TD – Clustering

novembre 2024

1. **Mise en pratique**
   On considère un ensemble de points 1D :

   $$\mathcal{D} = \{-5, -3.5, -2.75, -0.5, 0, 0.2, 0.5, 2, 3, 5, 7\}$$

   On veut réaliser le clustering de ces points en utilisant diverses méthodes : le clustering hiérarchique ascendant (CHA), le K-means et DBSCAN.

   (a) Appliquez le CHA aux données, dessinez le dendogramme et montrez que le nombre raisonnable de clusters est 3. On utilisera comme métrique entre deux clusters

   $$d(C_i, C_j) = \min_{x \in C_i, z \in C_j} \|x - z\|$$

   puis la métrique :

   $$d(C_i, C_j) = \max_{x \in C_i, z \in C_j} \|x - z\|$$

   (b) Appliquez l'algorithme des K-means à partir des initialisations suivantes $\mu_1 = -1$, $\mu_2 = -0.25$, $\mu_3 = 1$.

   (c) Proposez d'abord un $NbPoints$ et un ε et explorez les points en ordre inverse (du plus grand au plus petit) pour l'exécution de DBSCAN.

2. **Indice de Silhouette**
   Pour un partage un clusters $C_1, C_2, \ldots, C_k$ on calcule pour chaque point x deux valeurs :

— $a(x)$, $x \in C_i$, la distance moyenne du point à son groupe :

$$a(x) = \frac{1}{|C_i| - 1} \sum_{y \in C_i, y \neq x} d(x, y)$$

— $b(x)$ la distance moyenne du point au cluster le plus proche :

$$b(x) = \min_{j, j \neq i} \frac{1}{|C_j|} \sum_{z \in C_j} d(x, y)$$

L'indice Silhouette d'un point :

$$s_{Sil}(x) = \frac{b(x) - a(x)}{\max\left(a(x), b(x)\right)}$$

L'indice Silhouette d'une solution de clustering est la moyenne de la moyenne des indices d'un cluster :

$$S(C_1, C_2, \ldots C_k) = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{x \in C_i} s_{Sil}(x)$$

Calculez l'indice de Silhouette pour une solution de clustering obtenue au point précédent.