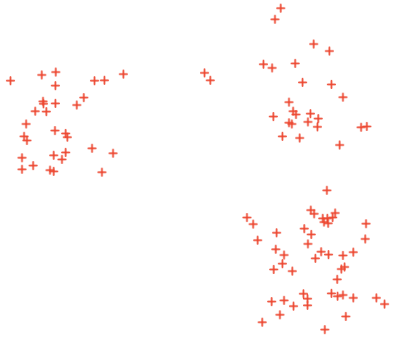
Licence d'Informatique 2
Analyse de Données Utilisateur (C5–160412)
TP 4 – Apprentissage Non Supervisé (*Clustering*)

Carl FRÉLICOT – Dpt Info / Lab MIA

1. *K-means Clustering*



(a) Dessinez trois ou quatre groupes de points en dimension 2, par exemple selon le modèle ci-contre.
(b) Trouvez le composant permettant d'exécuter l'algorithme des *k-means* pas à pas, c'est-à-dire de manière interactive et connectez-le.
(c) Indiquez le bon nombre de clusters, puis cliquez plusieurs fois sur `Randomize Positions`.
(d) Pour exécuter une itération de l'algorithme, il suffit de cliquer sur l'algorithme `Reassign Membership`. Comptez le nombre d'itérations nécessaire à l'obtention d'une partition stable.
(e) On peut aussi initialiser en positionnant les centres initiaux. Trouvez une configuration qui vous semble difficile et exécutez de nouveau pas à pas.
(f) Que va-t-il se passer si on demande un nombre de groupes différent de que ce que suggère le visuel (un de plus ou un de moins) ? Testez.
(g) Trouvez le composant `k-Means` et connectez-le aux données. Vous offre-t-il une solution pour le problème du nombre de groupes ?

2. *Clustering* Hiérarchique

(a) Chargez les données du fichier `TD4-xyzt.xlsx`, trouvez le moyen de calculer le tableau de distances (euclidienne usuelle) entre tous les points, puis de le visualiser.
(b) Calculez les distances de *Manhattan* et vérifiez visuellement (nuage dessiné en TD) ainsi qu'à l'aide du composant.
(c) Trouvez le moyen de calculer la distance cosinus entre les deux variables.
(d) Trouvez le composant permettant de réaliser une classification hiérarchique, puis retrouvez la hiérarchie de Ward vue en TD.
(e) Quelle semble être la meilleure partition ? Trouvez le moyen de récupérer une variable indicatrice du résultat.
(f) Utilisez cette indicatrice pour visualiser les points.
(g) Changez de distance et constatez le résultat.
(h) Exécutez les *K-means* avec $K = 2$, puis trouvez le moyen de comparer la partition obtenue avec celle déduite de la hiérarchie de Ward. Si c'est la même partition, changez un paramètre de sorte d'en comparer deux différentes.

3. Données Réelles

(a) Chargez les données `wine.tab`, et réalisez leur classification hiérarchique.
(b) Obtenez une partition en un nombre de groupes égal au nombre de qualités des 178 vins.
(c) Comparez le résultat (indicatrice de *cluster*) et la vérité-terrain (donnée d'expertise).
(d) Calculez (manuellement) le taux de désaccord.
(e) Les méthodes d'Analyse de Données ne sont pas mutuellement exclusives. Reprenez l'ACP de ces données, puis la réalisez le partitionnement par *k-means* du nuage réduit à un nombre judicieux de composantes principales.
(f) Recommencez la comparaison avec la vérité terrain (taux de désaccord).
(g) Le taux de désaccord est-il meilleur ? Peut-on affirmer que c'est toujours ainsi ?
(HW) Reprenez la même analyse (a) à (g) en inversant l'ordre : *k-means* dans l'espace initial et *hierarchical clustering* dans l'espace des composantes principales.
(h) Cherchez, par *k-means* et/ou par *hierarchical clustering*,

4. Digits Images

(a) Décompressez le fichier `digits0137.zip` et importez les images qu'il contient à l'aide du composant approprié.
(b) Connectez un `Data Table` afin de savoir ce qu'il en est, puis trouvez le moyen de visualiser les images.
(c) Un réseau de neurones profond a été codé dans le composant `Image Embedding` afin d'extraire 1 000 caractéristiques numériques de chaque image quelle que soit sa taille. Connectez ce composant et regardez le tableau de données résultant.
(d) Réalisez la *meilleure* classification hiérarchique au sens de la vérité-terrain en jouant sur la distance entre points et celle entre groupes. Vous pourrez visualiser les images par groupe.
(e) Comment réduire intelligemment le nombre de variables en amont ? Testez.

5. Ponctuation

Chargez le fichier `auteurs.xlsx`, et... débrouillez-vous ;)