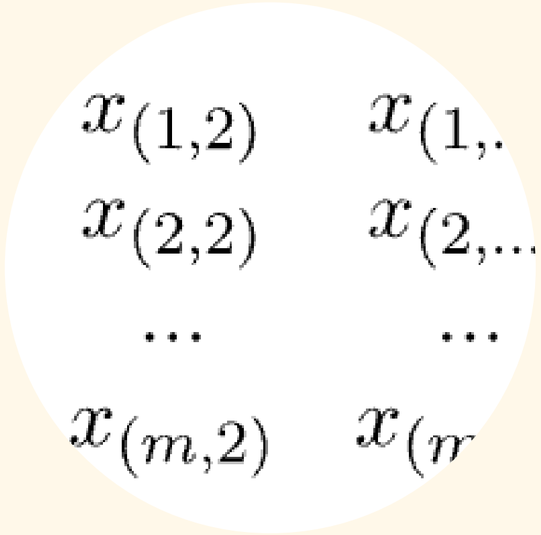
Ophélie Sabanowski

L'ALGOITHME K-MEANS

(k-moyennes)



SOMMAIRE

QU'EST CE QUE LE CLUSTERING ?

Le clustering est une méthode d'apprentissage non supervisé (unsupervised learning). Ainsi, on n'essaie pas d'apprendre une relation de corrélation entre un ensemble de features X d'une observation et une valeur à prédire Y, comme c'est le cas pour l'apprentissage supervisé.

Le clustering va regrouper en plusieurs familles (clusters) les individus/objets en fonction de leurs caractéristiques.

Il existe deux types de clustering :

- Le clustering hiérarchique
- Le clustering non-hiérarchique (partitionnement)

QU'EST CE QUE LE K-MEANS ?

K-means est un algorithme non supervisé de clustering non hiérarchique.

Il permet de regrouper en clusters distincts les observations du data set. Ainsi les données similaires se retrouveront dans un même cluster. Par ailleurs, une observation ne peut se retrouver que dans un cluster à la fois (exclusivité d'appartenance). Une même observation, ne pourra donc, appartenir à deux clusters différents.

Pour pouvoir regrouper un jeu de données en cluster distincts, l'algorithme K-Means a besoin d'un moyen de comparer le degré de similarité entre les différentes observations. Ainsi, deux données qui se ressemblent, auront une distance de dissimilarité réduite, alors que deux objets différents auront une distance de séparation plus grande.

La distance Euclidienne

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{n} (x_{1n} - x_{2n})^2}$$

CHOISIR K : LE NOMBRE DE CLUSTERS

Pour un même jeu de données, il n'existe pas un unique clustering possible. La difficulté résidera donc à choisir un nombre de cluster qui permettra de mettre en lumière des patterns intéressants entre les données. Malheureusement il n'existe pas de procédé automatisé pour trouver le bon nombre de clusters.

La méthode la plus usuelle pour choisir le nombre de clusters est de lancer K-Means avec différentes valeurs de et de calculer la variance des différents clusters. La variance est la somme des distances entre chaque centroid d'un cluster et les différentes observations inclues dans le même cluster. Ainsi, on cherche à trouver un nombre de clusters de telle sorte que les clusters retenus minimisent la distance entre leurs centres (centroids) et les observations dans le même cluster. On parle de minimisation de la distance intra-classe.

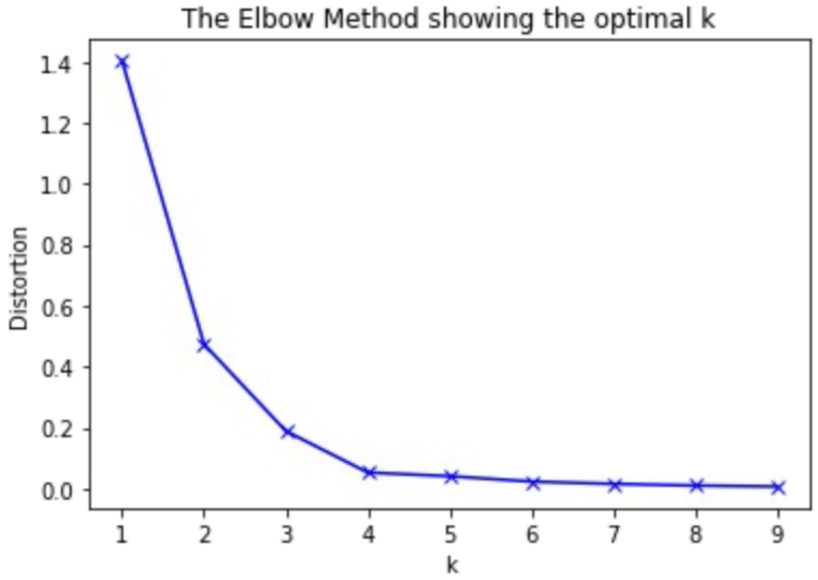
La variance des clusters se calcule comme suit :

$$V = \sum_j \sum_{x_i \to c_j} D(c_j, x_i)^2$$

Avec :

- c_j : Le centre du cluster (le centroïd)
- x_i : la ième observation dans le cluster ayant pour centroïd c_j
- $D(c_j, x_i)$: La distance (euclidienne ou autre) entre le centre du cluster et le point x_i

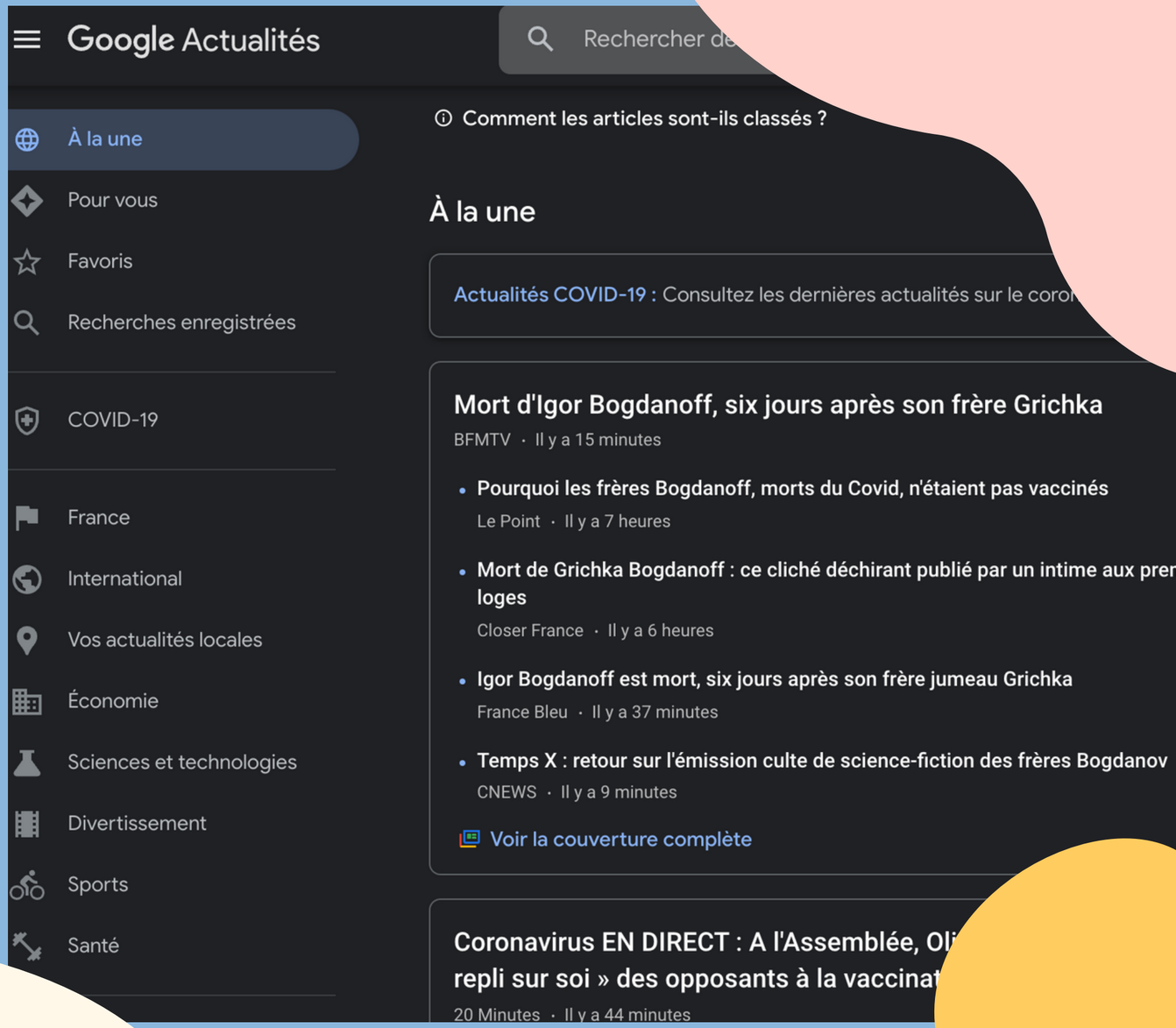
Généralement, en mettant dans un graphique les différents nombres de clusters K en fonction de la variance, on retrouve un graphique similaire à celui-ci :



CAS D'UTILISATION K-MEANS

K-Means en particulier et les algorithmes de clustering de façon générale ont tous un objectif commun : Regrouper des éléments similaires dans des clusters. Ces éléments peuvent être tous et n'importe quoi, du moment qu'ils sont encodés dans une matrice de données.

Exemples :
La segmentation de la clientèle en fonction d'un certain critère
Clustering de documents
Lors de l'exploration de données pour déceler des individus similaires



FONCTIONNEMENT DE L'ALGORITHME

k-means est un algorithme itératif qui minimise la somme des distances entre chaque individu et le centroïd. Le choix initial des centroïdes conditionne le résultat final.

Entrée :

- K le nombre de cluster à former
- Le Training Set (matrice de données)

DEBUT

Choisir aléatoirement K points (une ligne de la matrice de données). Ces points sont les centres des clusters (nommé centroïd).

REPETER

Affecter chaque point (élément de la matrice de donnée) au groupe dont il est le plus proche au son centre

Recalculer le centre de chaque cluster et modifier le centroide

JUSQU'A CONVERGENCE

OU (stabilisation de l'**inertie totale** de la population)

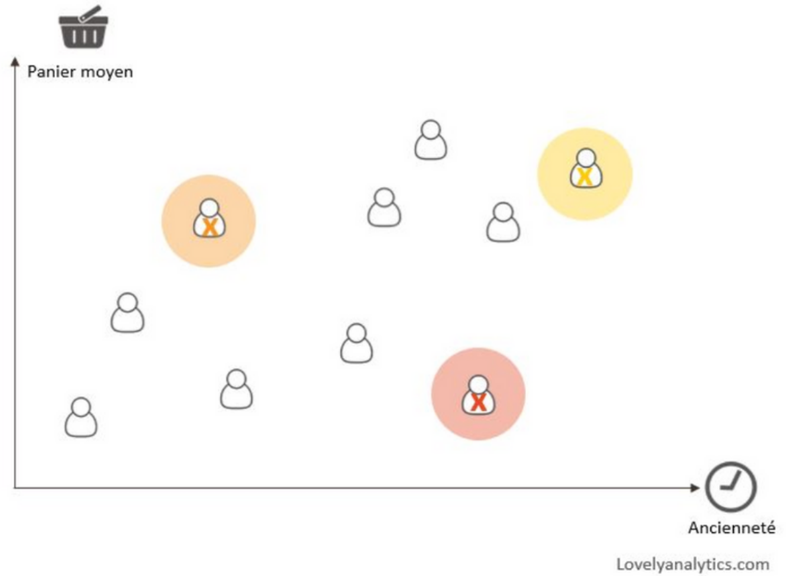
FIN ALGORITHME

EXEMPLE SIMPLE

une base de 10 clients pour lesquels on connait l'ancienneté et le panier moyen. On souhaite créer 3 groupes en utilisant la méthode des k-means.

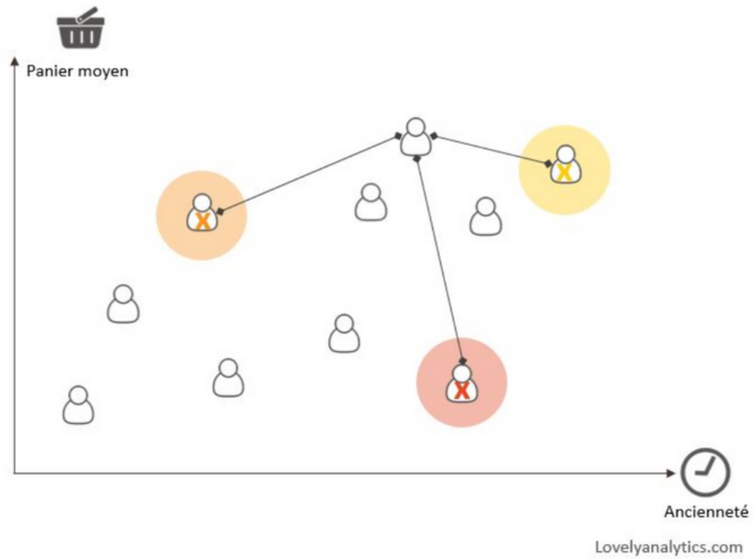
Etape 0 : Initialisation

On tire aléatoirement 3 individus. Ces 3 individus correspondent aux centres initiaux des 3 classes.



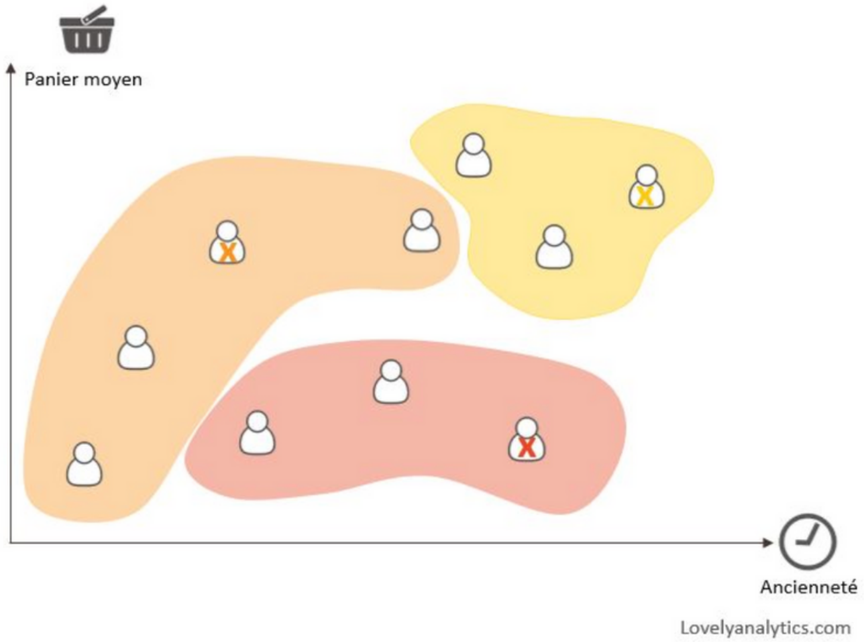
Etape 1 :

On calcule la distance entre les individus et chaque centre. Plusieurs métriques existent pour définir la proximité entre 2 individus. La méthode “classique” se base sur la distance euclidienne, vous pouvez aussi utiliser la distance Manhattan ou Minkowski.



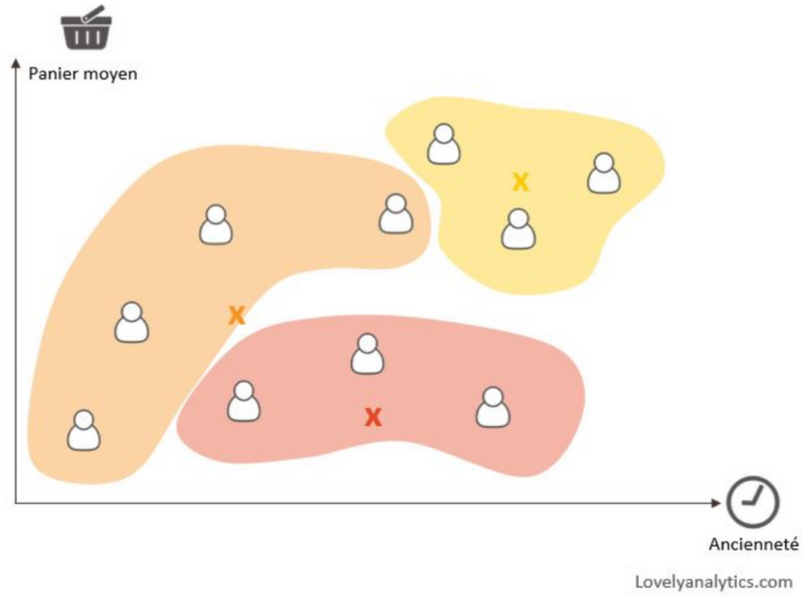
Etape 2 :

On affecte chaque individu au centre le plus proche.



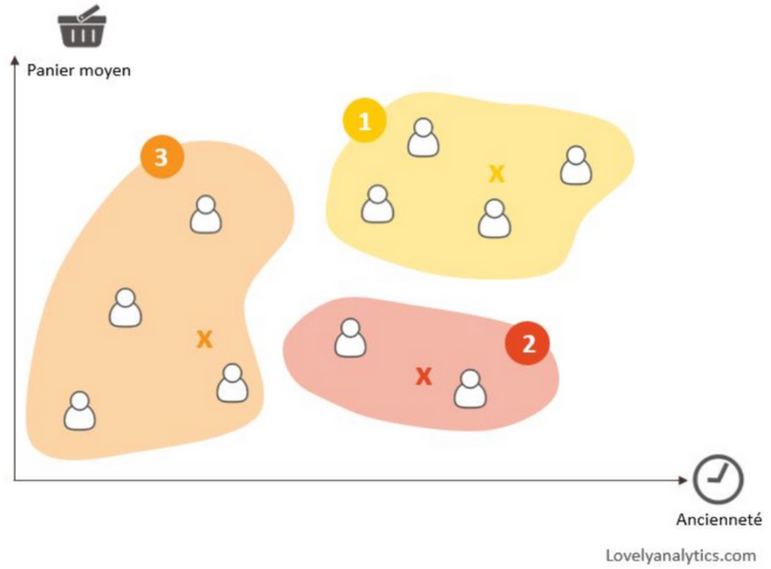
Etape 3 :

On calcule les centres de gravité des groupes qui deviennent les nouveaux centres



Boucle itérative :

On recommence les étapes 1, 2 et 3 tant que les individus sont réaffectés à de nouveaux groupes après une itération.



OPTIMUMS LOCAUX

En analysant la façon de procéder de l'algorithme de K-means, on remarque que pour **un même jeu de données, on peut avoir des partitionnements différents**. En effet, L'initialisation des tous premiers centroids est complétement aléatoire. Par conséquent l'algorithme trouvera des clusters différents en fonction de cette première initialisation aléatoire. **De ce fait, la configuration des clusters trouvés par K-Means peut ne pas être la plus optimale. On parle d'optimum local.**

Afin de palier aux problèmes des optimums locaux, **il suffit de lancer K-means plusieurs fois sur le jeu de données** (avec le même nombre de clusters et des initialisations initiales des centroids différentes) et voir la composition des clusters qui se forment. Par la suite garder celui qui convient à votre besoin.

SOURCES :

www.lovelyanalytics.com
https://mrmint.fr/algorithme-k-means

AVEZ VOUS DES QUESTIONS ?

MERCI