RER
CLUSTERING

Contexte : Découvrir l'apprentissage non supervisé au travers les algorithmes **k-means**, **k-medoids** et **DBSCAN**. Analyser avec les métriques d'évaluations.

Problématiques :

1. Comment utiliser clustering dans l'apprentissage non supervisés ?
2. Comment construire des modèles de clustering ?
3. Comment utiliser la technique de la standardisation ?
4. Comment utiliser les métriques d'évaluations ?
5. Comment visualiser le clustering ?
6. Comment mesurer la similarité ?
7. Comment avoir un nombre de clusters optimal ?

Mots clés :

Apprentissage non-supervisé : la pratique consistant à donner à l'algorithme uniquement des données d'entrée sans une donnée cible, classes prédéfinis **(target**) afin qu'il détermine lui-même une cohérence structurelle dans celles-ci.

Clustering : méthode d'apprentissage non-supervisé qui consiste à regrouper les données similaires (se ressemblent). Il y a trois grandes méthodes de clustering : 1. méthode à partitionnement (**k-means, k-medoids**); 2. méthode hiérarchique (agglomérative (AGNES); divisive (DIANA) ; 3. Méthodes à densité (**DBSCAN**).

K-means : algorithme de clustering qui se base sur le barycentre (centre de gravité, barycentre n'est pas un point des données) pour créer des clusters. On doit choisir initialement k (nombre de clusters). Il fait des clusters de forme convexe.

K-medoids : comme k-means sauf qu'ici le centre de chaque cluster est nécessairement est un point du data set des données. Le calcul du centre se repose ici sur la médiane.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) : algorithme de clustering basé sur la densité des points pour former des clusters. Le cluster est un ensemble des points connectées. Sa forme n'est pas forcément convexe. Il y a deux hyperparamètres : 1. Epsilon est la distance pour déterminer pour chaque observation l'epsilon-voisinage (le rayon du cercle) ; 2. le nb minimal de voisins nécessaire pour considérer qu'une observation est un centre.

Similarité : est une mesure de similitude (distance) entre deux points de données. Plus deux points sont similaires - plus la distance entre eux est petite.

Distances :

1. **Euclidienne :** plus court chemin entre deux points et il est unique.

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{n}(x_{1n} - x_{2n})^2}$$

2. **Manhattan :** est la distance entre deux points parcourus selon un réseau ou un **quadrillage**. On se déplace d'un nœud du réseau à un autre en utilisant les **déplacements horizontaux** et **verticaux** du réseau. On peut trouver les différentes distances entre deux points (n'est pas unique).

Algorithme de réduction de dimension : bibliothèque pour afficher une image de 3D en 2D

Standardisation : est une méthode de normalisation (Z-score) de mise en échelle des écarts entre les enregistrements, elle s'applique sur les données du type intervalle et non sur les types nominaux. Pour les types ordinales – il faut les transformer en numérique d'abord.

Ground Truth (Vérité Terrain) : un ensemble de données dans lequel chaque point de données est déjà étiqueté avec un groupe ou une classe prédéfinie. Ces données étiquetées sont ensuite utilisées pour évaluer la qualité des groupes produits par un algorithme de clustering.

Rand Index : est une métrique d'évaluation de clustering (Ground Truth) - mesure la similarité entre deux partitions (entre deux clusterings), Il mesure à quel point deux clusters sont en accord en comparant toutes les paires de points et en mesurant la proportion de paires qui sont regroupées dans le même cluster dans le clustering produit et dans le "Ground Truth". **On calcul la proportion de deux points correctement répartis.**

$$\mathrm{RI} = \frac{a + b}{C_2^{n_{samples}}}$$

Adjusted Rand Index (!! Rand – n'est pas un random): permet de normaliser Rand Index (en tenant compte le hasard)

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]}$$

AMI (Adjusted Mutual Information) : Comme le Rand Index, elle se focalise sur la variation entre les points. (Basé sur l'entropie des points).

Inertie : mesure la somme des distances au carré entre chaque point de données et le centre de son cluster. Plus précisément, l'inertie est calculée comme la somme des variances intra-cluster,

c'est-à-dire la somme des distances euclidiennes au carré entre chaque point de données et le centre de son cluster, pour tous les clusters.

Silhouette : Est un coefficient qui mesure la cohésion des clusters. Le score de silhouette pour un point de données est calculé comme suit : (distance du cluster le plus proche - distance intra-cluster) divisé par le maximum de ces deux valeurs.

Analyse de silhouette :

$$SilhouetteAnalysis = \frac{b^i - a^i}{max(b^i, a^i)}$$

Ceci est utilisé pour déterminer le degré de séparation entre les clusters. Pour chaque échantillon. **a_i** représente la distance moyenne de tous les points de données d'un même cluster. **b_i** représente la distance moyenne de tous les points de données du cluster le plus proche.

Le coefficient de SA peut prendre des valeurs dans l'intervalle [-1, 1].

- SA = 0 : l'échantillon est très proche des grappes voisines.
- SA = 1 : l'échantillon est éloigné des grappes voisines.
- SA = -1 : l'échantillon est affecté aux mauvais clusters.

Maximal Distance to Cluster Center : distance maximale entre chaque point et le centroïde du cluster, le nombre élevé indique grande dispersion des points à l'intérieur dans le cluster. N'est pas utilisé seul, il est utilisé avec Avarage Distance to Cluster Center

Average Distance to Other Center : indique la distance, en moyenne, qui sépare chaque point dans le cluster des centroïdes de tous les autres clusters.

Average Distance to Cluster Center : indique la distance, en moyenne, qui sépare chaque point dans le cluster du centroïde de ce cluster.

Number of Points : nombre de points affectés au cluster.

Méthode Elbow : est une technique utilisée pour déterminer le nombre optimal de clusters dans un ensemble de données pour un algorithme de clustering donné. Elle utilise inertie pour déterminer k (nombre de clusters).

Hypothèses :

1. L'utilisation de l'algorithme k-means est plus simple que l'utilisations de k-medoids(Osman) **Faux**
2. La standardisation s'apparente à la normalisation (Jean Paul) **Vrai**
3. On est obligé d'utiliser Rand index dans clustering (Tetyana) **Faux**
4. Le clustering implique d'utiliser les mesures de dispersions (Étienne) **Faux**
5. La distance Euclidienne est la mesure la plus utilisée dans clustering (Adeline) **Faux**
6. Optimiser le nombre de cluster permet d'éviter l'over-fiting (Briand) **Faux**
7. L'entropie est utilisée pour la mise en place des clusters (Seydou) **Faux**
8. Plus l'entropie est faible plus le l'algorithme de clustering est efficace (Axel) **Faux**
9. Le clustering non-supervisé est l'unique méthode dans la classification non-supervisé (Loïc) **Faux**
10. On ne peut pas utiliser le clustering pour des données avec target (Nicolas) **Faux**
11. L'inertie est liée à Average Distance to Cluster Center (Aude) **Vrai**
12. La silhouette est une forme de dispersion des données (Solenn) **Vrai**
13. On utilise les métriques de classification pour faire du clustering (Adrien) **Faux**

Plan d'action :

- Explorer les ressources
- Définir et comprendre les mots clés
- Répondre aux problématiques
- Vérifier les hypothèses
- Comparaison entre **k-means**, **k-medoids** et **DBSCAN**.

K- means	K-medoids	DBSCAN
- Clustering par mesure de distance - Utilise les barycentres - Utilise les hyperparamètres k (nombre de barycentres) - Sensible aux données extrêmes	- Clustering par mesure de distance - Utilise un centroïde - Utilise les hyperparamètres k (nombre de centroïdes) - Sensible aux données non sectorisées	- Clustering par mesure de densité - Utilise les hyperparamètres eps et min_samples
-	-	

Caractéristiques	K-means	K-medoids	DBSCAN
Type de clustering	Partitionnement	Partitionnement	Densité
Sélection du nombre de clusters	Pré-défini	Pré-défini	Automatique
Centres de cluster	Moyenne des points	Point le plus représentatif	Non applicable
Sensible aux outliers	Oui	Oui	Non
Forme des clusters	Sphérique	Non-sphérique	N'importe quelle forme
Efficacité sur des données de grande dimension	Efficace	Efficace	Peu efficace
Robuste aux données bruitées	Non	Non	Oui
Identification de clusters de densité différente	Non	Non	Oui
Détection de clusters de petite taille	Non	Oui	Oui
Convient aux données non-linéaires	Non	Non	Oui

- Faire les Workshops
- Rendu du RER