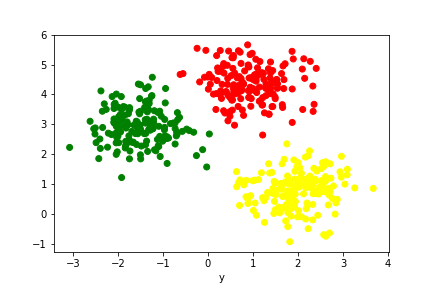


Rapport Projet 4ème année

Machine Learning

**Netflix Clustering**





Professeur référent : Victor Rabiet

**Sommaire**

1. **Introduction**

De nos jours les entreprises produisent une quantité astronomique de données et être capable de les analyser et les valoriser représente un enjeu crucial et un avantage compétitif indéniable. C’est pourquoi le machine learning et le métier de Data Scientist vont jouer un rôle important dans le monde que nous connaissons aujourd’hui. Le machine learning (ou apprentissage automatique en français) est l’étude d’algorithmes informatiques qui s’améliorent automatiquement par l’expérience, il est considéré comme un sous-ensemble de l’intelligence artificielle (*Wikipédia*). Le Data Scientist aura pour but d’utiliser des algorithmes de machine learning afin de créer par exemple des moteurs de recommandations, des prédictions pour son entreprise ou encore des intelligences artificielles.

Il existe deux types d’algorithme de machine learning : *l’apprentissage supervisée* et***l’apprentissage non-supervisée****.* La différence entre ces deux types d’algorithme est très simple. Dans le cas de l’apprentissage supervisée, le Data Scientist va guider l’algorithme en lui fournissant des exemples. L’algorithme va alors « apprendre » de chaque exemple en ajustant ses paramètres afin de limiter les erreurs que ce dernier pourrait générer. Concernant l’apprentissage non-supervisée, on ne va pas fournir d’exemples, l’apprentissage va se faire de façon autonome.

Il est important de signaler que le type d’algorithme utilisé dépend de la situation dans laquelle on se trouve, c’est-à-dire, de la base de données et de ce que l’on cherche à déterminer.

Le projet consiste à travailler sur une base de données contenant des données de la célèbre plateforme de streaming **Netflix**. Nous allons à partir de cette base de données nous intéresser à un problème connu d’apprentissage non-supervisée : **le** **clustering**. Avec le clustering et les données de Netflix, il sera possible de **prédire les films** que les utilisateurs vont préférer.

Mais qu’est-ce que le clustering ?

1. **Clustering et KMeans**
   1. **Définition**

Le clustering est une méthode d'analyse statistique utilisée pour organiser des données brutes en ce que l’on nomme des « clusters ». Les clusters vont regroupés des données qui possèdent des similarités entre elles. Pour avoir un bon clustering, l’algorithme utilisée va minimiser l’inertie intra (intérieur) de chaque cluster et maximiser l’inertie inter (extérieur) entre chaque cluster. En d’autres mots, plus les données d’un même cluster sont proches les uns des autres mais éloignées des données des autres clusters, meilleur sera le clustering effectuer.

Le clustering sert donc principalement à segmenter ou classifier une base de données. Il est utilisé par exemple pour structurer les mots clés d'un site. Pour appliquer le clustering on va utiliser un algorithme très répandu dans le monde de la Data Science : KMeans.

* 1. **L’algorithme de clustering : KMeans**

KMeans permet de regrouper en K clusters distincts les données en notre possession. Chaque donnée appartiendra à un cluster et ne pourra appartenir à deux clusters différents. Pour regrouper des données dans un cluster on va se servir de la notion de similarité entre ces dernières. Cette notion va se traduire par une distance dite « distance de dissimilarité ». Pour calculer cette distance on a plusieurs choix mais dans notre projet on s’est contenté d’utiliser les plus distances les plus utilisées.

* Distance euclidienne : La distance euclidienne est la distance classique que l’on connait depuis le collège. Soit une matrice X à n variables quantitatives. Dans l’espace vectoriel En. La distance euclidienne d entre deux observations x1 et x2 se calcule comme suit :



* La distance de Manhattan (taxi-distance) : est la distance entre deux points parcourus par un taxi lorsqu’il se déplace dans une ville où les rues sont agencées selon un réseau ou un quadrillage. Entre deux points A et B, de coordonnées respectives (XA, YA) et (XB, YB), la distance de Manhattan est définie par :



* Distance cosinus :

KMeans est un algorithme dit itératif ce qui signifie que c’est un algorithme qui va effectuer une boucle. Dans l’algorithme que nous allons utiliser la boucle aura pour but de minimiser la distance entre chaque individu et le centroïd. Le centroïd est un point qui sert de représentant entre chaque cluster. La fin de l’algorithme est atteinte lorsque que l’on a convergé vers la distance minimum entre chaque donnée et son centroïd qui lui a été affecté.

Voyons maintenant l’algorithme.

**Entrée :**

* K le nombre de cluster à former
* Base de données (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)

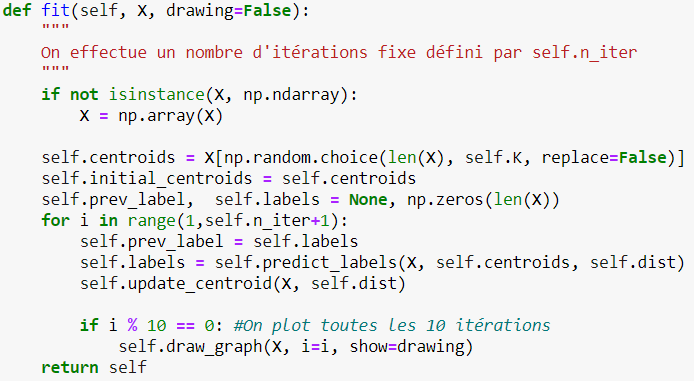
**FIN ALGORITHME**

* 1. **Implémentation de l’algorithme**

Une fois les bases assimilées nous avons implémenté l’algorithme en utilisant le langage informatique python. Pour ce faire nous nous sommes aidés de différents packages :

* pandas : permet de manipuler facilement des données à analyser.
* numpy : permet d’effectuer des calculs numériques avec une certaine facilité sur des tableaux de nombres pouvant atteindre de grandes dimensions.
* Scipy : pour le traitement d’images
* Matplotlib.pyplot : permet d’afficher des graphes.

La première étape établie a été de créer une classe Points sur laquelle nous effectueront les tests. Ensuite nous avons créé la classe KMeans qui contient, comme son nom l’indique, l’algorithme Kmeans. Cette classe est composée de la fonction fit(), c’est cette fonction qui réalise le clustering sur les données qu’on lui passe en paramètre.



Comme nous pouvons le voir, la fonction fit() fait appel à plusieurs autres fonctions que nous avons créé.

**Bibliographie**

* *Différence entre apprentissage supervisé et apprentissage non supervisé,* **ActuIA,** [www.actuia.com/vulgarisation/difference-entre-apprentissage-supervise-apprentissage-non-supervise/](http://www.actuia.com/vulgarisation/difference-entre-apprentissage-supervise-apprentissage-non-supervise/)
* *A quoi sert le clustering des données ?,* **JDN**, [www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203345-clustering-definition/](https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203345-clustering-definition/)
* *Tout ce que vous voulez savoir sur l’algorithme K-Means,* **Mr.Mint**, <https://mrmint.fr/algorithme-k-means>