Projet: Réduction de la dimenion, clustering et Modèles de mélange

Projet à réaliser par binome ou seul, A rendre au plus tard le 26 mars. Attention, ce projet permettra d'attribuer deux notes : une sur la réduction de la dimension et une autre sur le modèle 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
UMIST	360	644	20
MNIST5	3495	784	10
MFEA	2000	240	10
COIL20	1440	1024	20
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.
5. Réaliser du clustering à partir des deux premières composantes; la fonction `HCPC` peut être utilisée.
6. Réaliser un *spectral clustering* en utilisant un package approprié.
7. Quel nombre de classes peut-on proposer ?
8. 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.
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 terme de qualité de la partition. On utilsera dans un premier temps, le taux de mal classés issu de la table de confusion puis on évaluera cette qualité à l'aide la NMI et l'ARI.
11. Jusqu'à présent, la réduction de la dimension et la classificationont ont été réalisées d'une manière séparée, on décide cette fois-ci de réaliser ces deux tâches d'une manière simultanée. Pour ce faire on va utiliser le package `clustrd`[3] et la fonction **cluspca**. Cette fonction permet de réaliser 2 algorithmes combinant (simultanément) l'ACP et k-means, ces deux algorithmes sont notés respectivement FKM et RKM. Un paramètre alpha peut être utilisé pour combiner ces deux algorithmes. Pour $\alpha = 0.5$ le problème se réduit à l'utitilisation de RKM, et pour $\alpha = 0$ à celui de FKM. Lorsque $\alpha = 1$ la solution se réduit à l'application de l'ACP suivie par kmeans sur les composantes principales. Attention, ne pas négliger l'option *rotate*, par exemple la rotation *varimax* permettra dans certaines situations de mieux caractériser les classes.

[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