Travaux Pratiques - Fouille de Données

MIAGE

Année universitaire 2018-2019

1 Avant-propos

Lors de ces travaux pratiques, nous allons utiliser le logiciel KNIME, disponible gratuitement à l'adresse `https://www.knime.org/downloads/overview`, sans besoin de s'enregistrer (étape 2). Si vous n'êtes pas administrateur de la machine sur laquelle vous travaillez, pensez bien à télécharger la version self-extracting, sinon prenez la version adaptée à votre système d'exploitation. Tout au long de ces séances, vous prendrez des notes, essaierez des fonctions qui ne sont pas nécéssairement demandées dans le sujet, bref vous fouillerez le logiciel ! Ces notes vous aideront à réaliser le comtpe-rendu de TP, à rendre par binôme. La date de remise est fixée au **7 Décembre**.

2 Introduction

Pour cette introduction, nous allons utiliser un jeu de données provenant du recensement américain, collecté en 1994. Ce jeu de données est disponible à l'adresse suivante
`https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data`,
et vous trouverez une description de celui-ci à l'adresse
`https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names`.
Prenez le temps de bien lire cette description, et identifiez le volume et la dimension de ce jeu de données. Pour quel type de tâche ces données ont-elles été initialement utilisées ?

Pour créer le premier workflow, il faut charger des données, ce que l'on peut faire via la section IO. En double cliquant sur File Reader, le bloc correspondant se place dans l'espace du workflow. Un clic droit sur ce bloc permet de le configurer. Charger le jeu de données

Réaliser un premier workflow dont l'objectif est de discrétiser un des attributs numériques du jeu de données, l'attribut Age. Cherchez dans la section Manipulation un outil vous permettant de mettre chaque valeur de Age dans un sac. En particulier, vous procéderez à un binning à fréquence égale en 5 sacs, et chaque valeur sera remplacée par le nom des extrémités des intervalles.

Pour aller un peu plus loin,

- Filtrer les lignes pour ne conserver que les individus dont l'âge est compris entre 18 et 60 ans (inclus).
- Excluez les lignes dont la colonne Education correspond à l'expression régulière "*th.*".
- Supprimer la colonne Age une fois qu'elle aura été discrétisée.

Nous n'avons pour le moment pas visualisé les données. Cette visualisation revêt un aspect important en fouille de données. La section Views contient un certain nombre d'outils permettant cette visualisation.

- visualiser la répartition en âge des individus
- visualiser la distribution du salaire en fonction de l'âge ou du niveau d'étude

Tout ceci est un peu terne, ajoutons un peu de couleur pour aider la visualisation. Ajoutez une coloration en fonction de l'attribut Income, et visualisez le nuage de points, fonction de cette variable et le secteur professionnel. Pour terminer cette introduction, nous allons ajouter quelques éléments statistiques. Connectez ce bloc, puis étudiez les trois sorties et la visualisation qu'il permet.

3 Arbres de décision

Dans cette partie, nous allons étudier un nouveau jeu de données concernant les abonnements téléphoniques. Lors du premier abonnement d'un client, un téléphone est souvent offert, mais cela coute cher de répéter ce processus à chaque renouvellement. L'objectif est de savoir, trois mois à l'avance, si le client a des chances de partir. Si c'est le cas, on peut être tenté de lui offrir un nouveau téléphone. Les données sur lesquelles nous allons travailler sont disponibles sur Madoc (cell.csv). Avant de vous aventurer avec Knime, ouvrez ce fichier avec un éditeur de texte, et essayer de déterminer quels sont les caractéristiques de ce jeu de données (volume, dimension, attributs décrits, prédiction).

Créez un nouveau workflow dans lequel vous aller charger les données, puis procédez à une première analyse statistique de chacun des variables descriptives. Visualisez ces données en vous aidant des outils utilisés dans la partie précédente (répartition, distributions conditionnelles, etc ...)

La variable Churn indique si oui ou non le client renouvelle son contrat (on dit qu'il churn s'il ne renouvelle pas). L'objectif est d'être capable de prédire, en fonction des autres attributs, si un client va ou non renouveler son contrat. Mettez en place le workflow permettant d'apprendre un arbre de décision sur ces données. Attention, souvenez-vous que l'on n'apprend pas un modèle sur l'ensemble des données disponibles, mais que l'on sépare ces données en deux sous-ensembles d'apprentissage et de validation.

Analyser l'arbre obtenu: quelles sont les variables importantes ? quelle est la qualité de cet arbre ?

Nous allons maintenant modifier les paramètres du modèle. Par exemple, faites varier les éléments suivants

- mesure de sélection du meilleur critère,
- élagage ou non,
- nombre minimum d'observations par nœud,

et analysez l'impact de ces paramètres sur le modèle produit.

En procédant à la même analyse sur le jeu de données Adult de la partie précédente, que peut-on conclure par rapport à ces différents paramètres ?

4 Règles d'association

Dans cette partie, vous ne serez pas guidé comme c'était le cas dans les deux parties précédentes. L'objectif est de procéder à l'analyse de panier, et en particulier de trouver des règles de type $x \to y$ où x et y sont des produits que l'on peut acheter dans une épicerie. Vous trouverez sur Madoc un fichier groceries.csv, que vous chargerez et dont vous apprendrez des règles. La manière de retranscrire votre analyse et les résultats obtenus sur ce jeu de données est laissée libre à votre jugement (des illustrations, des statistiques exploratoires sont tout de même toujours les bienvenues).

5 Regroupement

Dans cette partie, nous allons travailler sur un jeu de données portant sur des films. Ce jeu de données est disponible sur Madoc, sous le nom de fichier movie.data. Comme d'habitude, déterminez le volume, et la signification de chacun des attributs utilisés pour décrire les films. Avant de passer à la fouille exploratoire par statistiques et visualisation, faites en sorte que chaque ligne soit identifiée par le titre du film correspondant.

On observe beaucoup de données manquantes dans ces données. En particulier, le budget des films est relativement souvent manquant. Dans la suite de l'étude, nous ne considérerons que les films pour lesquels on dispose de cette information. De même, l'attribut MPAA est souvent manquant, vous ne tiendrez pas compte de cet attribut dans la suite du travail. Discrétiser les années de sortie de films par décennies, et les votes par unité. Vous trouverez ensuite une série de questions, données à titre d'exemple, pour vous aider à en apprendre un peu plus. Libre à vous de vous poser d'autres questions.

Que peut-on dire de l'évolution du nombre de films par rapport au début du XXième siècle ? Y-a-t-il des genres de films plutôt mieux notés ? Quels sont les genres de films les plus fréquents ? Quelle a été l'évolution de la production de film d'animation depuis 1900 ? Chacun des ces films n'est pas spécifiquement rattaché à une classe, et nous allons donc essayer de les regrouper, de manière non

supervisée. Vous utiliserez l'algorithme vu en cours pour procéder à ce genre de regroupement. Vous ferez particulièrement attention au fait que cet algorithme utilise une distance entre observations, ce qui nécessite un certain nombre de pré-traitements sur les données. Visualisez les groupes obtenus sur différents plans.

Chaque film est décrit par un certain nombre d'attributs, et la multiplicité de ces attributs fait que la visualisation des données n'est pas toujours facile. Nous allons donc procéder à une analyse en composantes principales pour observer ces données en deux dimensions. Visualiser, sur le plan principal, les clusters obtenus par l'algorithme k-means.