Travail 1

IFT3700 - Science des données - A2022

Proposer deux notions de similarité originales et spécifiquement construites pour être utilisée avec **MNIST** et **ADULT**. L'objectif est d'augmenter la performance de divers algorithmes de partitions et classifications. Étudier la performance des algorithmes suivants avec les deux similarités sur les deux jeux de données.

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

Pour MNIST comparer la performance de votre similarité avec celle obtenue en utilisant la distance euclidienne.

Conseils et indications

- Le travail doit être réalisé et remis en groupe de 2 ou 3.
- Le travail doit être produit dans le format PDF et .py.
- L'utilisation de la librairie scikit-learn est recommandée.
- Pour MNIST une légère translation de l'image ne devrait pas affecter sa similarité.
- Pour Adult, la distance euclidienne ne peut être appliquée que sur les variables continues, cependant votre mesure de similarité peut aussi considérer les variables discrètes.
- 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 directement reliée à une distance, mais elle doit se comporter de façon raisonnable.
- Il est parfois nécessaire dans la phase exploratoire (ou même finale) de travailler avec des jeux de données de taille réduite.
- Dans le cas de **Partition binaire,** utilisez la variation basée sur la moyenne des distances.

Critères d’évaluation

- Originalité de l’approche
- Clarté et rigueur du rapport
- La qualité des résultats obtenus avec votre mesure de similarité
- Une bonne performance en termes de temps de calculs peut être un avantage de votre technique mais si cela est un désavantage cela ne signifie pas que votre technique ne peut pas être intéressante à d'autres niveaux. Maximiser l’exactitude est l’objectif usuel. Cependant, il est aussi possible de vouloir minimiser le temps de calcul et/ou la quantité de mémoire utilisée par le modèle. Si c’est votre cas, analysez le compromis entre la perte d’exactitude et le gain en temps/mémoire par rapport à l’algorithme standard.
- Le rapport doit mettre en lumière de façon claire et honnête les **forces et faiblesses** de la similarité proposée.