Algorithme des k-moyennes

Véronique Tremblay

Regroupement

L'idée générale

On veut partitionner les n clients en K groupes de façon à ce que

- les observations à l'intérieur d'un groupes soient le plus similaire possible
- les observations de deux groupes différents soient le plus différent possible

Comment?

- Aucun algorithme ne garantit de trouver un optimum global.

- Il faudrait faire tous les regroupements possibles...

$$\frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\binom{K}{k}k^n$$

À titre d'exemple, il y a 8.5896253×10^{46} façons de partitionner 100 individus en 3 groupes.

Algorithme des k-moyennes

Conditions

- Variables quantitatives
- Distances euclidienne

Exemple

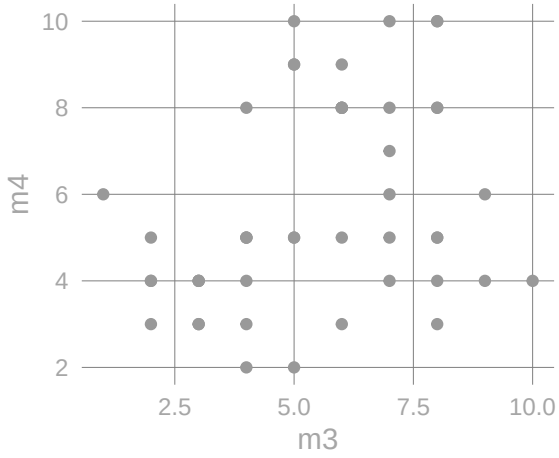
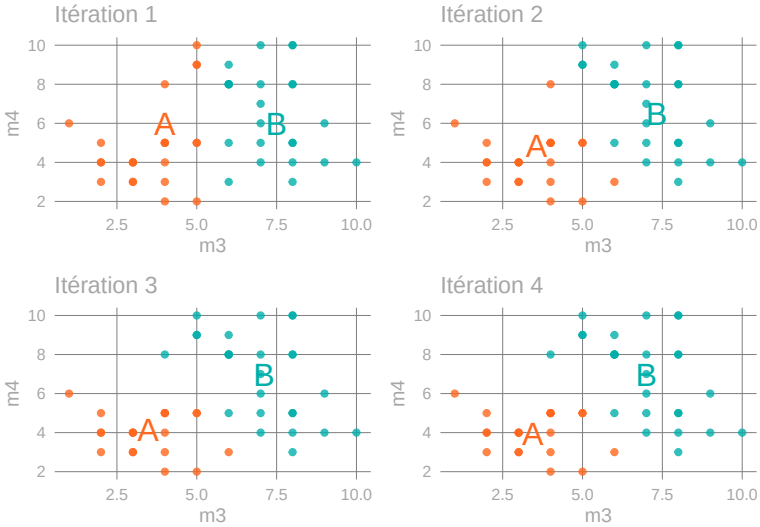
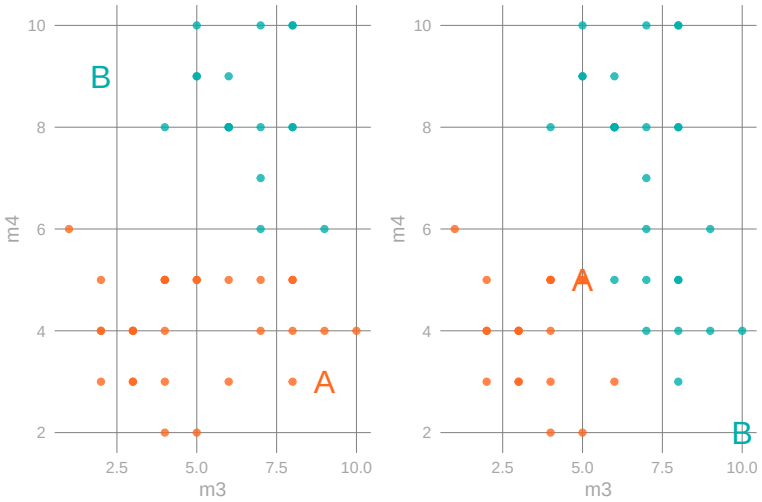


Illustration de l'algorithme des k-moynne

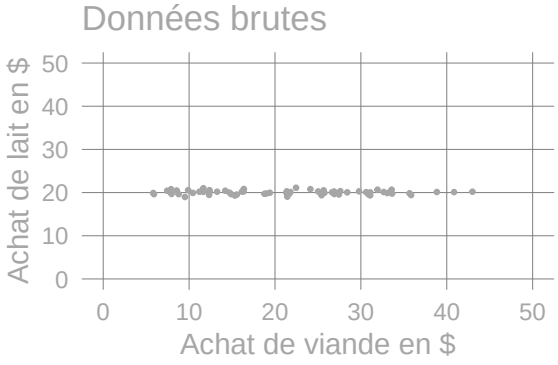


Sensibilité au choix des centroïdes initiaux



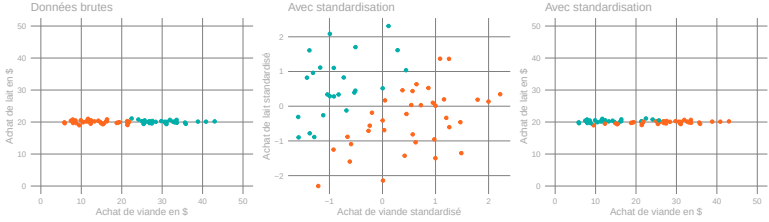
Impact de la standardisation

Un épicier souhaite faire une segmentation de sa clientèle sur la base des achats effectués par les clients.



Effet de la standardisation

Résultat de l'algorithme des k-moyennes



k-moyenne avec R

On retire les variables qui ne sont pas continues et l'identifiant

On utilise ensuite la fonction `kmeans`.

```
kmoy <- kmeans(var_kmoy, # Jeu de données
           centers = 6, # Nombre de groupes ou
                        # Centroïdes initiaux choisis
           iter.max = 10, # Nombre maximal d'itération
           nstart = 5, # Nombre de centroïdes initiaux testés
           algorithm = c("Hartigan-Wong") # Algorithme
       )
```

Faiblesses des k-moyennes

- Sensible au choix des centroïdes initiaux
- Il faut connaître le nombre de groupes
- Nécessite une mesure de distance recalculée à chaque itération
- N'accepte en théorie que les **variables continues**
- Assez sensible aux valeurs extrêmes

Résumé

- Méthode des k-moyennes
- Plusieurs limites, mais encore très utilisée et souvent très efficace.