Caractérisation des classes issues d'un clustering

Création d'un package pour R

Objectif du projet

- Création d'un package permettant
  1. De caractériser de manière univariée (variables prises individuellement) ou multivariée les classes issues d'un processus de classification automatique (clustering). Elle doit comporter une composante graphique forte.
  2. De calculer les mesures d'évaluation des partitions.
- Package que l'on peut installer directement à partir de GitHub
- Le package intègre un fichier d'aide en anglais aux normes R
- C.-à-d. description des fonctions, de leurs paramètres, des objets fournis en sortie, de la lecture des résultats, avec des exemples d'utilisation (voir par ex. ?lm du package stats)
- **A réaliser en groupes de 3 étudiants** (sauf 1 groupe)

Cahier des charges

- Ecriture d'une ou plusieurs fonctions (ex. une fonction pour la caractérisation numérique univariée, une autre pour le multivarié, une fonction pour les graphiques, etc. ; à vous de voir l'organisation la plus pertinente)
- Elles prennent en entrée :
  1. Une variable catégorielle indiquant les classes attribuées par un algorithme de clustering
  2. Plusieurs variables descriptives catégorielles et/ou quantitatives (mélangées ou prises de manière distinctes)
  3. (Eventuellement) Une variable catégorielle indiquant les vraies classes d'appartenance
  4. Des objets que vous avez créés spécifiquement avec d'autres fonctions
- Pour les indices (mesures) d'évaluation des partitions, vous ne devez pas utiliser les packages existants spécialisés dans le domaine (ex. clValid, …)
- Pour les fonctionnalités graphiques, vous devez utiliser le package « ggplot2 » de la galaxie « tidyverse ».

Cahier des charges – Remarques

- Attention aux calculs à mettre en place selon le type des variables descriptives. Bien lire la bibliographie à ce sujet. A vous de définir la stratégie à mettre en place pour appréhender correctement les différentes configurations.
- Les représentations graphiques peuvent intégrer des calculs additionnels (ex. passer par une représentation factorielle, …)
- Vous avez toute liberté pour intégrer des paramètres additionnels à vos fonctions
- S'il y a lieu de créer des classes (plutôt conseillé), vous travaillez avec la norme S3.

Bibliographie - Références

Quelques pistes :

- http://tutoriels-data-mining.blogspot.com/2016/09/clustering-caracterisation-des-classes.html
- http://tutoriels-data-mining.blogspot.com/2008/04/interprter-la-valeur-test.html
- http://tutoriels-data-mining.blogspot.com/2017/05/comprendre-la-taille-deffet-effect-size.html
- http://tutoriels-data-mining.blogspot.com/2013/11/classification-automatique-sur-donnees.html (ex. d'utilisation du composant GROUP CHARACTERIZATION de TANAGRA, inspiré de DMOD du logiciel SPAD)

- https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

… Vous trouverez pas mal de littérature en cherchant avec les mots clés adéquats… (ex. performance metrics for clustering, etc.)

A rendre

- Un **rapport** en français au format PDF de présentation de votre travail. **Il doit être rédigé en LaTeX** (source .tex doit être fourni).
- Il doit indiquer les formules et stratégies utilisées pour produire les résultats.
- Il doit décrire également l'architecture de votre programme, modules R, fonctions, détail des objets générés, description de vos implémentations en pseudo-code.

- Le projet doit être hébergé sur GitHub.
- Un tutoriel (reproductible) montrant l'utilisation des fonctionnalités de votre package doit être disponible sur GitHub.
- Le package doit pouvoir être installé directement en ligne à partir de GitHub. Il doit comporter les données exemples utilisées dans le tutoriel.
- Le code source du package et les documents associés (aide, etc.).
- Une copie du package au format ZIP directement utilisable sous R (plan B au cas où l'installation en ligne est défaillante).

Critères d'évaluation

- Qualité et clarté du rapport (en français)
- Qualité de la documentation du package (en anglais)
- Qualité de la programmation – Commentaires / documentation du code source

- Originalité des solutions proposées
- Pertinence des indicateurs proposés, de leur organisation
- Qualité des sorties graphiques
- Facilité pratique des fonctions implémentées
- Richesse fonctionnelle (options supplémentaires éventuelles)

Calendrier

- Diffusion du sujet : mardi 20 octobre 2020
- Retour attendu : mardi 8 décembre 2020 au soir
- Soutenances : mercredi 16 décembre 2020

- A faire :
  - Mettre votre projet complet (rapport, package, source, etc.) sur un drive quelconque
  - M'avertir par e-mail et m'envoyer le lien à l'adresse : ricco.rakotomalala@univ-lyon2.fr
  - Sujet : [SISE – Prog. R] Noms des étudiants