# Model-based learning

## Projet sous le logiciel R - 2024-2025

L'objectif du projet est de créer un package R permettant de réaliser un algorithme de clustering pour des données quantitatives.

**L'algorithme** Le modèle considéré sera un modèle de mélange gaussien, estimé à l'aide d'un algorithme EM.
Vous implémenterez l'estimation de deux modèles gaussiens : le modèle complet (matrices de variances libres) et le modèle homoscedastique (matrices de variances communes pour toutes les classes).
Votre fonction devra prendre en entrée :

- le jeu de données,
- le nombre de clusters,
- le nombre d'initialisation de l'algorithme EM (en spécifiant une valeur par défaut),
- le modèle utilisé (en spécifiant une valeur par défaut).

En sortie, l'algorithme devra retourner :

- les probabilités a posteriori pour chaque individu d'appartenir à chaque clusters,
- la partition estimée par maximum a posteriori,
- les proportions, moyennes et variance de chaque clusters,
- la valeur du critère BIC.

**Le package** Votre fonction R deva être incorporé à un package R.
Ce package comportera donc une unique fonction, et comme tout package R il devra comporter une aide pour cette fonction, incluant un exemple d'utilisation (un exemple simple qui soit rapide d'execution).
Pour la création du package R, vous pourrez suivre ce document pour savoir comment procéder.

**Application** Enfin, vous utiliserez votre algorithme pour réaliser un clustering sur les données MNIST (http://yann.lecun.com/exdb/mnist/).
Le jeu de données sur lequel vous travaillerez sera `train-images-idx3-ubyte.gz`. Vous pourrez travailler dans un premier temps sur un sous-échantillon de taille réduite de ces données (le temps de calcul peut être long).
Le nombre de clusters sera choisi en fonction du critère BIC (n'hésitez pas à introduire du calcul parallèle lorsque vous devez lancer l'algorithme pour différents nombres de clusters).
Vous interprèterez les images en affichant les moyennes de chaque cluster, et vous pourrez également comparer vos résultats aux vraies étiquettes des données (indiquant à quel chiffre correspond chaque image, disponible avec les données).

**Rendu pour le 01/11/2024**

- le package R en .tar.gz qui soit installable sur tout type de machine (comme tout package disponible sur le CRAN)
- un rapport de maximum 10 pages écrit en LaTeX.