IFT3700 : Science des données

Travail 1

Révisé le 7 octobre 2019

Description

Proposer une notion de similarité originale et spécifiquement construite pour être utilisée avec **MNIST**.
L'objectif est d'augmenter la performance de divers **algorithmes de partitions**.
Comparer la performance des algorithmes suivants en utilisant la distance euclidienne et votre mesure proposé.

***k*-medoïde**	**Partition binaire** (Regroupement hiérarchique)	**PCoA** (c'est un cas particulier de MDS)	**Isomap**	**KNN**

Conseils et indications

- L'utilisation de la librairie scikit-learn est recommandée.
- Si on effectue une légère translation de l'image, cela ne devrait pas affecter sa similarité.
- Il est permis de faire un prétraitement des données pour accélérer le calcul de la similarité.
- La notion de similarité n'a pas besoin d'être une distance, mais elle doit se comporter de façon similaire.
- Il est parfois nécessaire dans la phase exploratoire (ou même finale) de travailler avec des jeux de donnée de taille réduite.
- Pour l'algorithme **k-moyenne** vous devez utiliser une version de l'algorithme où le
  - **centroïde** est **l'élément** du **groupe** qui **maximise** la **similarité**
  - un **élément** est **dans** le **groupe** qui **maximise** sa **similarité** avec le **centroïde**.
  - ✓ **Cette version sera/fut présentée au TP**
- Dans le cas de **Partition binaire, utilisez** la
  - **variation** basée sur la **moyenne** des **distances**.

Critère d'évaluation

- Format **Jupyter Notebook**
- En groupe de 3 ou 4
- Originalité
- La qualité des résultats obtenus avec votre mesure de similarité
- La perte de performance si on compare avec la distance euclidienne (ou le gain si c'est le cas).
- Le rapport doit mettre en lumière de façon claire et honnête les forces et faiblesses de la similarité proposée.
- Le rapport peut être rendu en français ou en anglais.