Université de Bordeaux Année 2025-2026
Master 2 Bio-Informatique

Visualisation d'Information
Projet – Analyse critique d'algorithme de réduction de dimension –

Objectif : L'objectif de ce projet est de comprendre, expliquer et proposer une analyse critique d'un algorithme de réduction de dimension parmi ceux choisis et d'en produire une analyse quantitative et éventuellement qualitative.

Pour cela, vous trouverez plusieurs attendus ci-dessous. Le résultat sera remis sous la forme d'un **rapport** (pdf ou notebook Jupiter) décrivant le travail que vous avez fourni et du **code** de l'analyse quantitative que vous aurez faite de l'algorithme de réduction que vous étudierez.

Modalités : Projet à réaliser en groupe de 4. Les projets seront évalués sur la base du code remis et du rapport.

Date limite de remise du code et du rapport : 2 octobre, 23h59.

Liste des algorithmes de réduction de dimension : t-SNE [1] et UMAP[2]

Langages et technologies : Aucune contrainte, vous devrez toutefois fournir un fichier requirements.txt contenant toutes les dépendances de votre projet.

Attendus :
Dans ce projet, il vous est demandé de :

- [Lire/]Comprendre l'algorithme de réduction et l'**expliquer** dans votre rapport. Il est notamment important d'expliquer l'impact des différents/éventuels paramètres.

- Faire une **analyse quantitative** des résultats de l'algorithme : dans ce contexte, il est important de pouvoir quantifier la déformation induite par la réduction de dimension. Pour simplifier l'étude, vous vous focaliserez sur un type de distribution de données en *clusters* et une taille fixée qui semble vraisemblable dans vos métiers cibles. Il vous est demandé de mettre en place et d'exécuter un banc d'essai (*benchmark*) respectant les recommandations suivantes :

    - Générer des jeux de données en faisant varier le nombre de dimensions et le nombre de clusters.

    - Faire varier un ou plusieurs paramètres de l'algorithme choisi,

- Identifier et Implémenter des mesures permettant de quantifier cette déformation,
- Calculer les valeurs de ces mesures sur les différents jeux de données et pour différentes valeurs de paramètres.
- de proposer une (des) visualisation(s) permettant de mettre en évidence cette déformation

Dans l'analyse des résultats que vous fournirez, les données devront être agrégées par nombre de dimensions des données, nombre de clusters des données, valeurs de paramètre(s) de l'algorithme, *etc.*.

- Certaines mesures de déformation sont calculées en agrégeant les déformations sur chaque point des données. Proposez une visualisation permettant d'observer si cette déformation est du même ordre pour tous les points d'un jeu de données.

- [optionnel] Faire une **analyse qualitative** des résultats de l'algorithme en répondant à ce type de questions : à partir de données que vous connaissez, quelles informations pouvez-vous retrouver ? Quelle nouvelle information (vraisemblable) pouvez-vous en tirer ? Au contraire, quelle information vous paraît aberrante ?

- Les résultats de ces analyses devront être décrits dans votre rapport.

Remarque importante : Pour ce projet, vous devrez générer un grand nombre de jeux de données (car vous devrez faire varier la dimensionalité des jeux de données, le nombre de clusters et répéter la génération un certain nombre de fois pour une même configuration). Le temps d'exécution du banc d'essai ne doit donc pas être sous-estimé (possiblement jusqu'à plusieurs jours).

Références :

[1] van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data using t-SNE. Journal of Machine Learning Research 9:2579-2605, 2008.

[2] McInnes, L, Healy, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018