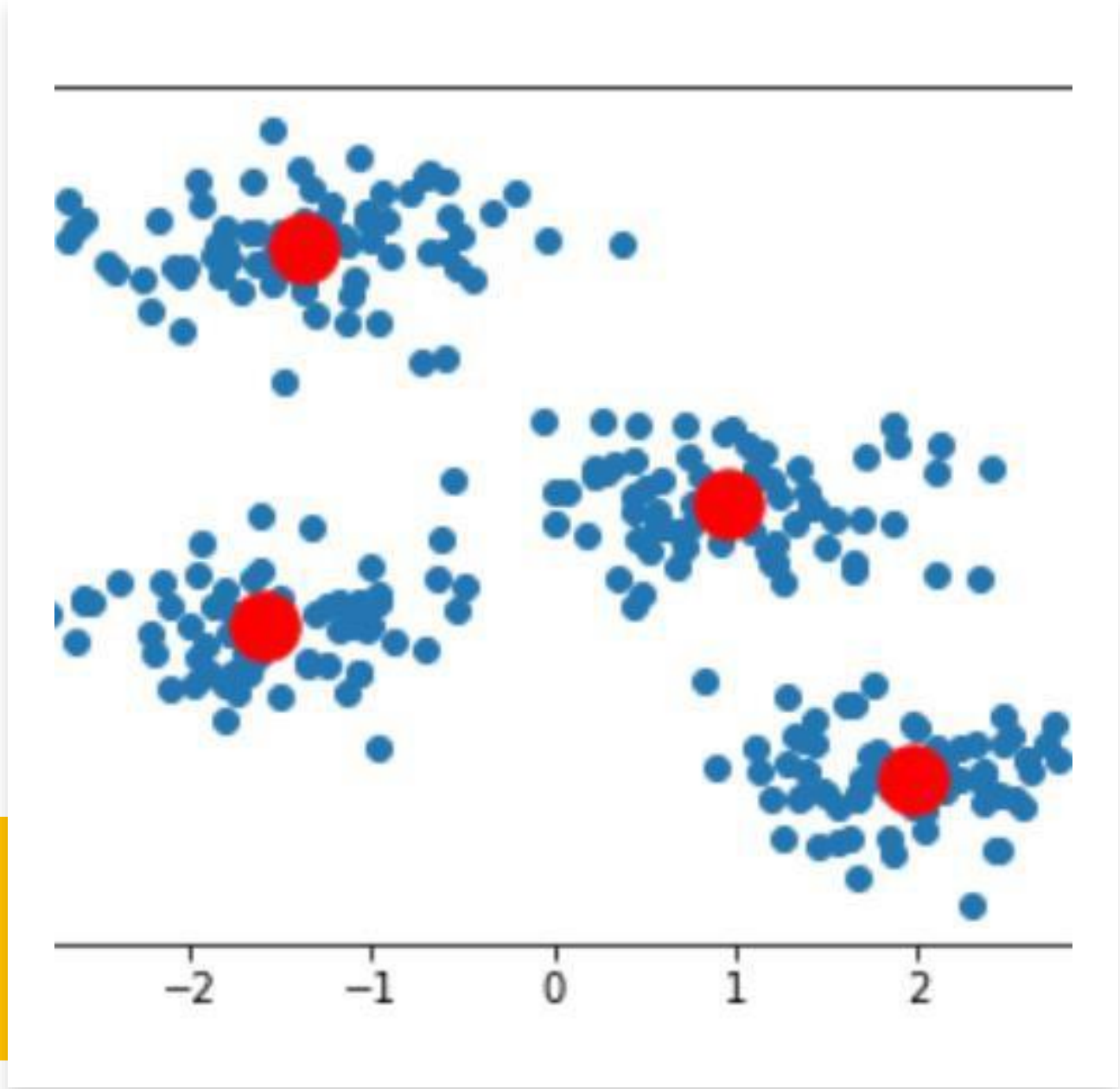


Machine Learning

Clustering

Algorithme K-means

L’apprentissage non supervisé – Machine Learning

Dans les cours précedents, nous avons parlé de l’apprentissage supervisé (**Supervised Learning**).

Dans ce cours, nous parlerons de **l’apprentissage non supervisé (Unsupervised Learning)** qui est la deuxième branche du Machine Learning.

Qu'est ce que l'apprentissage non supervisé (Unsupervised Learning) ?

A l'inverse de l'apprentissage supervisé (**Supervised Learning**) qui tente de trouver un modèle depuis des données labellisées $f(X) \rightarrow Y$ l'apprentissage non supervisé prend uniquement des données sans label (pas de variable à prédire Y).

Un algorithme d'**Unsupervised Learning** va trouver des patterns ou une structuration dans les données.

Les algorithmes de **Clustering** rentrent dans la catégorie de Unsupervised Learning. Ils permettent de regrouper en des ensembles, les données qui sont similaires.

Le Clustering

- A première vue, on pourrait penser que le Clustering a peu d'utilité dans les applications de la vraie vie. Mais détrompez-vous ! Les applications de cette technique sont nombreuses.

- Quand vous vous demandez comment **Amazon fait pour recommander** les bons produits, ou encore **YouTube** qui vous propose des vidéos en relation avec vos attentes, ou encore **Netflix** qui vous propose de bons films, tout ça c'est du **Clustering** !

- L'efficacité d'implémentation d'un algorithme de Clustering peut permettre **une augmentation significative du chiffre d'affaires** d'un site e-commerce comme pour le cas Amazon

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 d’une observation et une valeur à prédire v

- L’apprentissage non supervisé va plutôt **trouver des *patterns* dans les données.** Notamment, en regroupant les choses qui se ressemblent.

- En apprentissage non supervisé, les données sont représentées comme suit :

$$X = \begin{pmatrix} x_{(1,1)} & x_{(1,2)} & x_{(1,...)} & x_{(1,n)} \\ x_{(2,1)} & x_{(2,2)} & x_{(2,...)} & x_{(2,n)} \\ \dots & \dots & \dots & \dots \\ x_{(m,1)} & x_{(m,2)} & x_{(m,...)} & x_{(m,n)} \end{pmatrix}$$

- Chaque ligne représente un individu (une observation).

- A l'issu de l'application du clustering, on retrouvera ces données regroupées par ressemblance.

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

- les individus se trouvant dans un même cluster sont similaires et les données se trouvant dans un autre cluster ne le sont pas.

K-means
(1)

- **K-means** (k-moyennes) est un algorithme non supervisé de **clustering**, populaire en Machine Learning.

K-means
(2)

- Il permet de regrouper en K clusters distincts les observations du data set.

- 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.

K-means
(3)
Notion de similarité

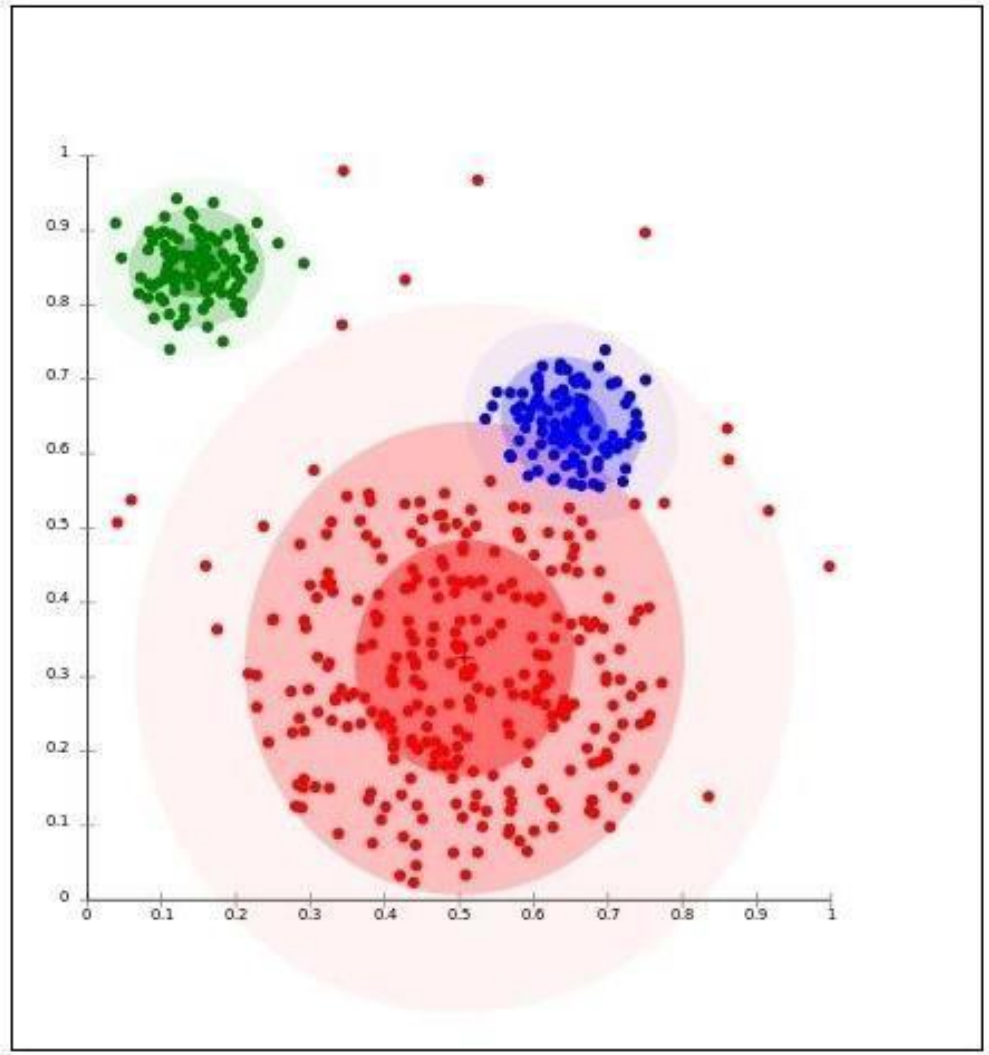
L'algorithme K-Means a besoin d'un moyen de **comparer le degré de similarité** entre les différentes observations.

➐ deux données qui se ressemblent, auront une **distance de dissimilarité** réduite, alors que deux objets différents auront une distance de séparation plus grande.

K-means (4)
Principe

- L'approche de K-Means consiste à affecter aléatoirement des centres de clusters (appelés **centroids**), et ensuite assigner chaque point de nos données au centroid qui lui est le plus proche.

- Cela s'effectue jusqu'à assigner toutes les données à un cluster.

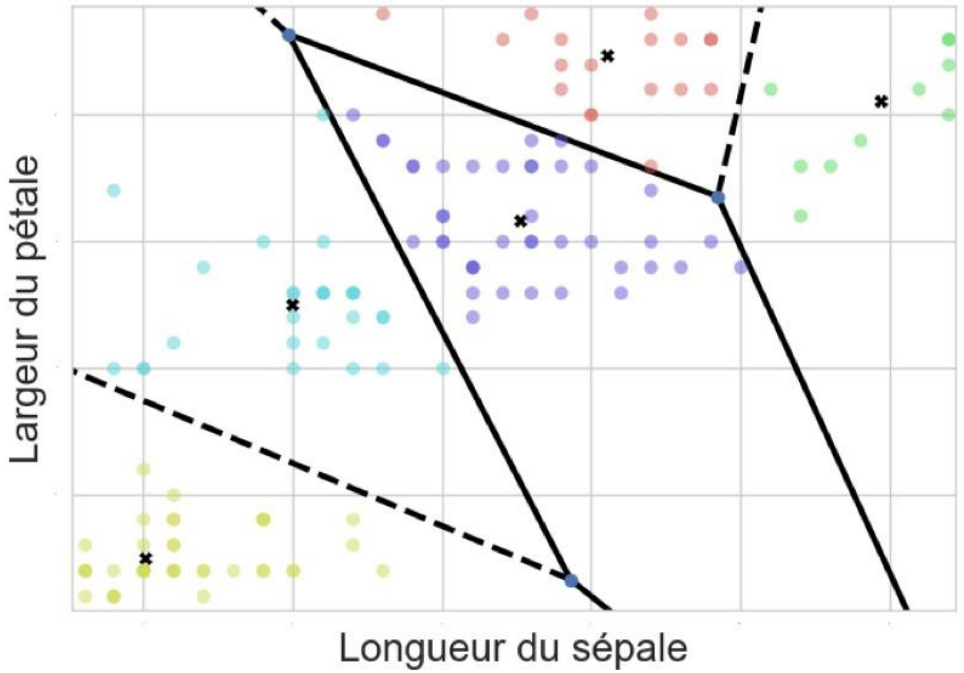


K-means
(5)
Exemple

- L'exemple suivant s'appuie sur le célèbre jeu de données "Iris" qui décrit des fleurs par l'intermédiaire des longueurs et largeurs de leurs pétales et sépales.

- Les descripteurs considérés ici sont la longueur des sépales et la largeur des pétales.

- Chaque point correspond à une fleur et la couleur associée au point reflète son appartenance à un groupe. Les centroïdes de chaque groupe sont représentés par des croix, les frontières par des traits.



Distance Euclidienne

La distance Euclidienne : C'est la distance géométrique qu'on apprend au collège.

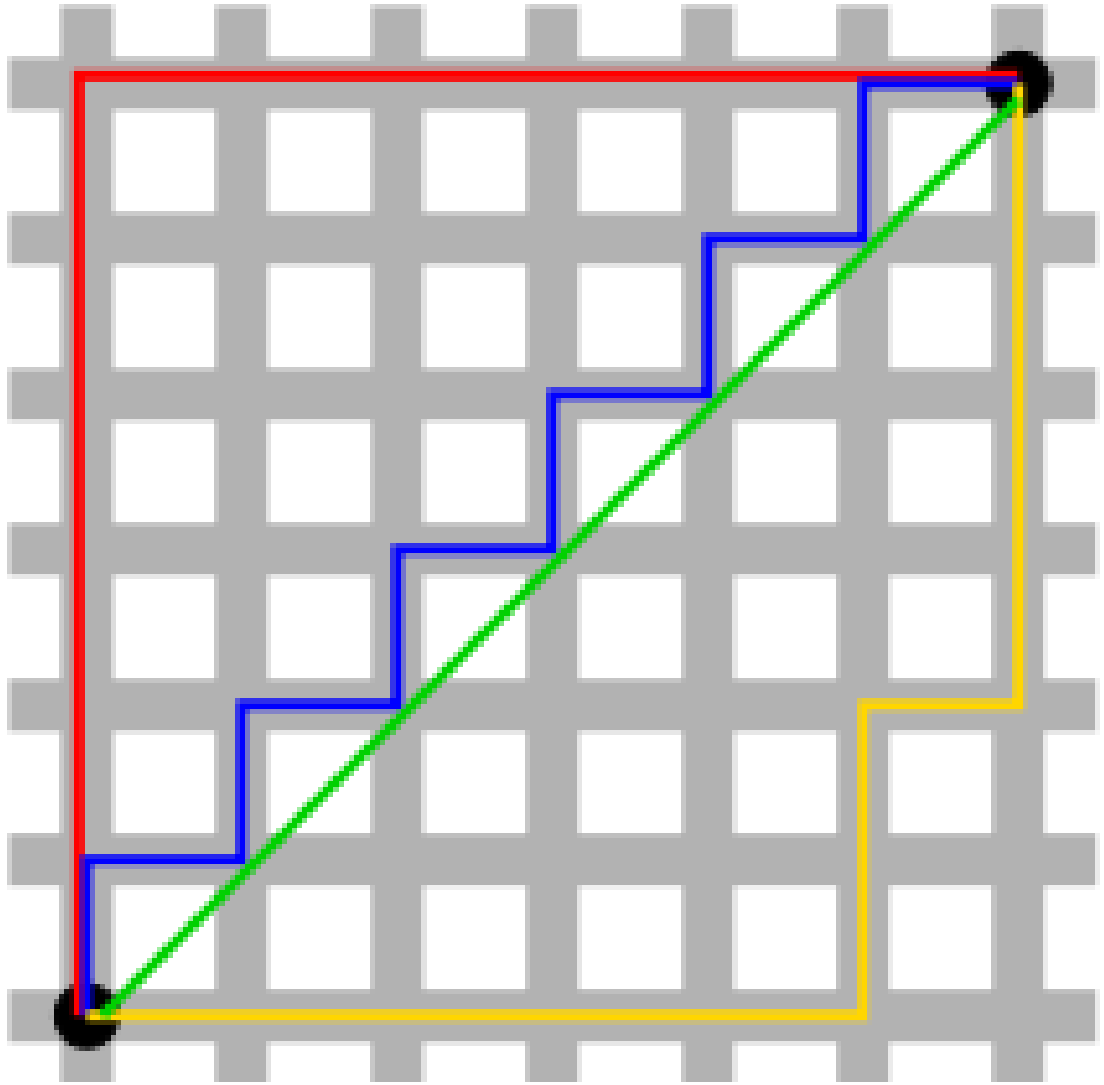
- Soit une matrice v à n variables quantitatives.
- La distance euclidienne d entre deux observations x_1 et x_2 se calcule comme suit :

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

Distance de Manhattan

La distance de Manhattan (taxi-distance) : est la distance entre deux points parcourue par un taxi lorsqu'il se déplace dans une ville où les rues sont agencées selon un réseau ou un quadrillage.

Un taxi-chemin est le trajet fait par un taxi lorsqu'il se déplace d'un nœud du réseau à un autre en utilisant les déplacements horizontaux et verticaux du réseau.



Choisir K : le nombre de clusters (1)

- Choisir un nombre de cluster K n’est pas forcément intuitif.
- Un nombre K grand peut conduire à un partitionnement trop fragmenté des données. Ce qui empêchera de découvrir des patterns intéressants dans les données.
- Un nombre de clusters trop petit, conduira à avoir, potentiellement, des cluster trop généralistes contenant beaucoup de données. Dans ce cas, on n’aura pas de patterns “fins” à découvrir.
- Pour un même jeu de données, il n’existe pas un unique clustering possible.
- il n’existe pas de procédé automatisé pour trouver le bon nombre de clusters.

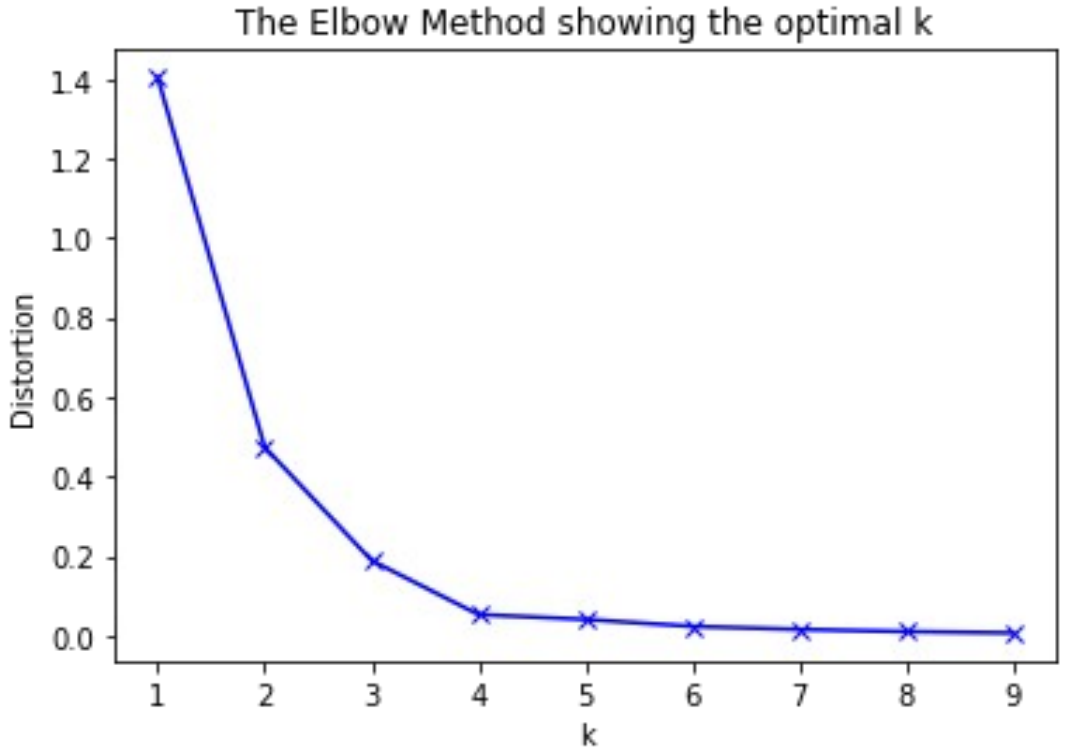
Choisir K : le nombre de clusters (2)

- Lancer K-Means avec différentes valeurs de K et 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.
- La variance des clusters se calcule comme suit : $V = \sum_j \sum_{x_i \to c_j} D(c_j, x_i)^2$
- 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

Graphique des clusters K en fonction de la variance

- Forme d'un bras où le point le plus haut représente l'épaule et le point où K vaut 9 représente l'autre extrémité : la main.

- Le nombre optimal de clusters est le point représentant le coude. Ici le coude peut être représenté par K valant 2 ou 3. C'est le nombre optimal de clusters.



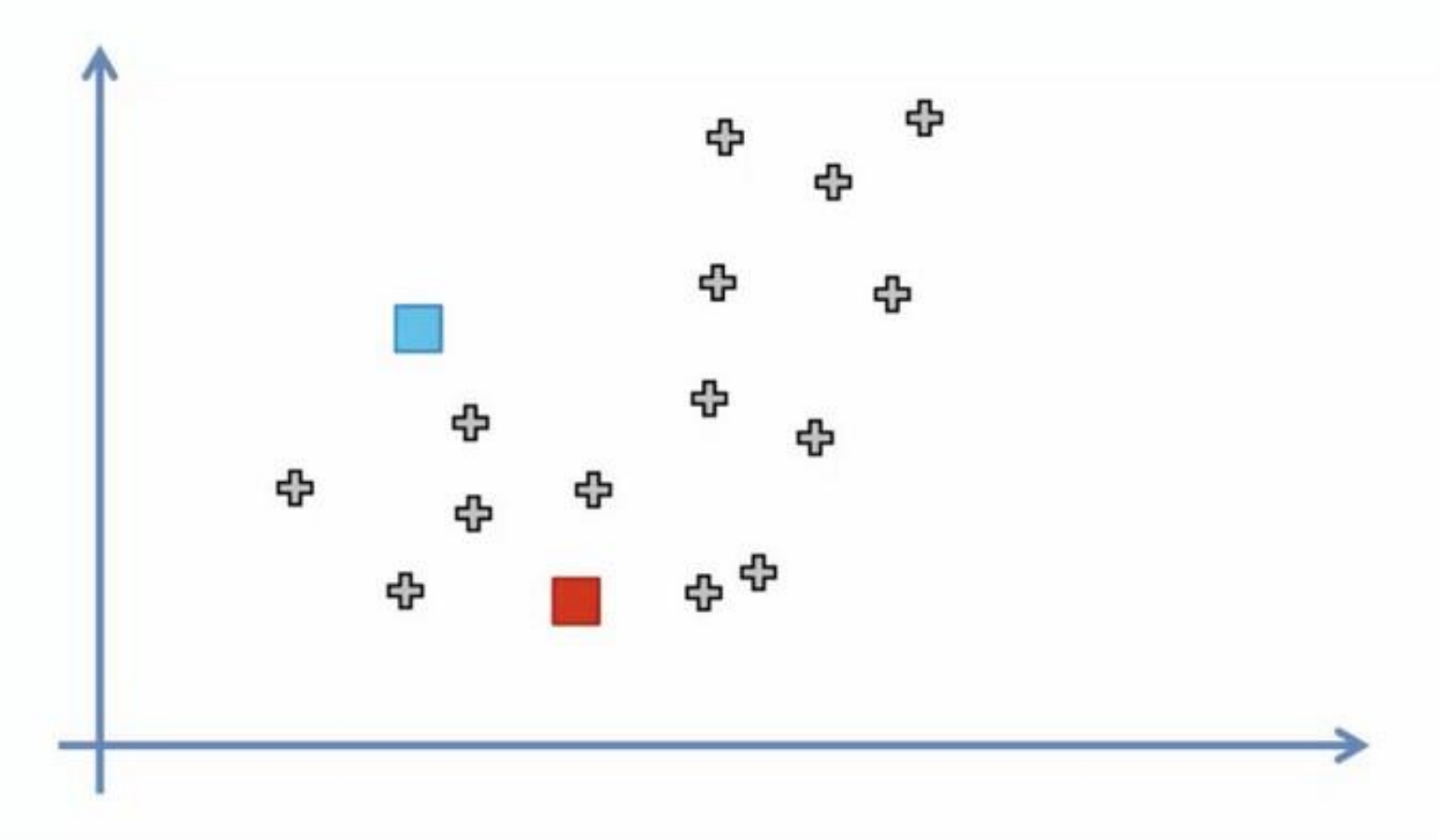
Cas d’utilisation K-means

- La segmentation de la clientèle en fonction d’un certain critère (démographique, habitude d’achat etc….)

- Utilisation du clustering en Data Mining lors de l’exploration de données pour déceler des individus similaires. Généralement, une fois ces populations détectées, d’autres techniques peuvent être employées en fonction du besoin.

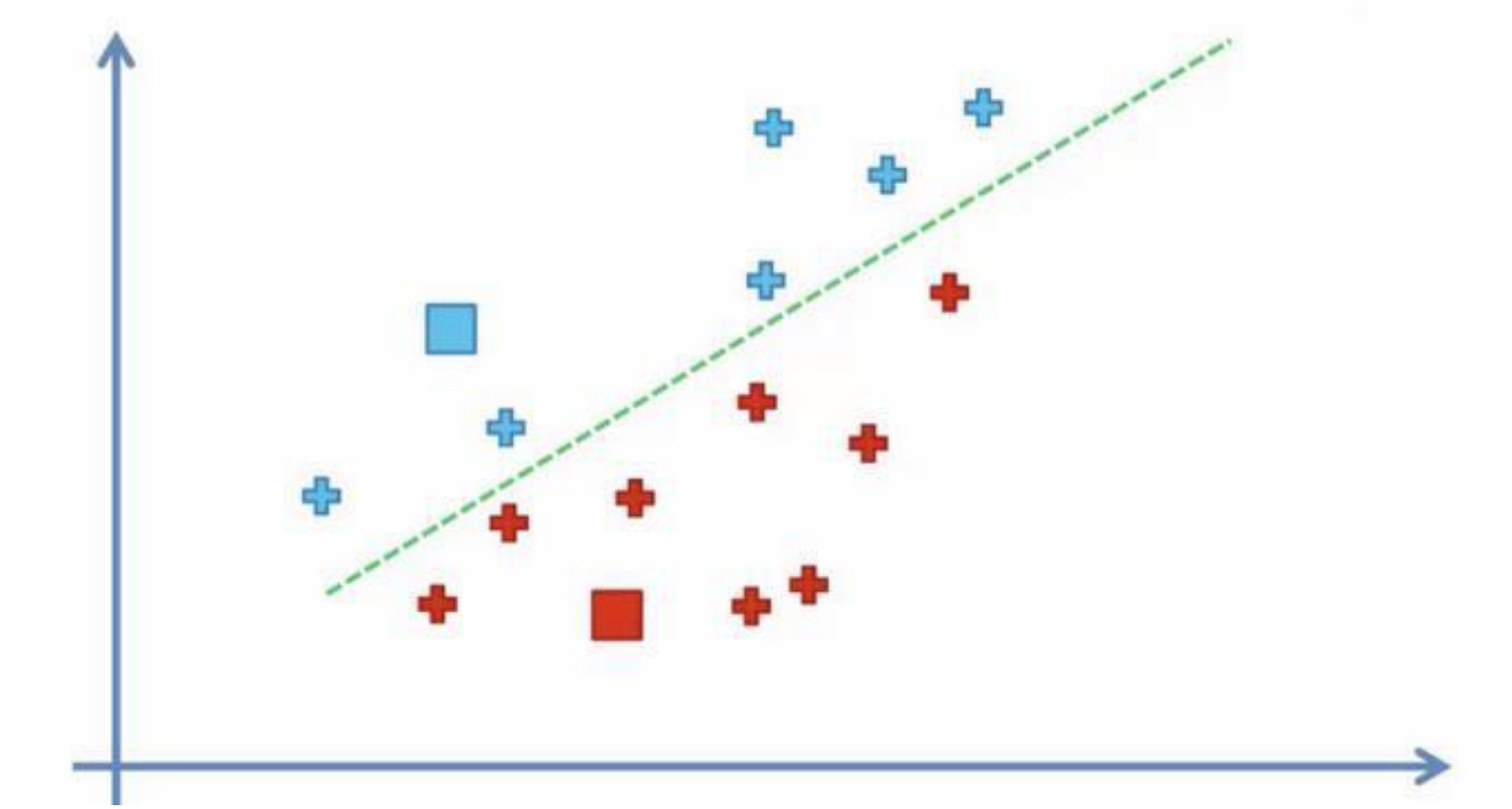
- Clustering de documents (regroupement de documents en fonction de leurs contenus. Pensez à comment Google Actualités regroupe des documents par thématiques.)

1. Select **K** (i.e. 2) random points as cluster centers called centroids



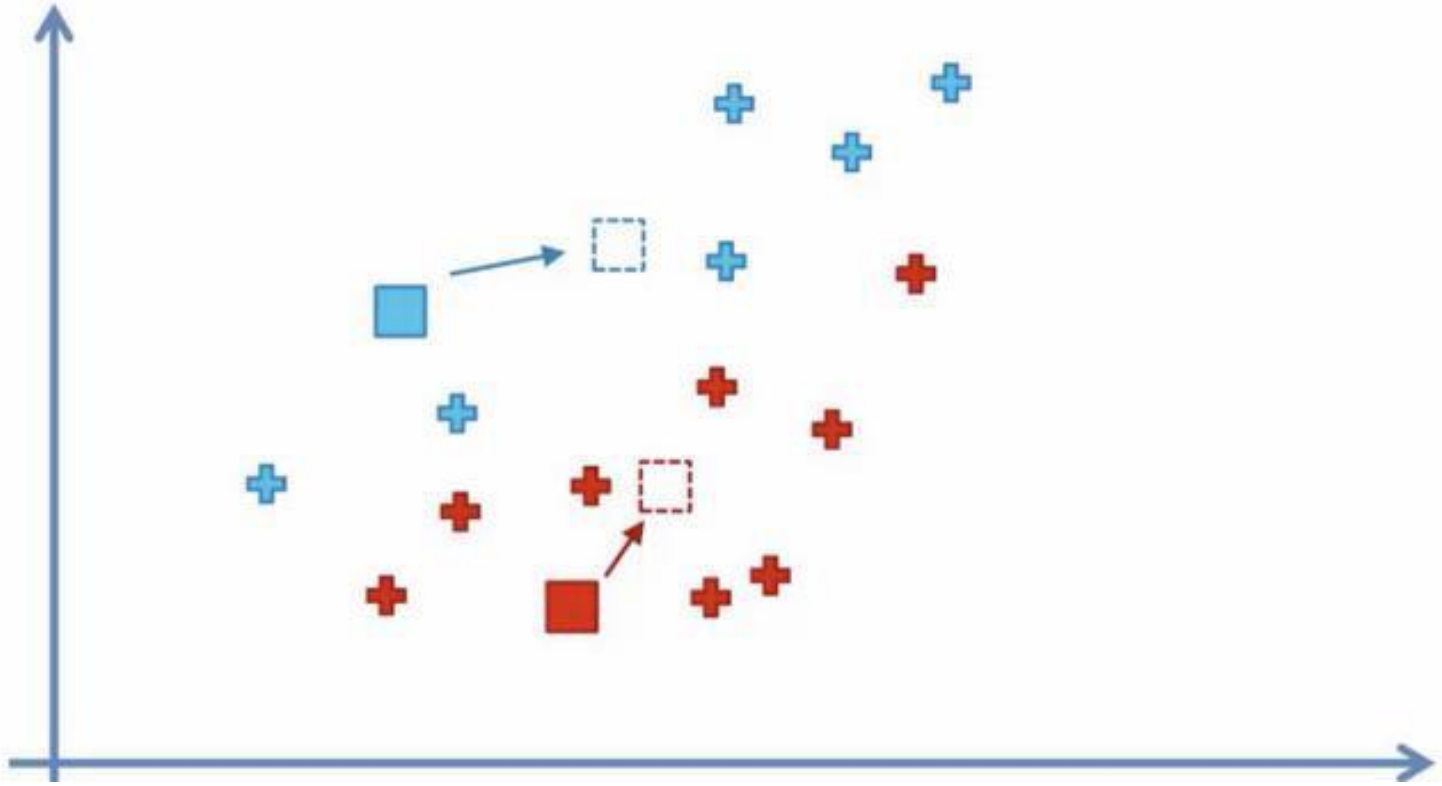
Algorithme Step1

2. Assign each data point to the closest cluster by calculating its distance with respect to each centroid



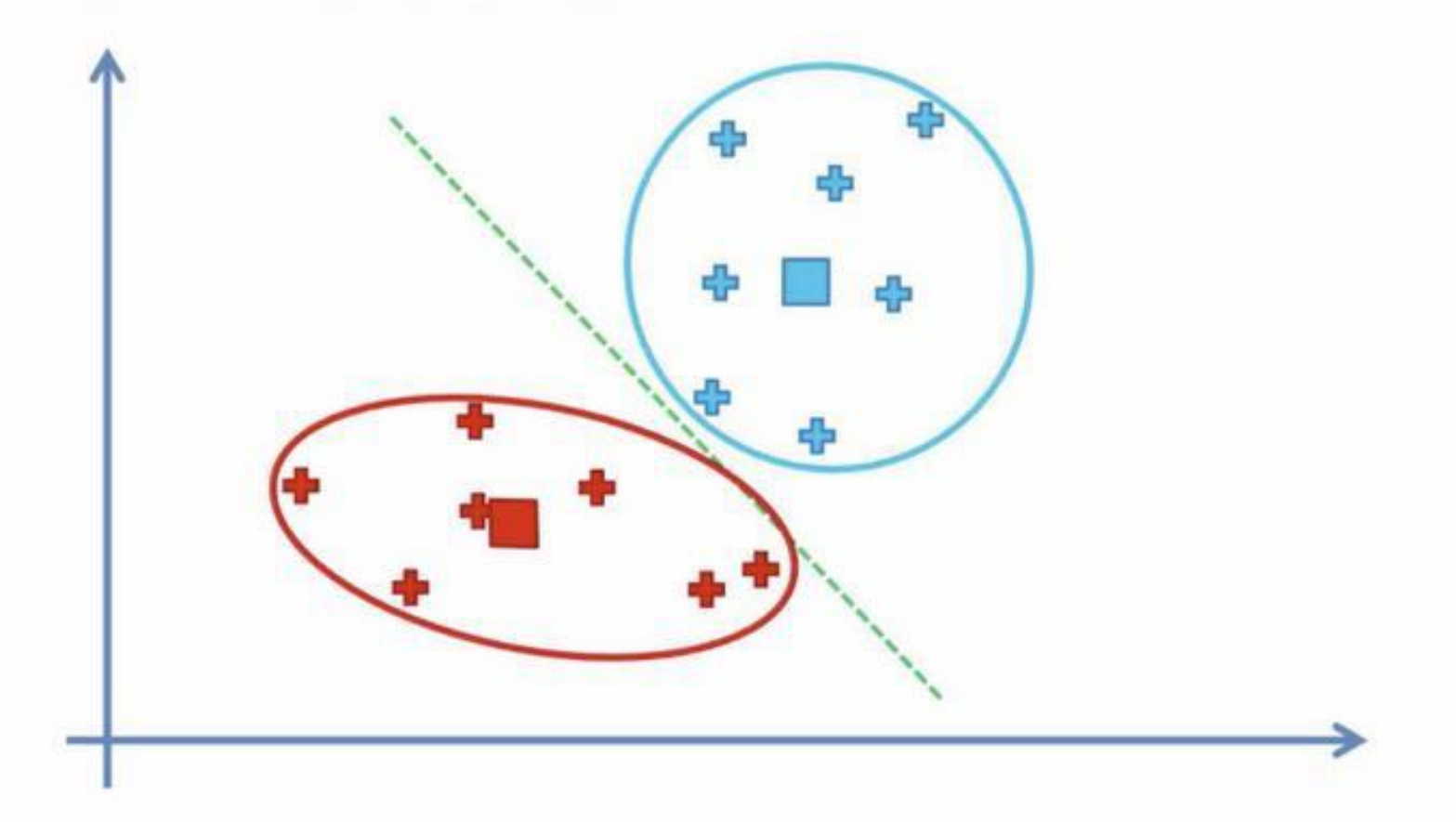
Algorithme Step 2

3. Determine the new cluster center by computing the average of the assigned points

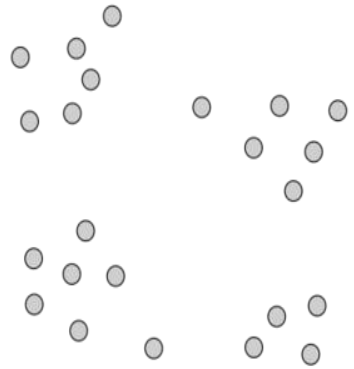


Algorithme Step 3

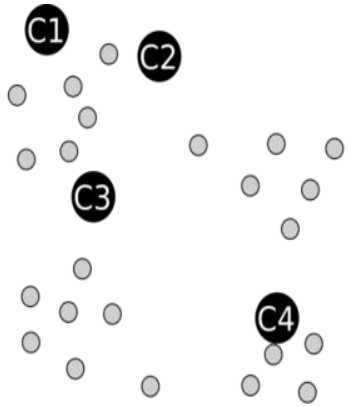
4. Repeat steps 2 and 3 until none of the cluster assignments change



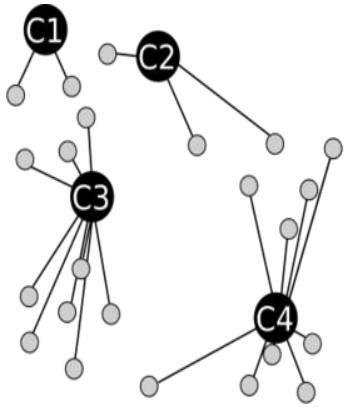
Algorithme Step 4



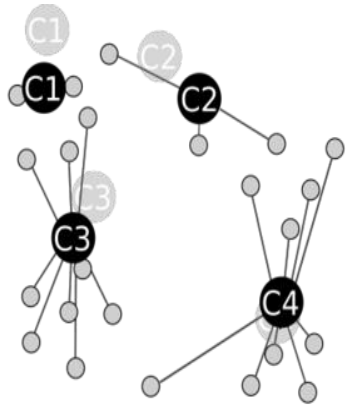
0a. Données d'entrée



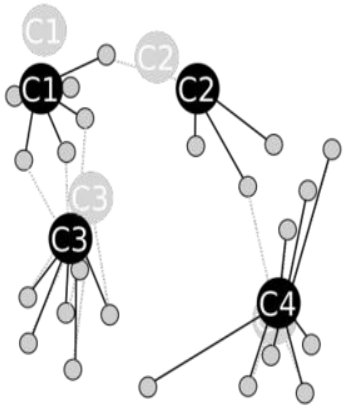
0b. intialisation



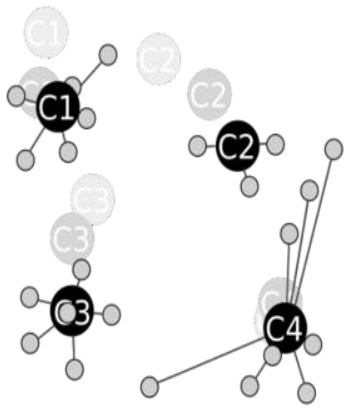
1a. assignation



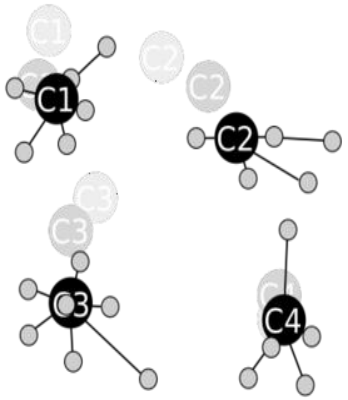
1b. calcul des points moyens



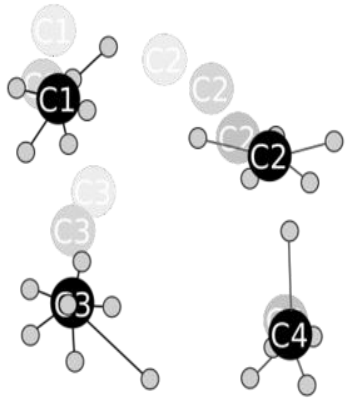
2a. assignation



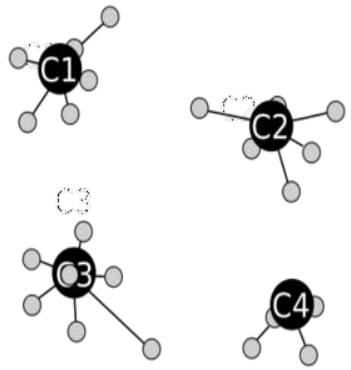
2b. calcul des points moyens



3a. assignation



3b. calcul des points moyens



4a. assignation

clusters stables (fin)

Algorithme K-means

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

Note 1: Lors de la définition de l'algorithme, quand je parle de "point", c'est un point au sens "donnée/data" qui se trouve dans un espace vectoriel de dimension n. Avec n : le nombre de colonnes de la matrice de données.

Note 2 : La convergence de l'algorithme K-Means peut être l'une des conditions suivantes :

- Un nombre d'itérations fixé à l'avance, dans ce cas, K-means effectuera les itérations et s'arrêtera peu importe la forme de clusters composés.
- Stabilisation des centres de clusters (les centroids ne bougent plus lors des itérations).