

TP EX23 : Algorithme des k-moyennes

Pour une réalisation en Python de cet exercice, utiliser le document « *triedpython* », et consulter la table d'équivalence Matlab/Python des noms de fonction.

Soit un ensemble de données d'apprentissage non labellisées en dimension 2 contenues dans le fichier **Data1.mat**. On dispose pour un ensemble d'exemples à classer de deux caractéristiques. Cet ensemble d'apprentissage comprend 132 exemples qui ont été simulés selon 3 gaussiennes.

1°) Expliquer le principe de fonctionnement, la nature et les différentes étapes d'un algorithme de k-moyennes, puis retrouver les différentes étapes spécifiquement associées à cet algorithme dans le programme **kmoys.m** fourni.

2°) Exécuter le programme **demokmeans.m** pour effectuer une classification non supervisée des données par un algorithme des k-moyennes. (Vous pouvez modifier le temps de pause de *kmoys.m* si l'exécution est trop lente).

En faisant varier le nombre de groupements réalisés par l'algorithme ($k=2, 3, 5, 10, 15, 20$), étudier les variations de « l'inertie intra ». Que remarquez-vous ? Qu'en pensez-vous ?

On fera une figure pour chaque cas et un tableau pour justifier les résultats, on indiquera la cardinalité de chaque groupe.

3°) On appelle forme forte les éléments de l'ensemble d'apprentissage qui ont toujours été classés ensemble au cours de plusieurs classifications (initialisations différentes). Ecrire un script qui permet de trouver le nombre de formes fortes et la cardinalité de ces formes. Déterminer les formes fortes pour quelques valeurs de k ($k=2, 3, 4, 5$ par exemple), et faire 25 essais à chaque fois.

Indication pour la réalisation du script :

- Vous aurez à effectuer une boucle sur le nombre d'essais dans laquelle la classification obtenue à chaque itération devra être stockée (dans un tableau).
- La fonction **unique** doit vous permettre ensuite de trouver les classifications qui identifient les formes fortes.
- Pour chaque forme forte identifiée, vous devez pouvoir, à l'aide des fonctions **ismember** et **find**, regrouper les individus qui les composent. Présenter les formes fortes sur une figure. Indiquer leur cardinalité.

4°) Frontière de décision (question Facultative) :

On vous demande de représenter les frontières de décision dans le cas qui présente le moins de formes fortes de la question 3° précédente. Pour cela, vous aurez besoin de définir un ensemble constitué des points d'un maillage 2D; on vous propose les points de -1.8 à 1.8 par pas de 0.05 par exemple. Vous pouvez soit réaliser une boucle imbriquée sur les 2 dimensions, ou utiliser les fonctions **meshgrid** et **reshape**.

Vous aurez besoin d'affecter les éléments de votre ensemble à une classe. Pour cela, vous pouvez coder cette affectation par vous-même en vous inspirant de la fonction **kmoys**, mais afin d'éviter une perte de temps trop importante, vous pouvez aussi utiliser la fonction **kclassif**.

A titre de divertissement et de curiosité, vous pouvez également essayer des valeurs de k plus élevées.