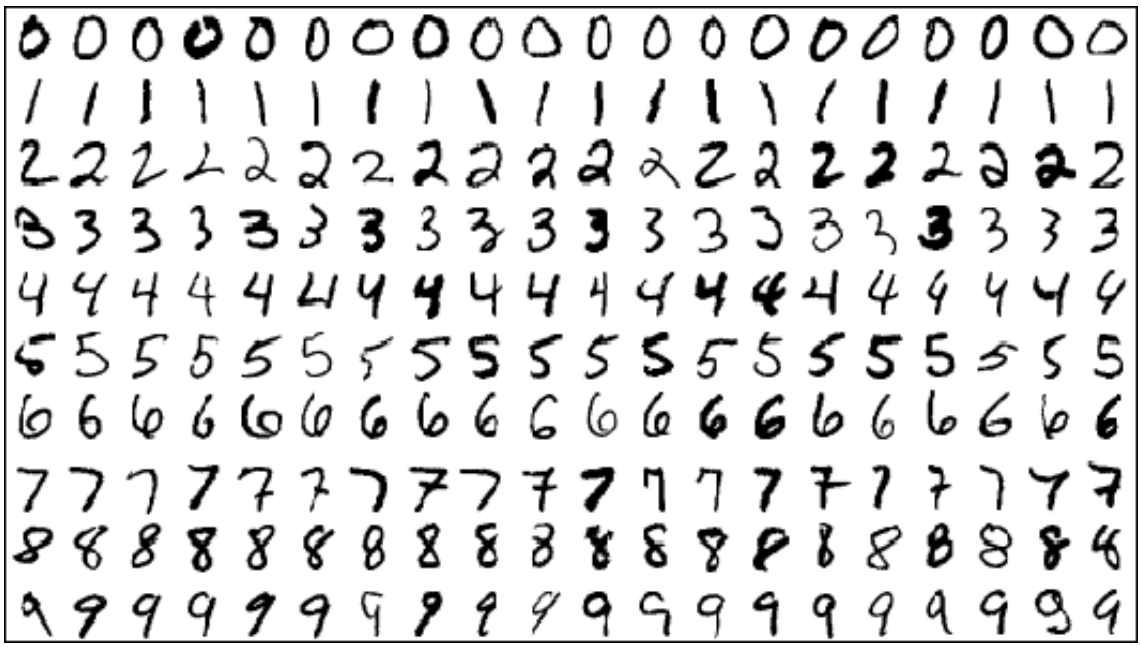
JULIEN JACQUES UNIVERSITÉ DE LYON

Fouille de données - Clustering

Travaux pratiques sous le logiciel R

Nous allons travailler sur les données MNIST (disponibles sur `http://yann.lecun.com/exdb/mnist/`).



Cette base de données comporte un jeu de 60000 images (`train`) et un second jeu de 10000 images (`test`). Chaque image est de taille 28x28 et représente un chiffre manuscrit.

Notre objectif sera de réaliser un clustering sur la base de données `train`.

1. Récupérer les données MNIST sur `http://yann.lecun.com/exdb/mnist/` ainsi que le fichier pour les manipuler `mnist.r` sur le site de J.Jacques.
2. Décompresser les fichiers `.gz`, et à l'aide du ficher `mnist.r`, charger les données sous R et afficher une des images.
3. Comment sont stockées les images sous R ?
4. Commencer par extraire aléatoirement un échantillon de 1000 images.
5. Réaliser un clustering sur cet échantillon à l'aide de la méthode CAH et de la méthode des kmeans. Utiliser une technique pour choisir le nombre de clusters.
6. Interpréter votre clustering en représentant les images moyennes de chaque cluster.
7. Recommencer cette fois avec la base de données complètes des 60000 images et avec 10 clusters.