

***k*-means**

Algorithme des *k*-moyennes :

1. On choisit le nombre de groupes K .
2. On partitionne aléatoirement les n observations en K groupes.
3. On calcule le vecteur-moyenne pour chacun des K groupes, soit

$$\mu_k = \frac{1}{N_k} \sum_{i: C(i)=k} x_i, \quad k = 1, \dots, K,$$

où N_k est le nombre d'observations dans le group k .

4. On calcule la distance entre chaque observation et chacun des K vecteurs-moyennes.
5. On assigne chacune des n observations au groupe dont le vecteur-moyenne est le plus près.
6. On répète les étapes 3 à 5 jusqu'à ce qu'aucune observation ne soit réassignée à un nouveau groupe.

Exemple

Inconvénients des *k*-means :

- Sensible au choix des centroïdes initiaux (à l'initialisation).
- Il faut connaître le nombre de groupes.
- Nécessite une mesure de distance recalculée à chaque itération.
- Assez sensible aux valeurs extrêmes.

Expliciter chaque nconvénients et de potentielles solutions.

k -médoides

On n'utilise plus le centre, mais l'observation qui minimise les distances dans chaque groupe.
La médiane plutôt que la moyenne.

Avantages :

- Permet d'intégrer des variables ordinales
- Robuste
- Permet de bien spécifier la matrix de distance

Inconvénients :

- Il faut connaître le nombre de groupes.