Licence d'Informatique 2 (C5–160412)
Analyse de Données Utilisateur – TP 1

Carl FRÉLICOT – Dpt Info / Lab MIA

Orange 3 est un logiciel libre spécialisé en *data mining* téléchargeable ici http://orange.biolab.si/

Développé en C++ et Python, il présente deux modes d'utilisation : écriture de scripts (python) et programmation graphique ; c'est ce second mode que nous allons utiliser. Il repose sur l'utilisation d'icônes (appelées *widgets*) représentant des éléments de traitements particuliers accessibles par menus/onglets spécialisés en :

- manipulation de données — *Data*
- visualisation graphique — *Visualize*
- classement (classification supervisée) — *Classify*
- régression — *Regression*

...

On les dispose dans une fenêtre graphique soit en cliquant dessus soit par glisser-déposer, que peut relier de sorte de créer des chaînes de traitements (appelés *workflow*). Le principe, très simple, consiste à construire un schéma de traitements (appelé *canvas*) à exécuter. Chaque *widget* est comme une fonction, avec des entrées et des sorties. Un double-clic permet de le paramétrer. Il s'exécute alors s'il est correctement paramétré, lorsque ses données d'entrées sont disponibles, c'est-à-dire lorsque celui qui le précède a terminé son traitement et rendu son résultat.

De nombreux jeux de données (*data sets*) sont à disposition dans un format natif (extension .tab), mais on peut également créer des données ou lire des fichiers dans d'autres formats (de tableur, par exemple). Les attributs/variables observé(e)s, appelés *features*, possèdent un type (*Numerical*, *Categorical/Nominal*, *String*, *Time*) et peuvent posséder un rôle (*feature*, *target*, *meta*). Des données peuvent aussi manquer...

1. Mon premier schéma

(a) Lancez `Orange`, puis ouvrez un nouveau schéma (*canvas*) en sélectionnant `New`.

(b) Dans l'onglet `Data`, prenez l'icône (*widget*) `File`, déposez-la dans le schéma, puis ouvrez-la de sorte de lire les données du fichier `heart_disease.tab`

(c) Connectez l'icône `Data Table` permettant de visualiser le tableau de données, puis identifiez les types et éventuels rôles des variables observées (colonnes). Vous vérifierez à l'aide de l'icône `Info`.

(d) En AD, on peut avoir besoin de discrétiser des données continues, ou inversement. Trouvez le moyen de discrétiser l'âge des indidividus

(e) Sauvegardez votre schéma sur votre compte.

2. Données de TD-1

(a) Essayez de charger les données contenues dans le fichier `TD1-ABC.xls`
S'il y a un problème, tentez de l'identifier ou demandez à l'enseignant.

(b) Connectez l'icône `Feature Statistics`. Retrouvez-vous tout ou partie des résultats numériques calculés en TD ?

(c) Trouvez le moyen de visualiser les *boxplots*. Retrouvez-vous ceux du TD ?

(d) La feuille 2 du fichier contient l'ensemble des valeurs indépendamnent du type (A, B ou C). Centrez et réduisez les données à l'aide de l'icône `Preprocess`. Comment vérifier ?

3. Données de TP-1

Chargez les données contenues dans le fichier `anscombe7.xlsx` et décrivez-les. L'objectif de cet exercice est de vous laisser aux commande de l'analyse, en vous appuyant sur nos échanges en TD/TP. Autrement dit : que pouvez-vous extraire comme information(s) pertinente(s) sur ces données, qui pourrai(en)t être très utile(s) pour leur compréhension au client/décideur qui vous les aurait confiées ?