

# Machine Learning

## Clustering : K-Means

### Objectifs

- Familiarisation avec le Clustering

### Enoncé

Le choix du langage de programmation est libre. Dans cet exercice il s'agit de programmer l'algorithme des K-Means décrit dans le document 02 du cours (pages 28-38).

- Initialiser k-centroïdes
- attribuer les points aux centroïde le plus proche (Manhattan / L2)
- Tant que les centroïdes bougent:
  - Pour chaque cluster calculer le point central du cluster (moyenne des coordonnées)
  - noter si ancien centroïde  $\approx$  nouveau centroïde
  - attribution des points aux nouveaux centroïdes
    - calcul de la somme des distances au sein d'un cluster
      - fonction objectif à minimiser
      - plot la somme de la somme des distances à chaque itération
  - afficher les nouveaux cluster à chaque itération

Deux jeux de données sont à disposition dans *CyberLearn* pour expérimenter l'algorithme : *Student* et *Iris*. Pour le jeu de données *Student*, faire des expériences en variant le nombre de groupes et assigner pour la visualisation une couleur distincte pour chaque groupe. La fonction objectif à utiliser est la distance au centroïde pour chaque point au cluster. Au fil des itérations la fonction objectif devrait baisser.

Les données *Iris* contiennent une étiquette indiquant la classe d'appartenance (c'est la dernière donnée de chaque ligne). Les données ne devront pas être visualisées. Cependant, à chaque itération de l'algorithme des K-means, chaque cluster sera étiqueté par la classe majoritaire et il faudra calculer le taux de classifications correctes. Spécifiquement, pour chaque cluster on aimerait savoir combien de cas sont corrects.

Ce travail pourra être réalisé par groupe de deux personnes (au maximum) ; il faudra le rendre sur *Cyberlearn* au plus tard le **10 octobre (à midi)**. Le rendu est un listing du programme (pas de rapport) indiquant au tout début les **noms** des personnes l'ayant réalisé.

Rappel : cette série d'exercices n'est pas notée, mais jugée suffisante ou insuffisante. **Une série jugée insuffisante baissera la note semestrielle de 0.5 points.**