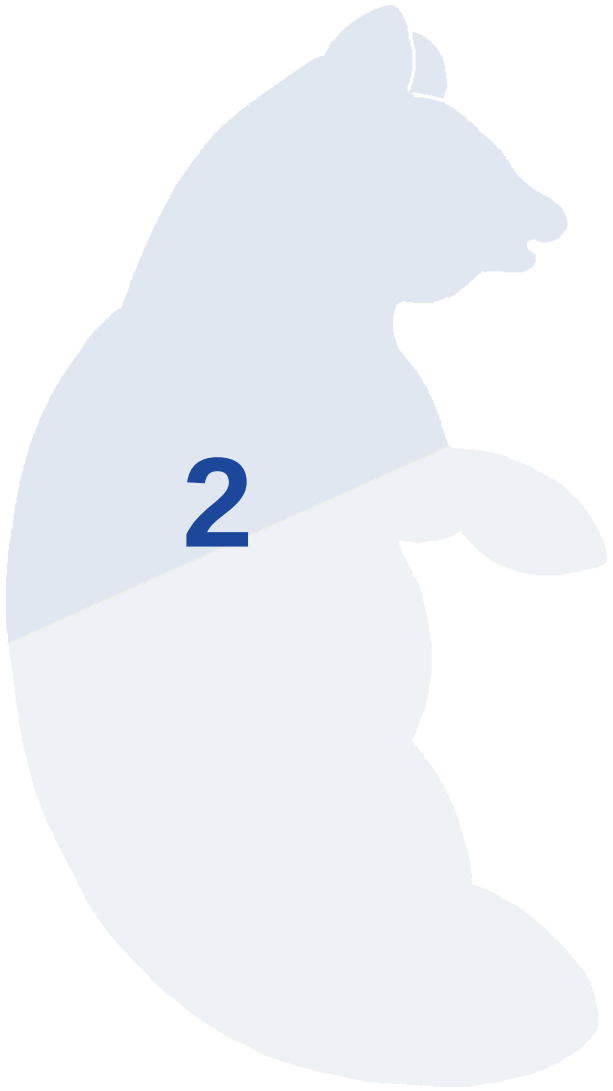

2

Custering Généralités

2. Clustering généralités

Clustering :

Les techniques de clustering cherchent à décomposer un ensemble d'individus en plusieurs sous ensembles les plus homogènes possibles sans connaître la classe des exemples (nombre, forme, taille...).

Méthodes de clustering :

- **Méthodes de partitionnement**
- **Méthodes hiérarchiques**
- **Méthodes de modélisation**
- **Méthodes à base de grille**
- **Recherche d'espaces latents**

2. Clustering généralités

Méthodes de partitionnement : construire une partition en classes d'un ensemble d'objets ou d'individus hiérarchiques.

Création d'une partition initiale de K classes, puis itération d'un processus qui optimise le partitionnement en déplaçant les objets d'une classe à l'autre.

- **méthodes K-moyennes** : une classe est représentée par son centre de gravité (ex : ***K-Means***).
- **méthode K-medoids** : une classe est représentée par son élément le plus représentatif (ex : algo ***PAM*** Partitioning around medoids, ***CLARA*** Clustring Large Applications).
- **partitionnement spectral** : principe = passer par graphe d'adjacence. Calculer matrice « Laplacienne » L=D-A du graphe adjacence. Trouver et trier valeurs propres de L (symétrique => valeurs propres réelles >=0, et vecteurs propres perpendiculaires. Projeter les points sur k vect propres de plus grandes valeurs propres, pour séparer plus facilement.

2. Clustering généralités

Méthodes hiérarchiques : les clusters sont organisés sous la forme d'une structure d'arbre (le nombre de clusters n'est pas défini) en utilisant les distances entre les classes.

- **approche hiérarchique ascendante RHA (ou agglomération)** : on commence avec un objet dans chaque classe, puis on fusionne successivement les 2 classes les plus proches et on s'arrête quand il n'y a plus qu'une classe contenant tous les exemples (recherche : ***BIRCH*** (1996) : basée sur la représentation d'une classe par ses traits caractéristiques et ***CHAMELEON*** (1999) : basée sur la théorie des graphes et une représentation plus riche des données).
- **approche descendante (ou par division)** : tous les objets sont initialement dans la même classe, puis division de la classe en sous classes jusqu'à ce qu'il y ait suffisamment de niveaux ou que les classes ne contiennent plus qu'un seul exemple.

2. Clustering généralités

Méthodes de modélisation :

- **à base de densité** : utilisation de la densité à la place de la distance. Les clusters sont des régions de l'espace qui ont une grande densité de points (ex : ***DBSCAN*** (1996), ***HDBSCAN***, ***OPTICS*** (1999), **DENCLUE** (1998), ***CLIQUE*** (1998)).
  - Un point est dense si le nombre de ses voisins dépasse un certain seuil
  - Un point est voisin d'un autre point s'il est à une distance inférieure à une valeur fixée.
- **mélange de gaussiennes** : clustering en utilisant les mixtures gaussiennes. La valeur de K correspond au nombre de composantes de la mixture (ex : algo ***EM*** Expectation Maximization, ***GMM***,).
  - Cartes de Kohonen (Self-Organizing Maps, SOM).

2. Clustering généralités

Méthodes à base de grille :

Méthode basée sur le découpage de l'espace des exemples suivant une grille.

Après initialisation, toutes les opérations de clustering sont réalisées sur les cellules, plutôt que sur les données. Construction des classes en assemblant les cellules voisines en terme de distance (ex : ***STING*** (a STatistical Information Grid approach (1997), ***Cluster*** (utilise la notion d'ondelettes) (1998)).

2. Clustering généralités

Recherche d'espaces latents :

Y'a t-il dans l'espace de description, des sous-espaces où la densité d'objets est plus importante que d'autres ?
Comment déterminer et caractériser ces régions ? (ex : ***ACP***).