Clustering via Non-Negative Matrix Factorization (NNMF)

Projet réalisé par :
Fezoui Yacine - Paul-Arthur NGUYEN - Stephane WU

Introduction

- **Motivation :**

  - La méthode NNMF : pour réduire la dimension des données, facilitant ainsi l'identification de clusters.

  - La segmentation : pour identifier des groupes au sein des données et de mettre en évidence des tendances importantes.

- **À propos du problème :**

  - Identifier des groupes aux comportements similaires à partir du jeu de données.

State of the Art

- **Méthodes courantes pour la réduction de dimensionnalité :**
  - **NMF simple : Réduction de dimensions** - Utilise la factorisation de matrice pour réduire les dimensions des données tout en conservant leur structure.
  - **Sparse NMF : Factorisation avec parcimonie** - Contraint les matrices factorielles pour obtenir des représentations plus interprétables et adaptées.
  - **Multiview PCA : Analyse multi-vues** - Combine des données provenant de différentes sources tout en appliquant une analyse en composantes principales.
- **Méthodes courantes pour le clustering :**
  - **K-means** : Méthode qui permet d'identifier les clusters dans un ensemble de donnée en fonction de k.
  - **DBSCAN** : Clustering basé sur la densité, adapté aux formes complexes de clusters.
  - **Agglomerative clustering** : Méthode hiérarchique créant un arbre de décision pour le clustering.
  - **Gaussian Mixture Models (GMM)** : Modèle probabiliste où chaque cluster suit une distribution normale.

Méthodologie

- **Étapes pour résoudre le problème :**

1. **Prétraitement des données**
2. **Application de NNMF**
3. **Nombre de segments**
4. **Clustering (K-means)**
5. **Analyse des clusters**

Experiments : Data et Protocol

- **Dataset :** Mall Customer Dataset (Standard)

  - 200 clients, colonnes : Age, Revenu annuel, Spending Score.

- **Protocole :**

❖ Exclusion de la variable **Genre** (qualitative).

❖ **Imputation des valeurs manquantes** par la **moyenne**.

❖ **Traitement des outliers** avec la méthode **IQR**.

❖ **Normalisation des données** avec la formule :

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Les données après prétraitement :

Age	Annual Income (k$)	Spending Score (1-100)
0.500000	0.567568	0.193878
0.423077	0.648649	0.122449
0.961538	0.297297	0.479592
0.307692	0.792793	0.224490
0.326923	0.531532	0.724490

Experiments : Data et Protocol

2. **Application de NNMF simple :** Réduction des dimensions.

- **Décomposition NMF** : fonction **NMF** du module **decomposition** : bibliothèque **sklearn**.
- Resultat : **W** (matrice des caractéristiques latentes) et **H** (matrice des bases).
- La matrice **W** est exploitée comme entrée pour l'étape :

de clustering.

- But ?

0.8	0.3
0.8	0.2
0.1	1.1
0.2	1.2
0.7	0.9
0.8	0.8

Experiments : Data et Protocol

3. Deuxieme methode sparse :

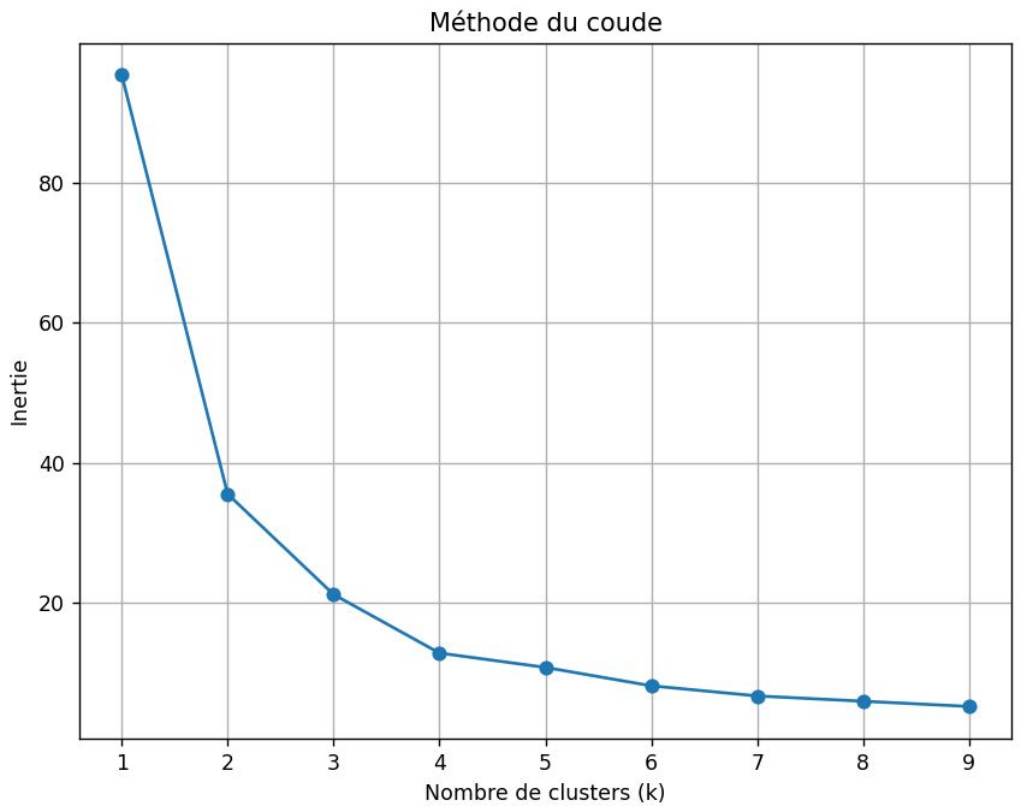
- La **Sparse NMF** est une variante de la NMF classique qui introduit une **contrainte de parcimonie** pour obtenir des matrices avec davantage de zéros.
- Réalisée avec **sklearn.decomposition.NMF**, ajoute le paramètre **l1_ratio=0.5** pour contrôler le niveau de parcimonie.
- Comparée à la NMF classique, elle produit des résultats similaires tout en forçant une représentation plus **sparse** (compacte).
- Cette méthode permet de mieux identifier les **composantes essentielles** en éliminant les contributions mineures des variables.

Experiments : Data et Protocol

4. **Nombre de segment :** Déterminer en combien de groupe on peut segmenter notre population

La méthode du coude :

- Mise en œuvre avec **KMeans** de `sklearn`.
- Le graphe montre un coude clair pour **k=3**, car **k=2** manque de granularité et **k=4** n'apporte pas une valeur ajoutée significative.
- Résultats confirmés par la méthode du **score de silhouette**, assurant la cohérence du choix.



Experiments : Data et Protocol

5. **Clustering (K-means) :** Regroupement des clients en clusters.

- Clustering effectué avec **KMeans** de `sklearn`.
- Nombre de clusters fixé à **3** à partir des analyses préalables (coude et silhouette).
- Utilisation de la matrice réduite **W** issue de **NMF** pour la classification.
- Les clusters sont attribués à chaque client dans le dataset initial.
- Résultat présenté avec : **CustomerID**, Genre, Âge, Revenu, Score de dépenses et Cluster :

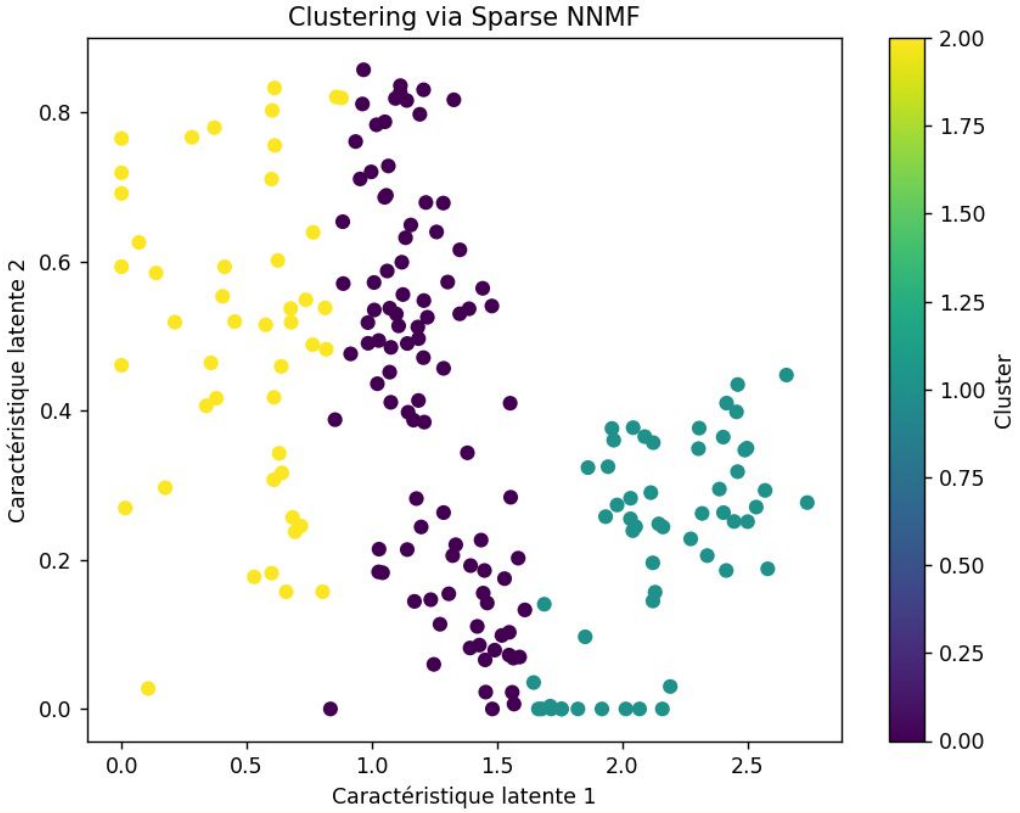
	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)	Cluster
0	1	Male	19	15	39	0
1	2	Male	21	15	81	1
2	3	Female	20	16	6	2
3	4	Female	23	16	77	1
4	5	Female	31	17	40	2

Resultats et Discussion

- **Analyse des clusters :**

Interprétation des groupes formés :

- Clusters formés basés sur l'âge, le revenu et le Score de Dépenses.



- Tableau des moyennes et des comptes par cluster :

Cluster	Age	Annual Income (k$)	Spending Score (1-100)	Nombre d'éléments	male	female
0	41.112245	58.704082	47.489796	98	37.755102	62.244898
1	30.196429	68.017857	82.357143	56	42.857143	57.142857
2	44.565217	55.434783	16.826087	46	58.695652	41.304348

Resultats et Discussion

- Discussion :

Cluster 0 (48.5%) :

- **Caractéristiques** : - **Classe moyenne**
  - Attentifs à leurs dépenses
  - Ils peuvent se permettre des achats plus élevés **occasionnellement**, mais uniquement si cela en vaut vraiment la peine ou si c'est un besoin essentiel.

- **Stratégie marketing** : Mettre en avant des **produits de qualité avec un bon rapport qualité/prix**, des **offres limitées** et des **promotions exceptionnelles** pour les inciter à faire des achats occasionnels plus importants.

Resultats et Discussion

Cluster 1 (28.5%) :

- **Caractéristiques** : **Jeunes adultes**
  revenu élevé.
  - Se laisse facilement séduire par des **tentations** et **dépense sans trop réfléchir**.
  - Prêts à faire des achats impulsifs et privilégient les **produits tendance** et **marques populaires**.

- **Stratégie marketing** : - Mettre l'accent sur des **produits très promus sur les réseaux sociaux ou marques connues**.
  - **Plats préparés** et des solutions rapides
  - **Promotions limitées** et des **offres attractives**

Resultats et Discussion

Cluster 2 (23%) :

- **Caractéristiques : - Dépensent très peu**.
  - **Achats spécifiques**.
  - Articles associé au magasin.
  - **Promotions attractives, produits de niche**, ou des articles difficiles à trouver ailleurs.
  - Réalisent leurs achats dans **d'autres enseignes**

- **Stratégie marketing : - Élargir la gamme** pour **couvrir davantage de besoins** et renforcer **l'image d'un "magasin complet".**
  - Proposer des **promotions ciblées**.
  - Une analyse des **données d'achat**.
  - **Programmes de fidélité** récompensant les achats fréquents avec des **avantages** (réductions cumulatives, cadeaux, ou accès à des offres exclusives) peuvent inciter ces clients à revenir régulièrement et à élargir leur **panier d'achat.**

Conclusion

- Résumé :

  - NNMF est une méthode efficace pour la segmentation.

  - Les clusters obtenus offrent des insights utiles pour des stratégies marketing ciblées.

- Perspectives :

  - Tester avec d'autres datasets pour valider la robustesse.

  - Intégrer d'autres méthodes de clustering comme DBSCAN.

Bibliographie

- On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering
- Mall Customer Segmentation Data
- Sklearn Documentation
- K Means Documentation