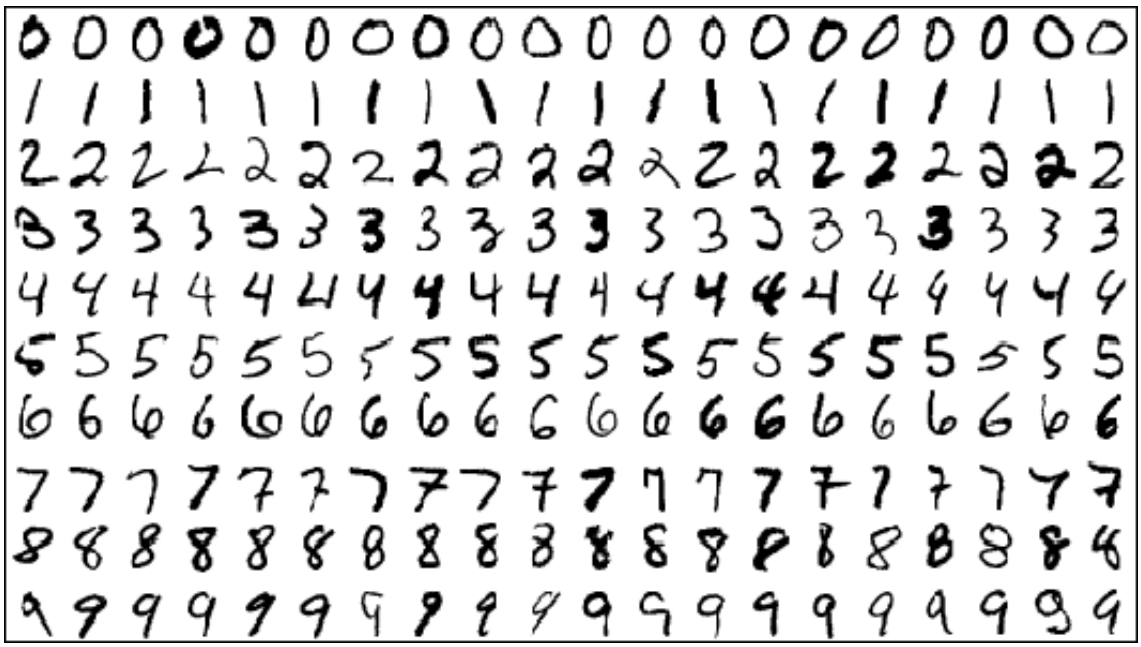
JULIEN JACQUES UNIVERSITÉ DE LYON

Fouille de données - Classification supervisée

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. Commencer par extraire aléatoirement un échantillon de 5000 images, qui serviront d'apprentissage, et un autre échantillon de 1000 images, qui serviront de test.
4. Réaliser une classification à l'aide des méthodes suivantes, et comparer leur qualité de prédiction sur l'échantillon test :
   (a) knn
   (b) arbre de décision (méthode CART)
   (c) forêt aléatoire
5. Tester différentes valeurs pour les hyper-paramètres des méthodes (nombre de plus proches voisins, nombre d'arbres de la forêt, nombre de variables choisies pour être candidates à chaque noeud).
6. Sélectionner la meilleure des méthodes, et essayer d'utiliser les 60000 données du jeux `train` pour prédire les 10000 du jeux `test`. Quelle est l'erreur de prédiction obtenue ? Comparez-vous aux résultats disponibles sur `http://yann.lecun.com/exdb/mnist/`.