ÉCOLE NATIONALE SUPÉRIEURE DE TECHNIQUES AVANCÉES

ROB311 : Apprentissage pour la robotique

Parcours robotique

Année universitaire : 2020/2021

TP6 : K-means

Auteurs :

M. AHMED YASSINE HAMMAMI

MME. HANIN HAMDI

Objectif

L'objectif de ce TP est d'implémenter l'algorithme K-means pour faire le *clustering* du dataset : *digital handwriting.*

Les données utilisées

Les données traités sont des images numériques pré-traitées qui représentent différentes façons de digitaux écrits à la main.

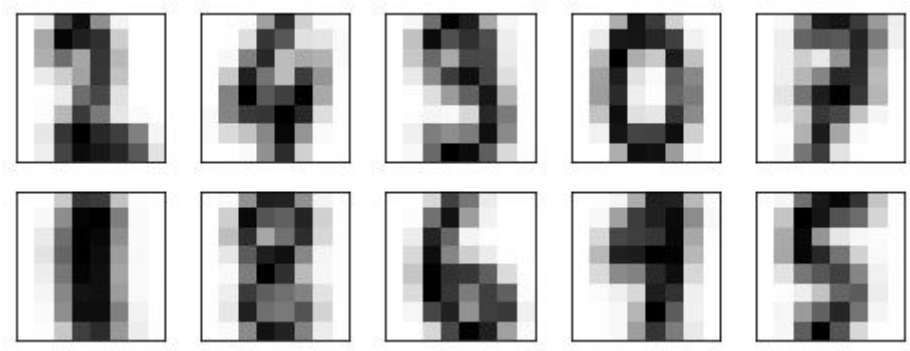

FIGURE 1 – Données traitées

Choix du nombre des clusters K

Pour ce dataset, on souhaite regrouper les images (codées en valeurs numériques) de 0 à 9. Donc, la valeur de K la plus logique pour ce dataset est 10.

Standardisation des données

Étant donné que le PCA produit un sous-espace de caractéristiques (dans notre cas les pixels de chaque image) qui maximise la variance le long des axes, il est logique de normaliser les données, en particulier si elles ont été mesurées à différentes échelles. Ainsi, la standardisation (ou la normalisation) consiste à redimensionner les caractéristiques de sorte qu'elles aient les propriétés d'une distribution normale standard avec une espérance nulle et un écart-type unitaire.

Réduction de la dimension par PCA

La taille des données du test et de l'apprentissage est très grande ce qui ne rend l'apprentissage très lent voire dépassant les capacités matérielles de nos machines utilisées pour faire l'apprentissage. Ainsi, on utilise le PCA pour réduire les dimensions des données et donc pour augmenter le speed-up de l'apprentissage.

Algorithme des k-means

L'algorithme implémenté est le suivant :

- Initialisation aléatoire des 10 centroïde.
- Pour chaque instance du dataset, déterminer le cluster le plus proche en s'appuyant sur la distance euclidienne entre cette instance et chaque centroïde.
- Répéter jusqu'à convergence :
  — Mettre à jour la liste des centroïdes : Pour chaque clustre le nouveau centre est la moyenne de toutes les instances dans le cluster.
  — Mettre à jour les clusters à partir de la nouvelle liste de centroïdes.
  — On considère qu'on a converge si la liste des centroïdes demeure inchangée.

Résultats

Pour ce type d'apprentissage on a utilisé la bibliothèque prédéfinie sur Python K-means pour réaliser l'apprentissage et la prédiction des données. Afin, de vérifier l'utilité de cet algorithme, on calcule son accuracy. On obtient le résultat suivant :

```
K-MEANS ALGORITHM

ACCURACY =  0.8074306645735218
```

FIGURE 2 – Accuracy de k-means

Pour comprendre encore plus les endroits où la prédiction est mauvaise ou bonne, on trace aussi la matrice de confusion :

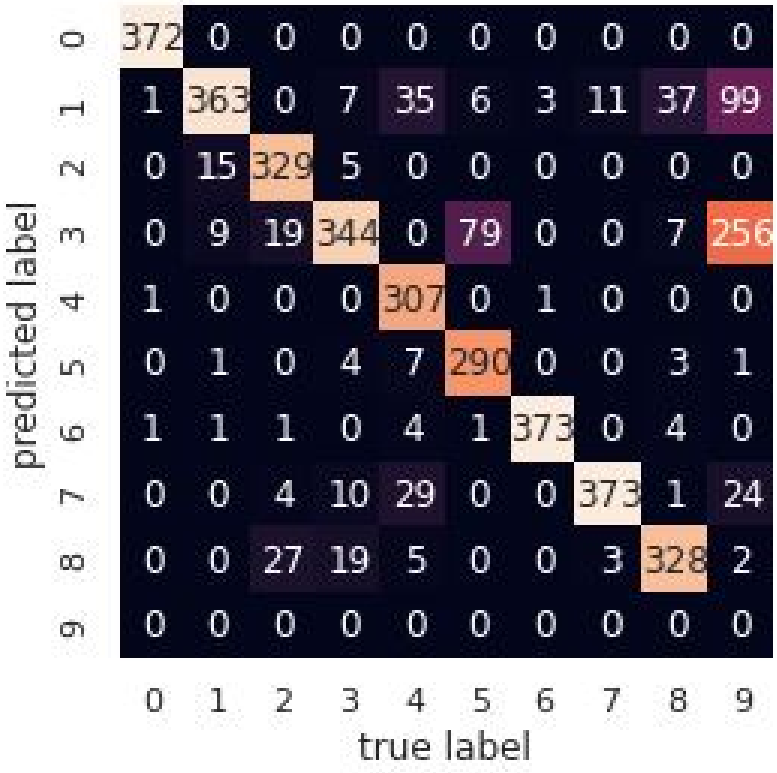

FIGURE 3 – Matrice de confusion de k-means

On obtient aussi la figure suivante qui résume les différents clusters :

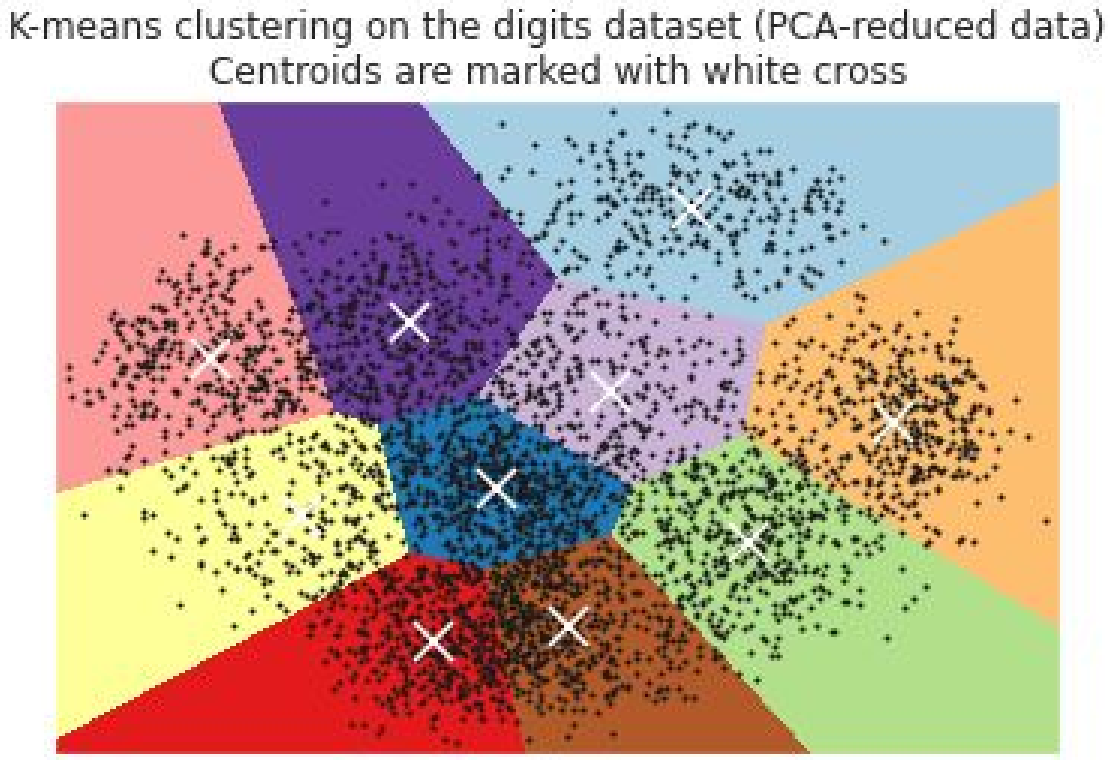

FIGURE 4 – Plot du résultat du clustering