Clustering

Introduction

Le **clustering**, ou analyse de regroupement, est une méthode d'**apprentissage non supervisé** utilisée pour découvrir la structure sous-jacente des données en les regroupant en **clusters** (groupes) homogènes.

Objectifs et Applications

Objectif principal

L'objectif principal du clustering est de **maximiser la similitude** entre les éléments au sein d'un même cluster tout en **minimisant la similitude** avec les éléments des autres clusters.

Applications

- *Text Mining* : Identifier des textes similaires pour le résumé automatique ou la détection de sujets.
- **Marketing** : Segmentation des clients pour des campagnes ciblées et la personnalisation de l'offre.
- **Analyse d'images** : Segmentation d'images pour isoler des zones homogènes ou détecter des objets.
- **Bio-informatique** : Identifier des groupes de gènes ou de protéines similaires pour comprendre leurs fonctions.
- **Web** : Regrouper des utilisateurs en fonction de leur comportement en ligne pour la recommandation de contenu.

Notion de Dissimilarité

La **dissimilarité** est une mesure de la différence entre deux échantillons ou clusters. Elle est essentielle pour :

- **Identifier les similarités** et différences entre les données.
- **Guider les algorithmes** dans la formation et la mise à jour des clusters.

Dissimilarité entre deux points

Plusieurs mesures de distance peuvent être utilisées :

- **Distance Euclidienne** : Mesure la distance directe (à vol d'oiseau) entre deux points dans un espace Euclidien.

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

- **Distance de Manhattan** : Somme des distances absolues sur chaque dimension, représentant le chemin parcouru dans une grille urbaine.

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

- **Distance de Mahalanobis** : Prend en compte la corrélation entre les variables et est utile pour pondérer certaines dimensions.

$$D(x_1, x_2) = \sqrt{(x_1 - x_2)^\top S^{-1} (x_1 - x_2)}$$

  où S est la matrice de covariance.

- **Distance personnalisée** : Utilise une matrice positive définie W pour pondérer les dimensions selon leur importance.

$$D^2(x_1, x_2) = (x_1 - x_2)^\top W (x_1 - x_2)$$

Dissimilarité entre clusters

Pour mesurer la dissimilarité entre clusters, plusieurs critères sont utilisés :

Distances entre clusters

Pour mesurer la dissimilarité entre deux clusters C_1 et C_2, différentes métriques peuvent être utilisées :

- **Distance minimale (Single Linkage)** : Distance entre les deux points les plus proches appartenant à des clusters différents.
$$D_{\min}(C_1, C_2) = \min\{D(x_i, x_j) \mid x_i \in C_1, x_j \in C_2\}$$
- **Distance maximale (Complete Linkage)** : Distance entre les deux points les plus éloignés de chaque cluster.
$$D_{\max}(C_1, C_2) = \max\{D(x_i, x_j) \mid x_i \in C_1, x_j \in C_2\}$$
- **Distance moyenne (Average Linkage)** : Moyenne des distances entre tous les points des deux clusters.
$$D_{\text{moy}}(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{x_i \in C_1} \sum_{x_j \in C_2} D(x_i, x_j)$$
- **Distance des centroïdes (Centroid Distance)** : Distance entre les centroïdes des deux clusters.
$$D_{\text{centroid}}(C_1, C_2) = D(\mu_1, \mu_2)$$
où μ_1 et μ_2 sont les centroïdes des clusters C_1 et C_2.
- **Distance de Ward** : Basée sur la minimisation de la variance intra-cluster après fusion des deux clusters.
$$D_{\text{Ward}}(C_1, C_2) = \sqrt{\frac{|C_1| \cdot |C_2|}{|C_1| + |C_2|}} \cdot D(\mu_1, \mu_2)$$

Évaluation de la Qualité des Clusters

Pour évaluer la pertinence d'un clustering, on utilise des critères quantitatifs :

Variance intra-cluster

Évalue la compacité des points à l'intérieur de chaque cluster (on cherche à la minimiser).

$$J_w = \sum_{\ell} \sum_{i \in C_\ell} D^2(x_i, \mu_\ell)$$

par exemple avec la norme 2:

$$J_w = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

où μ_k est le centroïde du cluster C_k.

Variance inter-cluster

Mesure la séparation entre les clusters (on cherche à la maximiser).

$$J_b = \sum_{k=1}^{K} N_k \|\mu_k - \mu\|^2$$

où N_k est le nombre de points dans le cluster C_k et μ est la moyenne globale des données.

Critère de bon clustering

Un bon clustering minimise la variance intra-cluster (J_w) tout en maximisant la variance inter-cluster (J_b).

Méthodes de Clustering

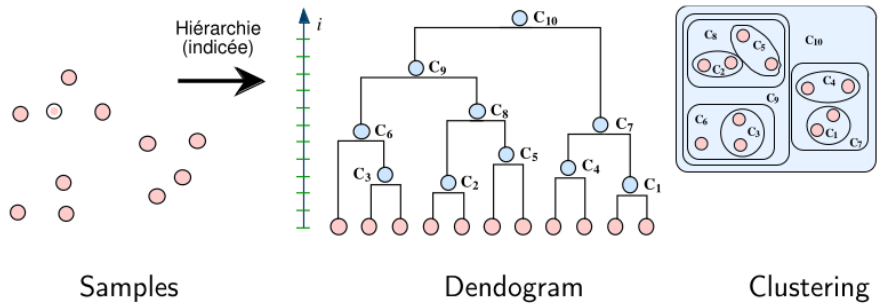
Clustering Hiérarchique

Principe : Construire une hiérarchie de clusters en regroupant successivement les données.

Algo :

1. Chaque échantillon est initialement considéré comme un cluster singleton.
2. À chaque étape, fusionner les deux clusters les plus proches selon un critère de liaison.
3. Répéter jusqu'à ce que tous les échantillons soient regroupés en un seul cluster.

Représentation : Le résultat est souvent visualisé sous la forme d'un **dendrogramme**



Avantages :

- Ne nécessite pas de spécifier le nombre de clusters à l'avance.
- Permet d'explorer différentes granularités de clustering en coupant le dendrogramme à différents niveaux.

Limites :

- Complexité computationnelle élevée pour de grands jeux de données.
- Sensible au bruit et aux outliers.
- Le choix du critère de liaison influence fortement le résultat.

K-means

Principe : Partitionner les données en K clusters en minimisant la somme des distances au carré entre les points et le centroïde du cluster.

Algo :

1. **Initialisation** : Choisir K centres (*centroïdes*) initiaux, souvent de manière aléatoire.
2. **Assignation** : Attribuer chaque point au cluster avec le centroïde le plus proche.
3. **Mise à jour** : Recalculer les centroïdes en prenant la moyenne des points de chaque cluster.
4. **Convergence** : Répéter les étapes d'assignation et de mise à jour jusqu'à ce que les centroïdes ne changent plus significativement.

Avantages :

- Simple et rapide pour des jeux de données de taille modérée.
- Facile à implémenter.

Limites :

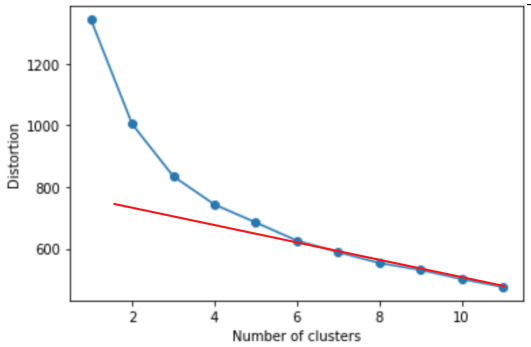
- Nécessite de spécifier K à l'avance.
- Sensible aux valeurs initiales des centroïdes (peut converger vers un minimum local).
- Ne fonctionne pas bien avec des clusters de formes non convexes ou de tailles très différentes.

Évaluation et Choix du Nombre de Clusters

Méthode du Coude (Elbow Method)

Consiste à tracer la courbe de la somme des distances intra-cluster J_w en fonction du nombre de clusters K. Le point où l'ajout d'un cluster supplémentaire n'améliore que marginalement J_w (forme de coude dans la courbe) est choisi comme le nombre optimal de clusters.

Hugo Tondenier



Indice de Silhouette

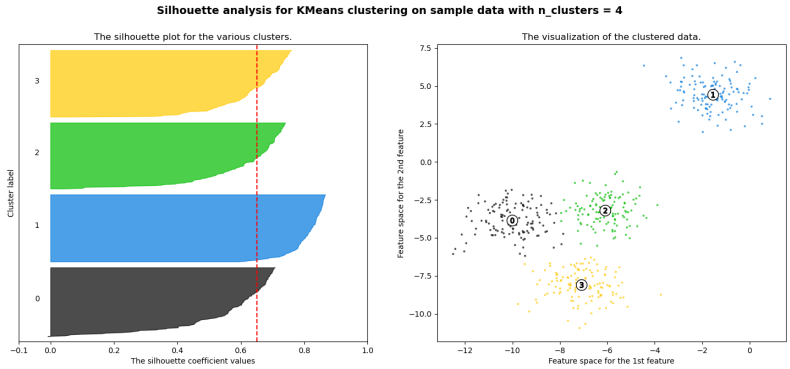
Mesure le degré de similarité d'un point avec son propre cluster par rapport aux autres clusters. Il varie de -1 à 1 :

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

où :

- $a(i)$ est la distance moyenne entre i et les autres points de son cluster.
- $b(i)$ est la distance moyenne entre i et les points du cluster le plus proche.

Un indice de silhouette moyen élevé indique un bon clustering.



Autres Méthodes de Clustering

Clustering par Densité (DBSCAN)

Principe : Identifie des clusters en se basant sur la densité locale des points, capable de détecter des formes arbitraires et de gérer le bruit.

Avantages :

- Pas besoin de spécifier le nombre de clusters à l'avance.
- Capable de détecter des clusters de formes arbitraires.
- Gère efficacement le bruit et les outliers.

Limites :

- Les performances dépendent du choix des paramètres de densité (ε et MinPts).
- Moins efficace dans les espaces de haute dimension.

Clustering Spectral

Principe : Utilise les valeurs propres du graphe de similarité des données pour effectuer le clustering. Particulièrement utile pour des données non linéairement séparables.