A.U 2025/2026
Semestre 3
Master : IIBDCC

Atelier 2:

Mesures de similarité et dis-similarité sur

MallcCustemers dataset

Cet atelier vous permet de comprendre les mesures de similarité et de dissimilarité appliquées à des données quantitatives. L'objectif alors est de :

- Comprendre la notion de proximité entre observations dans un espace vectoriel.
- Mettre en œuvre différentes mesures de similarité et de distance (euclidienne, Manhattan, cosinus, corrélation).
- Visualiser les relations entre les vecteurs de données et interpréter les résultats.

Importation et Exploration des Données

1. Importer les bibliothèques nécessaires :
2. Charger le dataset.
3. Afficher les premières et les dernières lignes du jeu de données.
4. Sélectionner les variables quantitatives.
5. Changer le valeurs de la variable Annual Income (k$) par la variable Annual Income ($) (par exemple 13→ 1300)
6. Afficher les premières et les dernières lignes du jeu de données.

Mesures de Dissimilarité

Les mesures de dissimilarité permettent de quantifier la différence entre deux observations.

7. Distance euclidienne

   dist_euc = euclidean_distances(data)
8. Distance de Manhattan

   dist_man = manhattan_distances(data)
9. Normaliser les données

   from sklearn.preprocessing import MinMaxScaler

   scaler = MinMaxScaler()

   data_n = ps.DataFrame(scaler.fit_transform(data), columns=data.columns)
10. Recalculer les distances euclidiens et Manhattan sur les données normalisées. Conclure

Mesure de similarité

Les mesures de similarité mesurent la proximité ou ressemblance entre deux vecteurs.

11. Similarité cosinus

    Cos_sim = cosine_similarity(data)
12. Similarité de corrélation (Pearson)

    sim_corr = data.T.corr()

Comparaison des Résultats

11. Comparer les matrices obtenues via la distance euclidienne et le cosinus similarité.
12. Visualiser les relations sous forme de heatmaps côte à côte. Conclure ?

Questions pour Discussion

1. Quelle est la principale différence entre une distance et une similarité ?

A.U 2025/2026
Semestre 3
Master : IIBDCC

Atelier 2:

Mesures de similarité et dis-similarité sur

MallcCustemers dataset

2. Pourquoi la normalisation des données est-elle importante avant le calcul de distances ?
3. Quelle distance est la plus sensible aux valeurs extrêmes ?
4. Dans quel cas la similarité cosinus est-elle plus adaptée que la distance euclidienne ?