TP : Clustering et Modèles de mélange

Ci-attaché une liste de tables de données décrites dans Table 1. Ces tables qui sont communément utilisées dans la communauté image sont utilisées pour évaluer des nouvelles méthodes de clustering. Le but de ce projet est de mettre en pratique certaines méthodes de clustering et de les évaluer en utilisant des indices appropriés.

Table 1: Description des données images: nombre de lignes, colonnes et classes

datasets	# samples	# features	# classes
JAFFE	213	676	10
MNIST5	3495	784	10
MFEA	2000	240	10
USPS	9298	256	10
OPTIDIGITS	5620	64	10

1. Faire une brève introduction concernant ces tables de données.
2. Importer ces tables en utilisant la librairie `R.matlab`.
3. Visualiser l'ensemble des observations (individus) sur votre premier plan factoriel en utilisant une analyse en composantes principales, que peut-on dire ?. Toute autre méthode de visualisation peut également être suggérée.
4. On cherchera à partitionner l'ensemble des observations, utiliser le package `Nbclust` pour réaliser un kmeans et des cah avec différents critères d'agrégation, soit un total de 5 méthodes (kmeans, average, ward, single complete). Sauvegarder toutes les partitions obtenues avec les 5 méthodes. Quel nombre de classes peut-on proposer ?
5. Réaliser du clustering à partir des deux premières composantes (ACP); la fonction `HCPC` peut être utilisée. Quel nombre de classes peut-on proposer ?
6. A l'aide de matrice de confusion, comparer les partitions obtenues avec `HCPC` et toutes les partitions obtenues avec `Nbclust`, que peut-on dire ?
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 et visualiser les classes dans le premier plan factoriel. Quel nombre de classes peut-on proposer ?
8. Comparer les partitions de `mclust` et `Rmixmod`, que peut-on dire ?
9. On décide de visualiser les classes de l'ensemble des observations avec la fonction **MclustDR** du package `mclust`. Préciser le rôle de cette fonction avant son utilisation.
10. Importer le vecteur des classes (vraie partition). Réaliser une étude comparative entre des résultats des différents algorithmes en termes de qualité de la partition avec le vrai nombre de classes. On utilisera dans un premier temps, le taux de mal classés issu de la table de confusion puis on évaluera cette qualité à l'aide des deux mesures la NMI et l'ARI.
11. Visualiser à l'aide de `t-SNE` et `UMAP` les classes obtenues.
12. On décide d'utiliser une approche deep learning pour la réduction de la dimension. Appliquer l'algorithme EM (`mclust`) sur les tables de dimension réduites obtenues via un autoencoder. Evaluer les performances en termes de clustering en utilisant l'accuracy (taux d'élements bien classés), la NMI et l'ARI. Que peut-on dire ?.
13. Question facultative. Il exite d'autres approches combinant simultanément une réduction de la dimesnion et un clustering via la méthode *mixture of factor analysis* (MFA) (disponible dans le package `EMMIX`). Faire une brève introduction de cette méthode puis l'appliquer sur les tables originales avec le vrai nombre de classes. Que peut-on dire ?

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