Exercice: Reconnaissance de chiffres manuscrits

Base de données MNIST

Vous pouvez utiliser « from sklearn.datasets import fetch_openml» pour importer la base MNIST.

La base MNIST comporte des images de chiffres manuscrits en format 28 × 28 = 784 pixels, et les étiquettes associées. La base contient 70 000 exemples (7000 exemples de chaque chiffre).

Figure 1 montre un exemple de ces images.

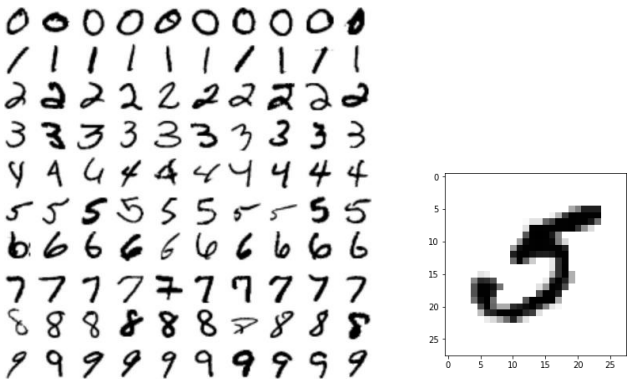

Figure 1 – Un exemple de 5 de la base MNIST.

Travail expérimental

Considérer uniquement les 10 000 premières images.

Visualiser quelques exemples de la base.

Indiquer les étiquettes qui existent dans la base.

Combien d'images y-a-t-il pour chaque étiquette ?

Partie 1: considérer les images de chiffres indépendamment de leurs labels et réaliser des regroupements de ces chiffres

Pour cela :

1. Appliquer la méthode de clustering ***K*-moyennes**.

- Discuter le nombre de classes et visualiser le centre des classes.
- Rappeler la fonction de coût pour l'algorithme *K*-Moyennes et visualiser son comportement en fonction de *k*
- Evaluer le nombre optimal de clusters k, en utilisant la « silhouette » comme métrique (indice de validité).
- Compléter par l'utilisation d'autres indices de validité : Calinski-Harabasz, davies_bouldin, entre autres. Préciser leur fonctionnement.
  Qu'en déduisez-vous ?

Partie 2: en prenant en compte les labels des classes

- Comment peut-on évaluer la qualité des clusters en se servant des étiquettes des classes ?
- Etudier et discuter les résultats associés.

Faire l'étude de la Partie 1 et Partie 2 pour différentes valeurs de *k* : 8, 10,12. Discuter.

Partie 3: Appliquer la méthode ***K*-medoids** et comparer avec la méthode des *K*-moyennes, en particulier en terme de temps de traitement, les centres obtenus avec *K*-medoids.