison avec la premiere choisie, puis la meilleure triple obtenue par la combinaison avec la meilleure couple choisie, etc. ;
— **Tâche 6** (**travaux supplémentaires pour IFT799**) En plus de la Tâche 5 qui consiste à choisir les varaibles en fonction des informations de classe, vous devez concervoir une façon de choisir les variables en supposant qu'il n'y pas d'information de classe durant le processus de clustering.

Comme pour les autres TPs, vous avez toujours beaucoup de liberté pour développer vos propres solutions. Vous devez faire preuve d'imagination. Il n'y a pas de meilleure solution et ce n'est pas le but du TP non plus. Le développement d'un esprit critique, la recherche de solutions et le savoir de "se défendre" (justification) sont bien plus importants. C'est pourquoi, vous devez mettre du temps et de l'energie pour bien rédiger votre rapport. Votre rapport doit décrire clairement les différentes étapes de traintement incluant les prétraitements effectués, les résultats obtenus, vos commentaires et vos conclusions. Parmi les résultats a presenter, vous devez inclure pour chaque valeur de K, les variables choisies, le taux d'erreur global d'apprentissage et le taux d'erreur global de test. N'oubliez pas de remettre aussi vos données modifiées. **Pour les étudiants de l'IFT799**, vous devez évidemment inclure aussi les variables choisies de façon non supervisée, de même que le taux d'erreur global d'apprentissage et le taux d'erreur global de test correspondants.