Vous avez le choix de l'outil : soit R, soit Python

1. Supports

[TUTO R] http://tutoriels-data-mining.blogspot.fr/2015/10/classification-automatique-sous-r.html

[TUTO Python] http://tutoriels-data-mining.blogspot.com/2016/03/classification-automatique-sous-python.html

2. Données

Les données https://archive.ics.uci.edu/ml/datasets/Image+Segmentation décrivent les propriétés (variables actives, REGION.CENTROID.COL … HUE.MEAN) de 2310 photos représentant des objets désignés par la colonne IMAGE, qui joue le rôle de variable illustrative dans notre exercice.

3. K-Means sur les données « segmentation »

La détection du nombre adéquat de classes pour la méthode des centres mobiles (K-Means) est le principal thème de cet exercice.

A rendre : un rapport PDF : impression d'un .docx issu de R Markdown si vous travaillez sous R, impression d'un notebook JUPYTER si vous travaillez sous Python.

1. Chargez le fichier « **segmentation.xlsx** ».
2. Quel est le nombre d'observations ? Le nombre de variables ? Enumérez les.
3. Centrez et réduisez les variables actives.
4. Pour un nombre de classes allant de 2 à 10, réalisez une classification à l'aide des K-Means. Construisez le graphique retraçant la part d'inertie expliquée en fonction du nombre de classes. Quel serait le bon nombre K* de classes de ce point de vue ?
5. On s'intéresse à l'indice « silhouette ». Que représente cet indicateur ? Construisez le graphique d'évolution de l'indice en fonction du nombre de classes. Quel est le nombre K** de classes suggérée par l'approche ?
6. Deux résultats semblent se démarquer (*j'imagine…*) : K' = 2 et K'' = 5. Pour ces deux solutions, réalisez la succession d'opérations suivantes :
    a. Lancez les K-Means en fixant le nombre de classes à K' puis K''.
    b. Croisez les groupes avec la variable illustrative IMAGE. Décrypter les correspondances entre les classes induites par la typologie et les groupes a priori.

    A la lumière des correspondances, quelle est la partition qui paraît la plus adaptée ?
7. On souhaite mettre en œuvre une stratégie mixte c.-à-d. K-Means + CAH. Mettons que nous démarrons avec un K-Means à 15 classes [Voir http://tutoriels-data-mining.blogspot.fr/2008/10/traitement-de-gros-volumes-cah-mixte.html, page 11 et suivantes, il n'est pas nécessaire de passer par une ACP préalable dans notre cas]. Le nombre

de groupes suggéré dans le dendrogramme de la CAH est-il cohérent avec ceux détectés précédemment (K' = 2 et/ou K'' = 5) ?

8. Réalisez une ACP (analyse en composantes proincipales) normée sur les variables actives. Combien de facteurs retiendriez-vous ?
9. Pour chaque paire de facteurs retenus, projetez les individus en les coloriant selon le groupe d'appartenance issu de la stratégie mixte. Quels commentaires pouvons-nous faire ?