Création d'un package pour R

M2 SISE 2025-2026

Clustering de variables

Algorithmes et Outils pour l'interprétation des résultats

Objectif du projet

- Création d'un package proposant
  1. Plusieurs méthodes de clustering de variables
  2. Accompagnées d'indicateurs pour l'interprétation des résultats
- 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. ?glm du package stats)
- A réaliser en équipes de 3 étudiants (2 étudiants pour une des équipes)

Cahier des charges - Obligatoire

Implémentation d'algorithmes de clustering de variables :

1. Implémenter 3 algorithmes de clustering de variables.
2. Parmi les 3 : au moins un basé sur les techniques réallocation ; au moins un consacré au traitement des (modalités) des variables qualitatives
3. Mise en place d'une stratégie pour l'identification du nombre de clusters (au moins des approches qui aident à l'identification)
4. Intégrer des outils (tableaux, graphiques) qui permettent d'interpréter les résultats : nature des partitions, nature des groupes, degré d'appartenance aux clusters
5. A vous d'arbitrer entre ce que vous implémentez vous-même et ce que vous utilisez d'autres outils (uniquement packages référencés, ex. hclust). Documentez vos choix.

Cahier des charges - Obligatoire

La classe de calcul doit être implémentée selon la norme R6. Elle doit présenter (alignez-vous sur le formalisme des librairies telles que « Scikit-Learn / Python ») :

1. Un constructeur, avec les paramètres éventuels (à vous de voir)
2. La méthode $fit(X) qui lance la modélisation sur les données d'apprentissage. « X » représente un data frame comportant l'ensemble des variables actives.
3. La fonction $predict(X) qui prend en entrée un data frame X de variables illustratives compatible [même nombre d'observations, type des variables ??? à voir ???] avec celui présenté à fit(), et qui rattache chaque colonne de X à l'un des clusters (avec éventuellement des indicateurs numériques).
4. La procédure $print() qui affiche les informations succinctes sur le modèle (à vous de voir le contenu)
5. La procédure $summary() qui affiche les informations détaillées (à vous de voir aussi)
6. Votre objet peut exposer une série de propriétés qui peuvent être exploitées par la suite, à vous d'en définir la teneur.

Cahier des charges - Obligatoire

- Une application R SHINY qui a pour vocation de présenter les fonctionnalités de votre package. Elle doit permettre de :
  - Sélectionner un fichier de données (CSV – séparateur tabulation – au moins ; autre éventuellement, ex. « xlsx »…)
  - Choisir les variables explicatives et les variables illustratives
  - Lancer les calculs et présenter les résultats, avec en particulier une mise en valeur de vos sorties
  - Autres… ?

Cahier des charges - Optionnel

- A vous de voir les fonctionnalités additionnelles que vous jugez utiles…

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, stratégies, algorithmes utilisés pour produire les résultats. Il doit décrire également l'architecture de votre programme, modules R, fonctions, détail des objets générés, description de vos algorithmes et implémentations en pseudo-code, des informations que vous souhaitez mettre en avant dans vos sorties (une douzaine de pages…)
- Soyez précis sur vos références (bibliographique). Je dois pouvoir les consulter.
- Le projet doit être hébergé sur GitHub.
- Le package doit pouvoir être installé directement en ligne à partir de GitHub. Il doit comporter les jeux de données exemples utilisés dans le tutoriel.
- Le code source du package et les documents associés (aide, etc.).
- Une copie du package au format ZIP (ou tar.gz) directement utilisable sous R (plan B au cas où l'installation en ligne est défaillante).
- L'application R SHINY avec tous les éléments permettant de le faire fonctionner.
- Sur GitHub, un tutoriel (reproductible) en anglais montrant l'utilisation des fonctionnalités de votre package doit être disponible sur GitHub. Montrez au moins le fonctionnement de $fit() et description des sorties ; puis montrez également le fonctionnement de l'application SHINY.

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
- Qualité de l'application interactive R SHINY

- Qualité du package, notamment l'installation en ligne
- Fonctionnement sécurisé (pas de plantage, calculs corrects)
- Rapidité d'exécution et appréhension des grandes volumétries (dimensionnalités)

  (attention, ce serait très dommage que votre package plante à cause d'une librairie externe que vous utilisez, à vous de vous assurer de leur robustesse et intégrité !)
- Richesse fonctionnelle (au-delà du cahier des charges obligatoire)
- Utilisabilité (facilité pratique) des fonctions implémentées

Calendrier

- Diffusion du sujet : vendredi 14 octobre 2024
- Retour attendu : dimanche 23/11 au soir
- Soutenances : semaine du 01/12

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