Clustering et Visualisation

Réaliser une classification n'est pas si simple. A travers un exemple nous allons évalué d'une part plusieurs algorithmes de classification vus en cours et d'autre part tester certaines méthodes de la réducton de dimension dans un un objectif de visualisation de la partition.

On dispose d'un fichier de 1000 observattions décrites par 15 variables quantitatives. Ce fichier a été simulé avec un certain nombre de classes qui reste à définir car on est dans le cadre de l'apprentissage non supervisé.

1. Importer votre fichier excel et vérifier la taille des données.
2. Visuliser les nuages des points en croisant les variables deux à deux.
3. Visualiser l'ensembe des observations (individus) sur votre premier plan factoriel en utilsant une analyse en composantes principales, que peut-on dire ?.
4. Visualiser les plans 1×3, 2×3, que peut-on dire ?
5. On chercher à partitionner l'ensemble des observations, utiliser le package `Nbclust` pour réaliser un kmeans et des cah avec différents critères d'grégation, soit un total de 5 méthodes (kmeans, average, ward, single complete). Sauvegarder toutes les partitions obtenues avec les 5 méthodes. Par exemple:

   res.NbClust.kmeans=NbClust(X, method="kmeans",min.nc=2, max.nc=8, index = "all")

   res.NbClust.kmeans$Best.partition

   res.NbClust.single=NbClust(X, distance="euclidean", method="single",min.nc=2, max.nc=8, index = "all")

   res.NbClust.single$Best.partition

6. Quel nombre de classes peut-on proposer ?
7. On décide d'utiliser les algorithmes issus de l'approche mélange. On retient l'algorithme `EM`. Utiliser les deux packages `Rmixmod`[1] et puis `mclust`[2]. Choisir le modèle approprié (avec le nombre de classes proposé). Sauvegarder les partitions obtenues à l'aide des deux packages.
8. On décide de visualiser les classes de l'ensemble des observations avec la fonction **MclustDR** du package `mclust`.
9. Importer le vecteur des classes (vraie partition). Réaliser une étude comparative entre des résultats des différents algorithmes en terme de qualité de la partition. On utilsera dans un premier temps, le taux de mal classés issu de la table de confusion.
10. Reprendre les questions 3) et 4), visualiser le plan 1×15, que peut-on dire ?
11. Jusqu'à présent, la réduction de la dimension et la classificationont été réalisées d'une manière séparée, on décide cette fois-ci de réaliser les deux tâches d'une manière simultanée. Pour ce faire on va tester le package `clustrd`[3] et la fonction **cluspca**.

[1] https://cran.r-project.org/web/packages/Rmixmod/Rmixmod.pdf
[2] https://cran.r-project.org/web/packages/mclust/mclust.pdf
[3] https://cran.r-project.org/web/packages/clustrd/clustrd.pdf