TP ANREC k-means

L'objectif de ce TP est d'abord d'étudier le comportement de l'algorithme du k-means en fonction des mesures de similarité ou de dissimilarité utilisées sur des jeux de données simples (deux variables). Dans un second temps, vous serez confrontés à un jeu de données complexes pour lequel vous produirez une ou plusieurs classifications et une interprétation de celle(s)-ci.

Le TP est à rendre le 1er février à midi. Il sera corrigé et interviendra en bonus/malus sur la note d'ANREC. Ce bonus/malus pourra aller jusqu'à 2 points. Tout plagiat entrainera automatiquement un malus de 2 pour l'ensemble des binômes impliqués. Les groupes (effectif maximum 2) devront m'être fournis pour aujourd'hui 9h30 et seront définitifs.

Travail attendu

Algorithme du k-means

La première étape est l'écriture dans un langage de programmation que vous choisirez d'un algorithme du k-means. Celui-ci devra être générique au sens où :

- la dimension des données (nombre de variables et nombre de lignes) ne devra pas être codée en dur
- la mesure de similarité/dissimilarité ne devra pas être incluse dans l'algorithme mais devra faire l'objet d'une fonction extérieure qui pourra être passée en paramètre.

Jeux de données simples

Sur deux jeux de données très simples à deux dimensions, vous devrez montrer en quoi la mesure de similarité/dissimilarité joue sur la classification produite. Pour cela, vous devez tester **au moins** deux mesures différentes sur chaque jeu de données. L'influence de l'étape d'initialisation du k-means doit être également être testée.
Les résultats seront montrés graphiquement (et éventuellement analysés) par des diagrammes de points (x,y) où les couleurs représenteront les clusters auxquels les points auront été affectés

Jeu de données complexes

Sur ce jeu de données complexe représentant les notes des étudiants (anonymisés) d'une promotion antérieure, vous devrez produire une ou plusieurs classifications et produire à chaque fois une analyse de la pertinence de cette classification. Le nombre de classes n'est pas connu a priori.

A rendre

- Code du k-means et des différentes mesures de similarité/dissimilarité (un lien github suffit)

- Graphiques et analyses sur les jeux de données simples
- Classification et analyse sur le jeu de données complexe

Annexe : format des jeux de données

Les fichiers de données comportent une donnée par ligne. Les valeurs de chacune des variables sont séparées par des tabulations. Le format des fichiers source peuvent être modifiés (par chercher/remplacer par exemple) si cela simplifie l'opération de lecture.