Université Cergy Paris


# Diplôme universitaire : Data Analyst

**Examen :**
**Réduction de dimensionnalité**

11 décembre 2022

**Haury Fabien**

# Examen Réduction de dimensionnalité

Haury Fabien

11 décembre 2022

# Table des matières

# Chapitre 1

# Questions générales

Comment décririez-vous la différence entre apprentissage supervisé et apprentissage non supervisé ?

La différence entre apprentissage supervisé et apprentissage non supervisé peut être vue comme la différence entre un étudiant suivant soit un cours en classe avec un professeur (supervisé), soit comme un étudiant apprenant en autodidacte (non supervisé).

Apprentissage supervisé :

- Entrainé sur des données étiquetées (label)
- Prédit la sortie (output)
- Les données d'entrée (input) sont fourni ainsi que la sortie
- Le but est d'entraîner le modèle pour pouvoir prédire l'output quand on lui donne des nouveaux inputs
- Séparé en deux grandes catégories : classification et régression

Apprentissage non supervisé :

- Entrainé sur des données non étiquetées
- Permets de trouver des patterns dans les données
- Seules les données d'entrée sont données
- Le but est d'entraîner le modèle pour trouver des patterns et avoir des perspectives utiles sur le jeu de données
- Séparé en différentes grandes catégories : clustering, association et réduction de dimensionnalité

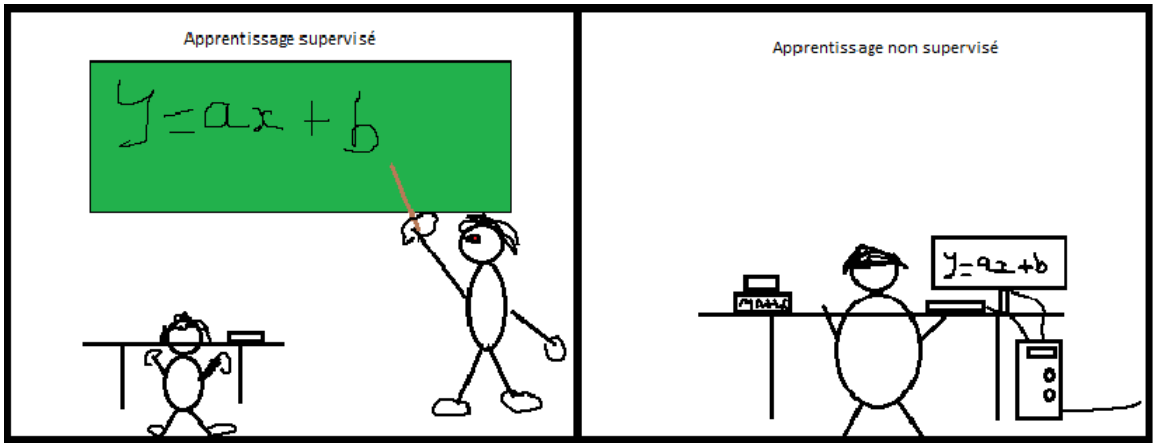


FIGURE 1 – Illustration entre l'apprentissage supervisé et non supervisé

# Chapitre 2

# Réduction de dimensionnalité

## 2.1 Dans une ACP, à quoi correspond la qualité de la représentation d'un individu ou d'une variable ?

La corrélation de chaque point sur un axe exprime la qualité de représentation du point sur l'axe. Elle prend des valeurs entre 0 (pas corrélé du tout) et 1 (fortement corrélé). Si cette valeur est proche de 1, alors le point est bien représenté sur l'axe.

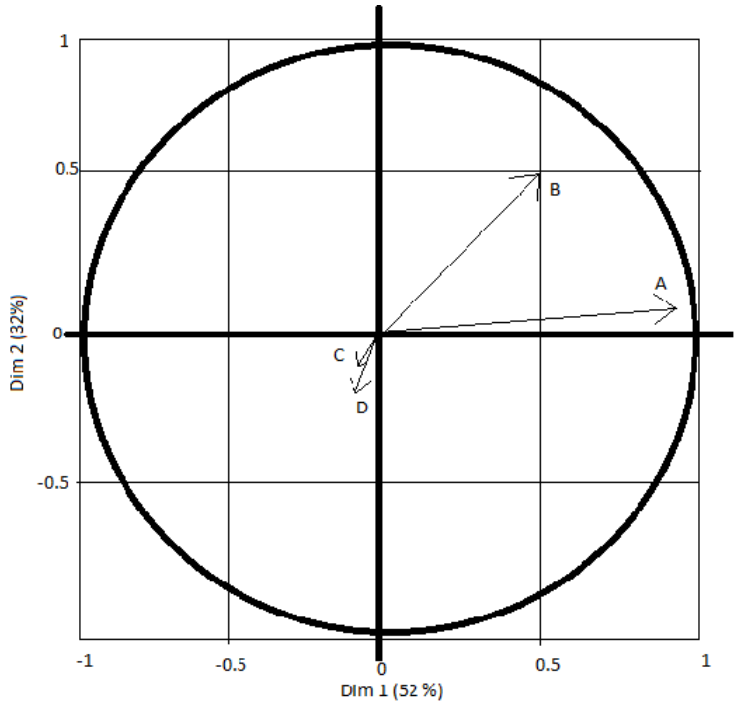


FIGURE 2 – Cercle de corrélation d'une ACP

Nous pouvons voir que sur la Figure 2, la variable A est la mieux représentée car sa flèche est presque égale à un. La variable B est la seconde mieux représentée. Les variables C et D sont les moins bien représentées car leurs flèches respectives sont courtes mais ces deux variables sont fortement corrélées entre-elles.

## 2.2 A quoi correspond une eigenvalue ?

Les valeurs propres (eigenvalues) mesurent la quantité de variance expliquée par chaque axe principal. Les valeurs propres sont grandes pour les premiers axes et petits pour les axes suivants. Autrement dit, les premiers axes correspondent aux directions portant la quantité maximale de variation contenue dans le jeu de données.

## 2.3 A quoi correspond la contribution d'un individu à un axe ?

La contribution d'un individu est la contribution relative de cet individu à la variance d'un axe factoriel. Ainsi, la contribution d'un individu est une mesure de l'importance d'un individu sur un axe factoriel. Plus la contribution d'un individu sera importante et plus il aura de poids sur ce facteur.

## 2.4 Comment définiriez-vous la saturation d'une variable sur un axe ?

Dès lors que les composantes constituent des sortes de « nouvelles » variables synthétiques, on peut examiner la relation entre les variables originales et les composantes. Plus cette relation est forte, plus la variable est « expliquée » par le facteur. Cette relation, qui peut s'exprimer par un chiffre variant de -1 à +1 s'appelle la « saturation » (factor loading) de la variable sur le facteur.

## 2.5 Dans quel contexte utilise-t-on l'analyse des correspondances multiples ?

On utilise une analyse des correspondances multiples (ACM) sur des données nominales. Les données nominales peuvent être à la fois qualitatives et quantitatives. Cependant, les étiquettes quantitatives sont dépourvues de valeur numérique ou de relation (par exemple, un numéro d'identification). D'autre part, divers types de données qualitatives peuvent être représentés sous forme nominale. Il peut s'agir de mots, de lettres et de symboles. Le nom des personnes, le sexe et la nationalité ne sont que quelques-uns des exemples les plus courants de données nominales.

# Chapitre 3

# Clustering

## 3.1 Expliquer en détails le fonctionnement de l'algorithme k-means

Le clustering K-means utilise des "centroïdes", K points différents pris au hasard dans les données, et attribue chaque point de données au centroïde le plus proche. Après l'attribution de chaque point, le centroïde est déplacé vers la moyenne de tous les points qui lui ont été attribués. Puis le processus se répète : chaque point est assigné à son centroïde le plus proche, les centroïdes sont déplacés vers la moyenne des points qui lui sont assignés. L'algorithme est terminé lorsqu'aucun point ne change de centroïde assigné.

1. Initialiser K centroïdes aléatoires.
2. Pour chaque point de données, regardez quel centroïde est le plus proche de celui-ci par calcul de la distance Euclidienne ou bien la distance de Manhattan, etc.
3. Assignez le point de données au centroïde le plus proche.
4. Pour chaque centroïde, déplacez le centroïde vers la moyenne des points assignés à ce centroïde.
5. Répétez les trois dernières étapes jusqu'à ce que l'affectation des centroïdes ne change plus.

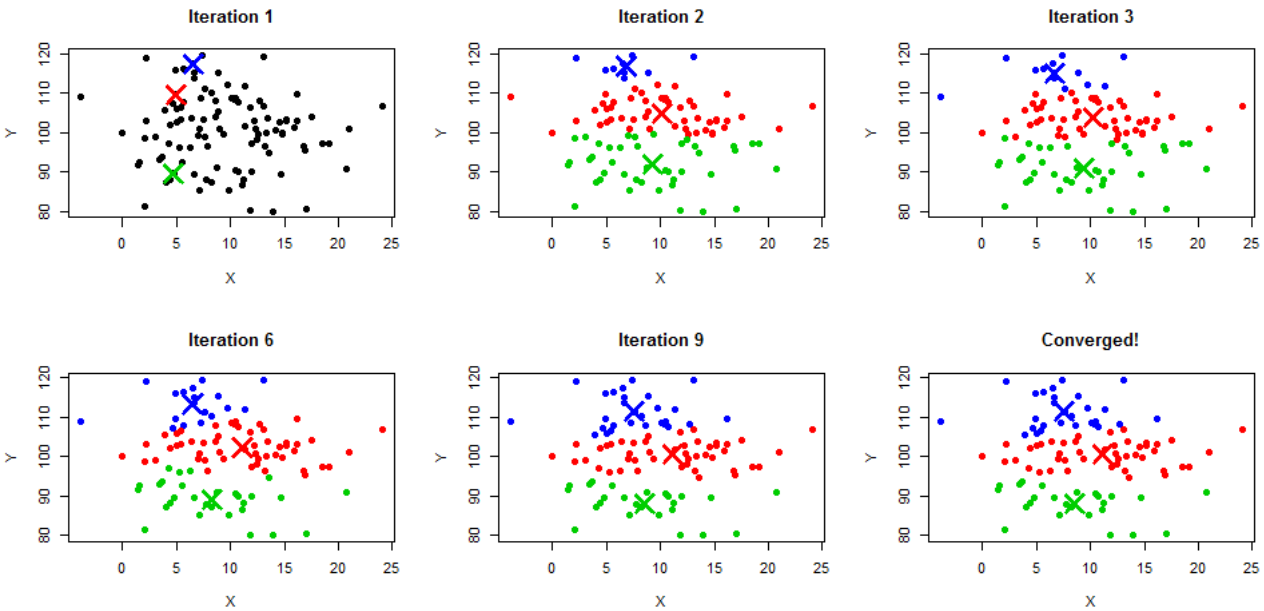


FIGURE 3 – Cheminement de fonctionnement d'un algorithme k-means

La Figure 3 représente le cheminement de fonctionnement d'un algorithme k-means. L'itération 1 correspond au point 1, c.à.d. l'initialisation des K centroïdes. Les points 2 et 3 sont ensuite effectués, cela correspond à la colorisation des points par rapport à la distance aux centroïdes. Le point 4 est ensuite effectué, chaque centroïdes est déplacé vers sa nouvelle position. Nous sommes maintenant à l'itération 2. Les étapes 2, 3 et 4 sont maintenant répétés autant de fois que nécessaires jusqu'au moment ou chaque centroïdes ne bouge plus.

## 3.2 A quoi correspondent au juste la WCSS et la BCSS dans le cadre du k-means ?

La somme des carrés à l'intérieur d'un cluster (within-cluster sum of squares, WCSS) est la mesure de la variabilité des observations à l'intérieur de chaque cluster. Un cluster qui a une petite somme des carrés est plus compacte qu'un cluster qui a une grande somme des carrés. Les clusters qui ont des valeurs plus élevées présentent une plus grande variabilité des observations au sein du cluster. La WCSS peut-être vue comme la mesure de compacité d'un cluster. Sur la Figure 2, pour l'itération 2, le groupe bleu possède un WCSS plus petit que le groupe vert.

La somme des carrés entre les clusters (Between Clusters Sum of Squares, BCSS) est la mesure de la distance moyenne au carré entre tous les centroïdes. La BCSS mesure la variation entre tous les clusters. Une grande valeur peut indiquer que les clusters sont dispersés, tandis qu'une petite valeur peut indiquer que les clusters sont proches les uns des autres. La BCSS peut-être vue comme la mesure de séparation entre cluster.

Leurs formules respectives sont les suivantes :

$$WCSS = \sum_{i=1}^{N_c} \sum_{x \in C_i} d(x, \bar{x}_{C_i})^2$$

$$BCSS = \sum_{i=1}^{N_c} |C_i| * d(\bar{x}_{C_i}, \bar{x})^2$$

$$C_i = Cluster,\ N_c = Numbers\ of\ cluster,\ \bar{x}_{C_i} = Cluster\ centroide,\ \bar{x} = Sample\ mean$$

## 3.3 Expliquer en détails le fonctionnement de la classification ascendante hiérarchique

Le clustering hiérarchique, également connu sous le nom d'analyse hiérarchique des clusters, est un algorithme qui regroupe les objets similaires en groupes appelés clusters. Le point final est un ensemble de clusters, où chaque cluster est distinct de tous les autres, et les objets au sein de chaque cluster sont largement similaires les uns aux autres.

Il existe deux façons d'appliquer l'algorithme de clustering hiérarchique :

- Agglomerative : La hiérarchie est créée du bas vers le haut.
- Divisive : La hiérarchie est créée du haut vers le bas.

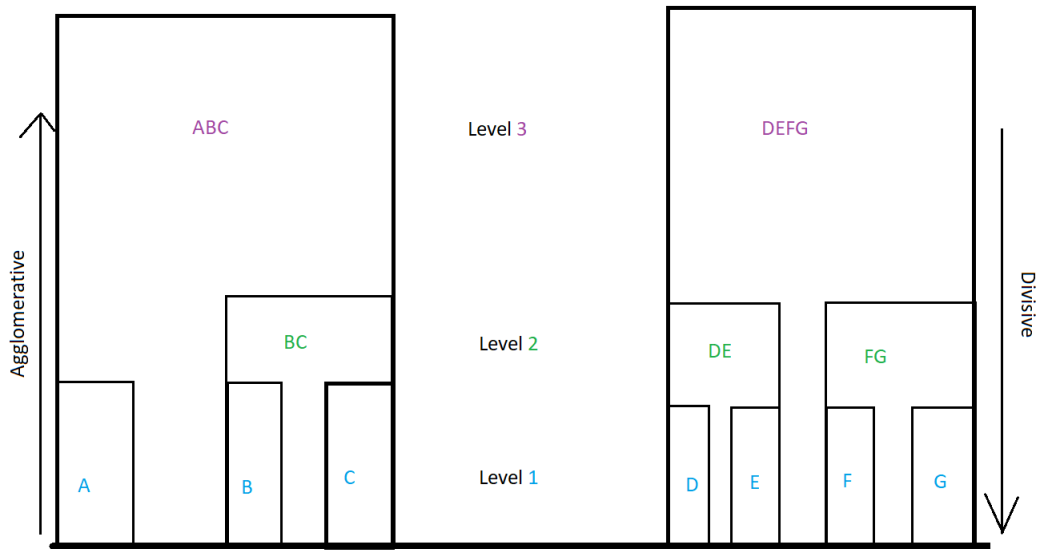


FIGURE 4 – Différents type de fonctionnement d'un algorithme de clustering hiérarchique

La Figure 4 rereprésente les différents types de fonctionnement d'un algorithme de clustering hiérarchique. La méthode agglomerative part du bas vers le haut, c'est-à-dire que le départ se fait en partant du fait que chaque donnée est son propre cluster, qui est ensuite fusionner jusqu'à obtention d'un seul cluster. La méthode divisive part du haut vers le bas, c'est-à-dire que le départ se fait en partant du fait que toutes les données sont inclusent dans un seul cluster, puis une succession de division est effectué jusqu'à obtenir un cluster par donnée.

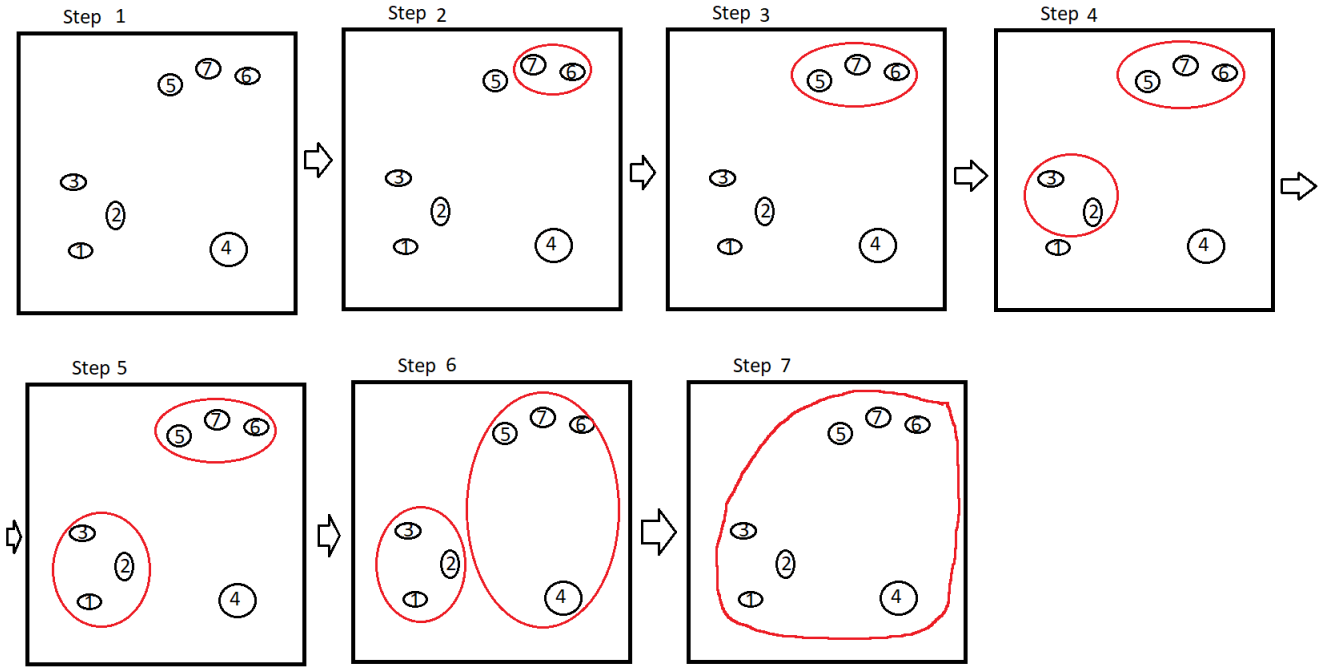


FIGURE 5 – Algorithme de clustering hiérarchique méthode agglomerative

La Figure 5 représente le fonctionnement pas-à-pas de l'algorithme de clustering hiérarchique méthode agglomerative. Les étapes sont les suivantes :

1. Faire de chaque point de données un single - cluster.
2. Prendre les deux points de données suivants les plus proches et en faire un cluster.

3. Encore une fois, prendre les deux points de données suivants les plus proches et en faire un cluster.
4. Répétez l'étape 3 jusqu'à ce qu'il ne reste plus qu'une seule cluster.

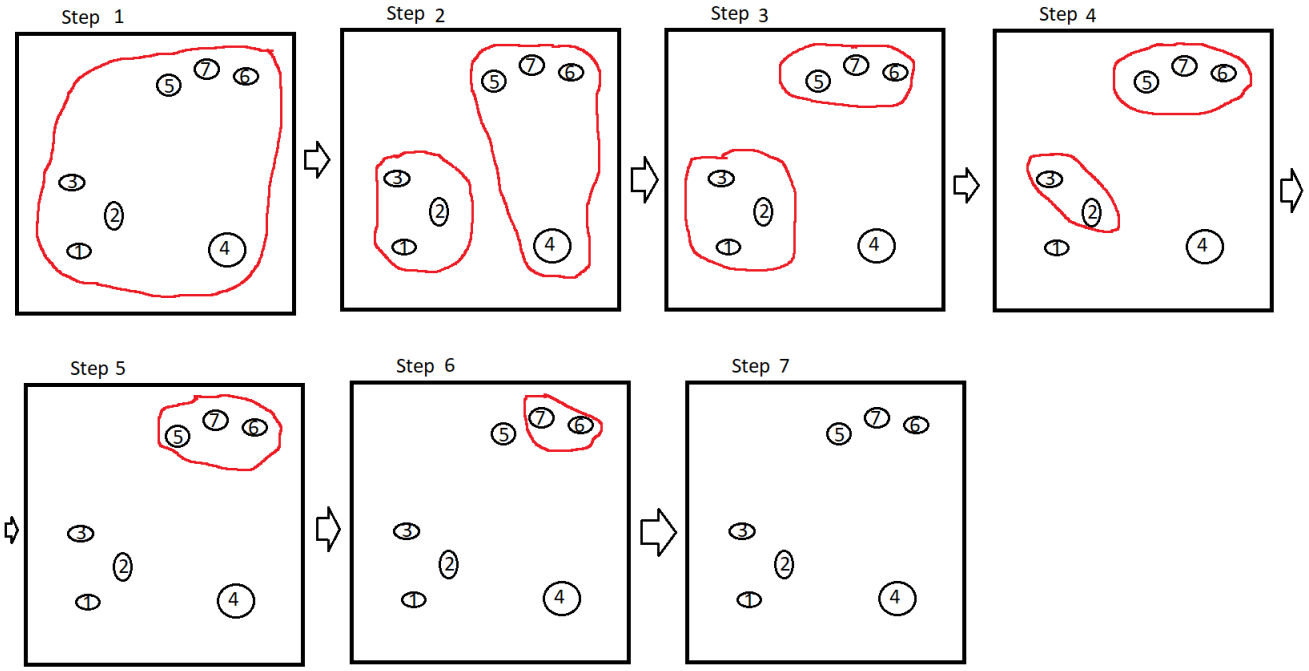


FIGURE 6 – Algorithme de clustering hiérarchique méthode divisive

La Figure 6 représente le fonctionnement pas-à-pas de l'algorithme de clustering hiérarchique méthode divisive. Les étapes sont les suivantes :

1. Mettre tous les objets ou points de l'ensemble de données dans un seul cluster.
2. Partitionner le cluster unique en deux clusters moins similaires.
3. Continuez ce processus pour former les nouveaux clusters jusqu'à ce que le nombre de clusters souhaité soit atteint, c'est-à-dire un cluster pour chaque observation.

Les différentes méthodes les plus utilisées pour calculer les distances entre clusters sont données à la Figure 7.

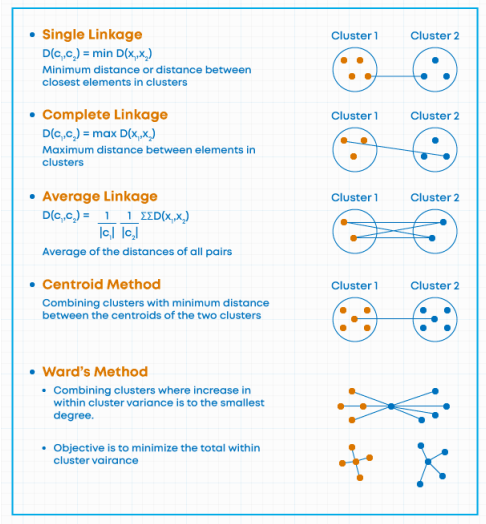


FIGURE 7 – Différentes méthodes de calcul des distances entre clusters

## 3.4 Quels sont les avantages et les inconvénients respectifs de ces deux techniques de clustering ?

K-means :

- Avantages :
  - Facile à comprendre et à mettre en œuvre.
  - Peut facilement s'adapter aux changements.
  - Adapté aux grands jeux de données.
  - Produit des clusters plus serrés, en particulier avec les clusters globulaires.
  - La technique est rapide et efficace en termes de coût de calcul.
- Inconvénients :
  - Décision du nombre de clusters prisent en amont.
  - La manière dont les données sont ordonnées lors de la construction de l'algorithme affecte les résultats finaux de l'ensemble de données.
  - La modification ou la remise à l'échelle de l'ensemble de données, par le biais de la normalisation ou de la standardisation, modifiera complètement les résultats finaux.
  - L'algorithme ne peut être exécuté que sur des données numériques.

CAH :

- Avantages :
  - Facile à comprendre et à mettre en œuvre.
  - Pas besoin de spécifier un nombre particulier de clusters à l'avance.
  - Facilement répétable.
  - Utilisation de dendrogramme pour la visualisation.
- Inconvénients :
  - Ne fonctionne pas bien sûr des grands jeux de données.
  - Une fois la décision prise de combiner deux clusters, elle ne peut être annulée.
  - Sensible aux bruits et outliers.

# Chapitre 4

# Algorithmes de recommandation

## 4.1 Donnez dans ses grandes lignes le principe du filtrage collaboratif

Le filtrage collaboratif repose sur l'adage : Si deux personnes ont aimé des contenus identiques par le passé, elles ont une probabilité élevée d'aimer les mêmes choses dans le futur. Les recommandations personnalisées issues du filtrage collaboratif peuvent être calculées de diverses manières. Notamment en se basant sur le profil des lecteurs (User-based), en utilisant les profiles de contenus (Item-based) ou encore en faisant de la factorisation de matrice.

## 4.2 Pourquoi préfère-t-on le filtrage collaboratif item-item au filtrage individu-individu ?

Le filtrage individu-individu (user-user) sur l'utilisateur revient à dire que "les personnes qui vous ressemblent ont aimé l'article X. Le filtrage item-item sur l'utilisateur revient à dire que les "personnes ayant acheté/consulté un objet X ont également acheté/consulté l'objet Y".

- Le filtrage collaboratif item-item peut être plus facile à mettre en œuvre que le filtrage individu-individu, car il nécessite moins de données sur les préférences des utilisateurs. Pour utiliser le filtrage individu-individu, il faut disposer d'une matrice de préférences complète pour chaque utilisateur, ce qui peut être difficile à obtenir et à gérer à grande échelle.
- Le filtrage collaboratif item-item peut être plus rapide à exécuter que le filtrage individu-individu, car il nécessite moins de calculs. Dans le filtrage individu-individu, il faut comparer les préférences de chaque utilisateur avec celles de tous les autres utilisateurs, ce qui peut être coûteux en termes de temps de calcul.
- Le filtrage collaboratif item-item peut fournir des recommandations plus précises que le filtrage individu-individu dans certaines situations.

# Table des figures