METHODES DE CLUSTERING

Sébastien Quinault / IRIAF / 2021

DBSCAN CLUSTERING

(DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS)

Caractéristiques

- Permet d'identifier des clusters de n'importe quelle forme (pas uniquement sphériques)
- Utilise la densité de points (Density Based) pour définir des clusters
- Détecte les outliers
- Ne nécessite pas de définir un nombre de clusters

Les paramètres de l'algorithme

- **eps**
  La distance de 'voisinage'. Deux points sont voisins si la distance est inférieure ou égale à ce paramètre.

- **minPts**
  Le nombre de points pour constituer un cluster.

Les paramètres de l'algorithme bonnes pratiques

- **eps**
  Si la valeur eps est choisie trop petite, une grande partie des données sera considérée comme aberrante. Si elle est choisie très grande, les clusters fusionneront et la majorité des points de données seront dans les mêmes clusters. Une façon de trouver la valeur eps est de se baser sur le graphique de la k-distance.

- **minPts**
  En règle générale, la valeur minimale de MinPts peut être dérivée du nombre de dimensions D dans l'ensemble de données comme suit :
  MinPts = 2*D

  La valeur minimale de MinPts doit être choisie au moins égale à 4.

Le fonctionnement de l’algorithme

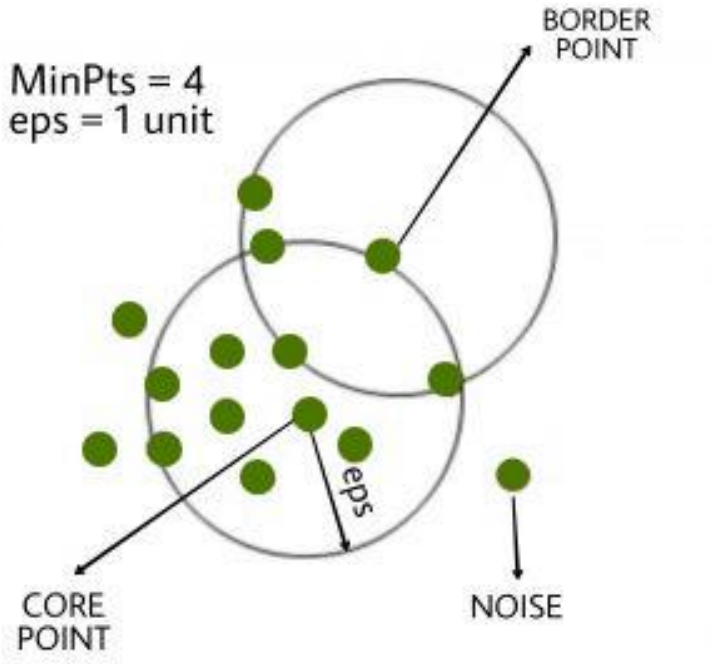
1. Un point de départ est choisi au hasard et son voisinage est déterminé en utilisant le rayon eps.
2. S'il y a au moins le nombre minPts de points dans le voisinage, le point est marqué comme point central et la formation d'un cluster commence. Sinon, le point est marqué comme bruit.

Le fonctionnement de l’algorithme

3. Une fois que la formation d'un cluster commence (disons cluster A), tous les points dans le voisinage du point initial deviennent une partie du cluster A.
4. L'étape suivante consiste à choisir aléatoirement un autre point parmi les points qui n'ont pas été visités lors des étapes précédentes. La même procédure s'applique alors.

Ce processus est terminé lorsque tous les points ont été visités.

Le fonctionnement de l'algorithme



Avantages

- Ne nécessite pas un nombre de clusters prédéfinis
- Forme des clusters de forme libre (pas uniquement sphérique)
- Permet d’identifier des outliers

Inconvénients

- Très sensible aux paramètres, donc difficile de les estimer
- La qualité dépend fortement de la mesure de distance

Code R

http://www.sthda.com/english/wiki/wiki.php?id_contents=7940

HIERARCHICAL CLUSTERING

Deux approches

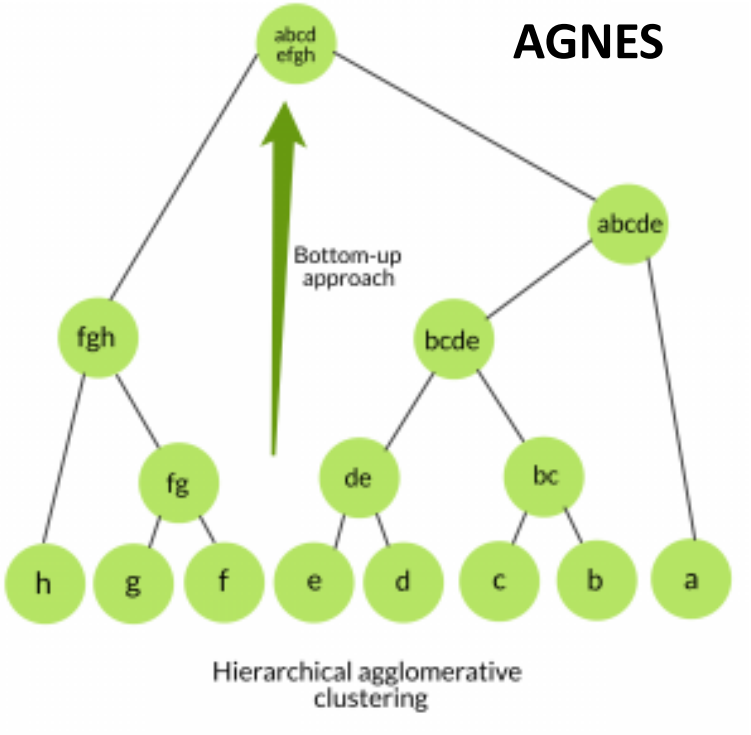
- Ascendante (AGNES)

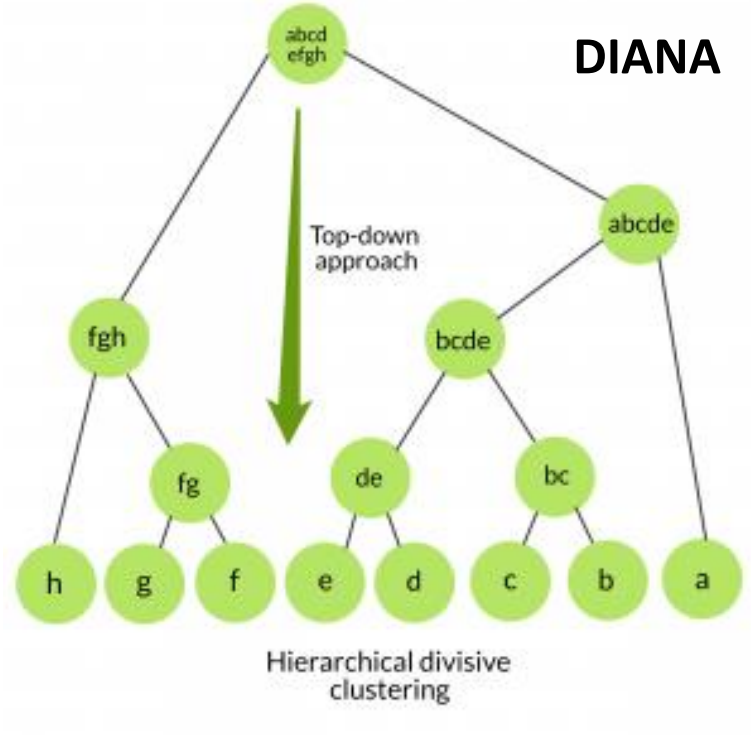
Chaque individu est initialement considéré comme un cluster à élément unique (feuille). À chaque étape de l'algorithme, les deux clusters les plus similaires sont combinés en un nouveau cluster plus grand (nœuds). Cette procédure est itérée jusqu'à ce que tous les points soient membres d'un seul grand cluster

- Descendante (DIANA)

L’algorithme commence par la racine, dans laquelle tous les objets sont inclus dans un seul cluster. A chaque étape de l'itération, le cluster le plus hétérogène est divisé en deux. Le processus est itéré jusqu'à ce que tous les objets soient dans leur propre cluster.

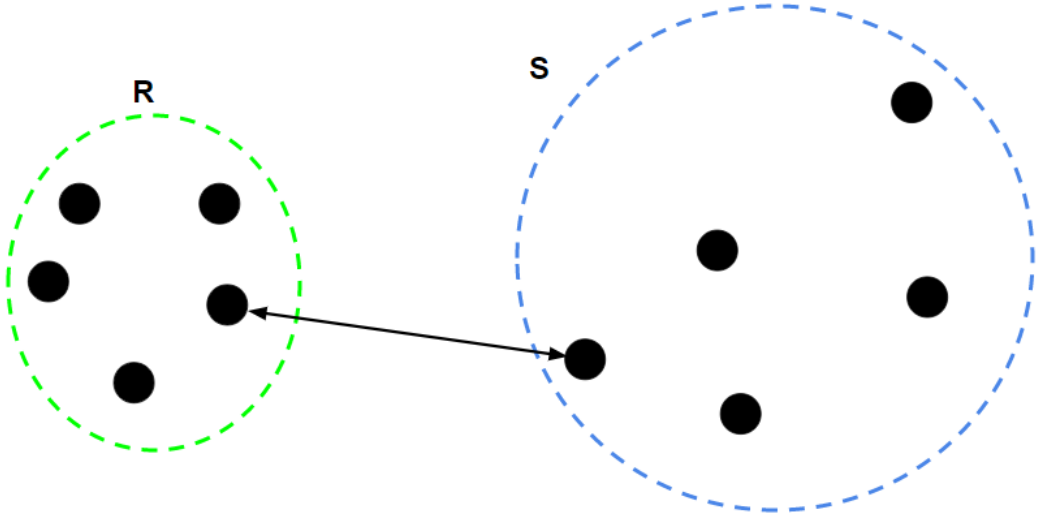
Deux approches





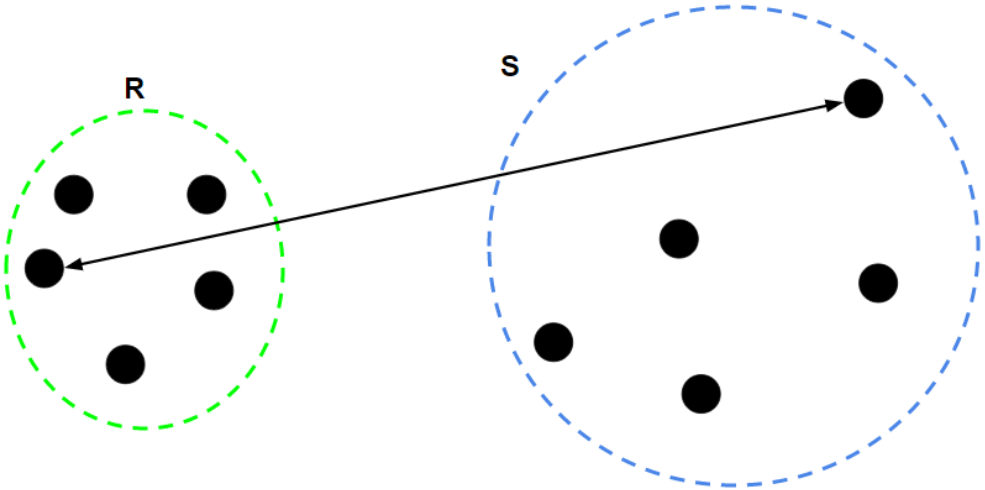
Mesurer la dissimilarité entre deux groupes d'observations ?

- **Single Linkage:** For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.



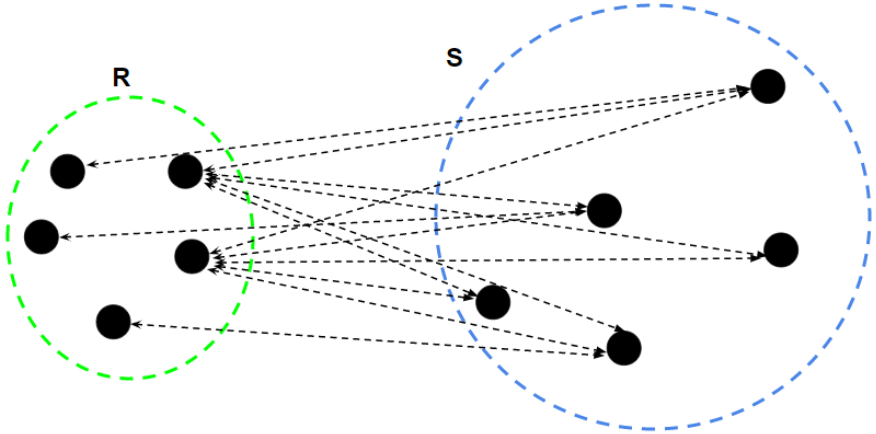
Mesurer la dissimilarité entre deux groupes d'observations ?

- **Complete Linkage:** For two clusters R and S, the complete linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.



Mesurer la dissimilarité entre deux groupes d'observations ?

- **Average Linkage:** For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.



Préparation des données

1. Traitement des valeurs manquantes

2. « scaling » des données (et oui on mesure des distances)

Dendogramme