Cours PCD – Labo 2 : Détection de données aberrantes (*outliers*)

Résumé du travail demandé

- Trouver les données atypiques (ou aberrantes) dans plusieurs jeux de données en appliquant la méthode LOF.
    - Ces jeux de données, ALOI, Glass et Ionosphere, sont disponibles en ligne et comportent une annotation des données atypiques.
    - Cette annotation vous permettra d'évaluer la méthode LOF en calculant son score F1, avec divers paramètres, sur chaque jeu.
- Indiquer les meilleurs paramètres que vous avez trouvés pour LOF sur chaque jeu de données (après plusieurs essais) et indiquer les scores F1 correspondants.
- Présenter la méthode et les résultats comme un court rapport dans un notebook Jupyter.

Indications

- Téléchargez les trois jeux de données depuis https://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/, section '*Datasets used in the literature*' ('ALOI', 'Glass' et 'Ionosphere'). Lorsqu'il y a le choix, prenez les versions normalisées et sans les doublons (*duplicates*).
- Commencez vos expériences avec le jeu de données le plus petit. Dans la mesure du possible, écrivez vos traitements sous forme de fonctions, pour pouvoir ensuite les appliquer facilement aux deux autres jeux de données.
- Cherchez sur Internet comment lire des données au format ARFF dans une *dataframe* Pandas.
- Pour chaque jeu de données, veuillez écrire dans une cellule de texte ce que les données représentent. Veuillez calculer le nombre d'items et le nombre *d'outliers* et comparez ces nombres avec les indications fournies sur la page DAMI.
- N'oubliez pas d'enlever l'attribut '*outlier*' avant de fournir les données à l'algorithme LOF. Comparez comment les *outliers* sont codés dans les données et dans l'algorithme LOF.
- À https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html vous trouverez des informations sur la méthode LOF. Elle a deux paramètres : le nombre de voisins (*n_neighbors*) et la proportion *d'outliers* attendue (*contamination*). Vous pouvez trouver leurs valeurs optimales par une recherche exhaustive sur l'intégralité de chaque jeu de données, ou par une recherche manuelle.
- Veuillez indiquer en conclusion les valeurs des paramètres qui maximisent les scores F1 sur chaque jeu de données, ainsi que les scores de rappel, précision et F1 obtenus avec ces paramètres.