Fiche TP 05 :
Analyse de données

Master 1 informatique
2018-2019

Exercice 1 : Meilleur sous-espace

Le fichier `hc_3.dat` contient les performances obtenues par la recherche locale hill-climbing first-improvement pour les 60 sous-espaces possibles synchronisant 3 cellules (voir projet). Le descriptif de ce fichier est donné dans README.txt.

Le but de l'exercice est de savoir quel est le sous-espace le plus performant.

Questions :
0 - Lire le fichier et enregistrer les données dans un data frame (cf. `read.table`)
a - Calculer pour chaque sou-espace la performance moyenne et enregistrer le résultat dans un data frame (utiliser `summaryBy` si possible).
b - Repérer les 3 sous-espaces les plus performants en moyenne.
c - Observer les distributions des performances pour ces 3 sous-espaces.
d - Réaliser un test statistique pour montrer que les moyennes sont ou non significativement différentes.
e - Conclure.

Exercice 2 : Corrélation

Le fichier `rnd_3.dat` contient le résultat de l'échantillonnage de chaque sous-espace de manière aléatoire. Le descriptif de ce fichier est donné dans README.txt.

Le but de l'exercice est de savoir si cet échantillonnage aléatoire est corrélé avec les performances du hill-climber.

Questions :
a - Quelles sont les mesures possibles de corrélation ?
b - Existe-t-il une différence entre corrélation et causalité ?
c - Après la lecture du fichier `rnd_3.dat` dans un data frame, ajouter une colonne qui correspond à la performance moyenne du hill-climber.
d - Tracer et calculer les différentes corrélations entre les variables "interessantes" (voir `pairs`, ou `ggpairs` de la librairie GGally).
e - Calculer le modèle linéaire entre les performances du hill-climber et le ou les variables vous semble pertinentes.
f - Conclure.