Exercice K-Means

1- INTRODUCTION

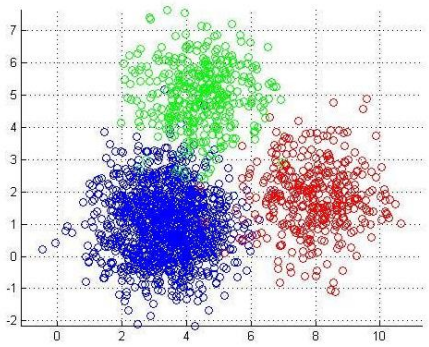
Il existe trois grands types d'apprentissage en Machine Learning : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement (cfr cours théorique).

Avec l'exercice du module 2 concernant l'attrition chez Orange, nous avons créé un algorithme d'apprentissage supervisé : une classification « maison ».

Dans ce module 3, nous allons travailler sur un algorithme très connu dans l'apprentissage non supervisé : l'algorithme des K-Means. Cet algorithme permet de réaliser des classifications non supervisées appelées segmentation ou clustering. Bien sûr, il existe beaucoup d'autres algorithmes de classification non supervisée mais les principes de base du K-Means sont assez simples à comprendre (cfr exercice ci-dessous).

1-1. Qu'est-ce que la segmentation (clustering) ?

La segmentation est une technique d'analyse des données qui consiste à regrouper un ensemble de données en sous-ensembles homogènes appelés clusters. L'objectif principal est donc de partitionner un ensemble de données de telle manière que les observations au sein d'un même groupe ont plus de ressemblance entre elles qu'avec celles des autres groupes. Cela permet de découvrir des structures dans des sous-groupes de données et de regrouper des éléments qui partagent des caractéristiques communes.



KDNuggets

1-2. Exemples d'utilisation du clustering

1-2.1 Segmentation client

Dans le secteur commercial, les entreprises font appel à ce genre d'algorithmes pour regrouper les clients en segments homogènes en fonction de leurs habitudes d'achats, leurs préférences ou leur comportement en ligne... Cela peut notamment aider à personnaliser les campagnes marketing de l'entreprise.

1-2.2 Compression d'images

K-Means peut être utilisé pour compresser des images en réduisant le nombre de couleurs utilisées. En regroupant les couleurs similaires, on peut réduire la taille de l'image tout en préservant les caractéristiques visuelles importantes.

1-2.3 Systèmes de recommandation

Cet algorithme peut être utilisé pour les systèmes de recommandation.

Par exemple, Bob lit un texte : on peut lui proposer un texte du même groupe donc calculé comme similaire ou un texte lu par des personnes similaires (càd regroupées avec) à Bob.

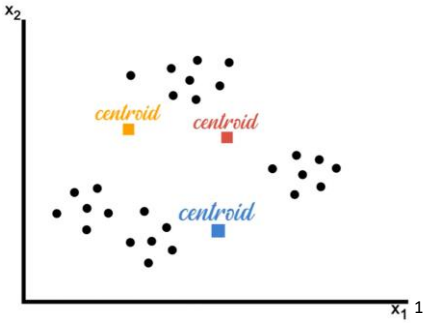
1-2.4 Analyse de texte

Dans le traitement du langage naturel, cet algorithme peut être appliqué pour regrouper des documents de texte similaires. Cela peut être utile pour organiser de grandes collections de documents, identifier des thèmes récurrents ou détecter du plagiat.

1-2.5 Principes du K-Means

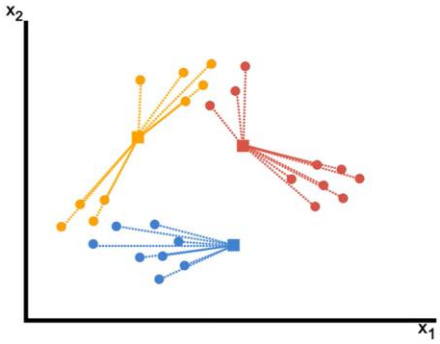
Etape 1 : Initialisation

Sur base du paramètre d'entrée k = nombre de clusters souhaité (nous n'entrons pas dans les détails du choix de k, ci-dessous k=3), l'algorithme choisit aléatoirement un certain nombre de points comme centres initiaux des clusters. Ils sont appelés centroïdes.



Etape 2 : Affectation

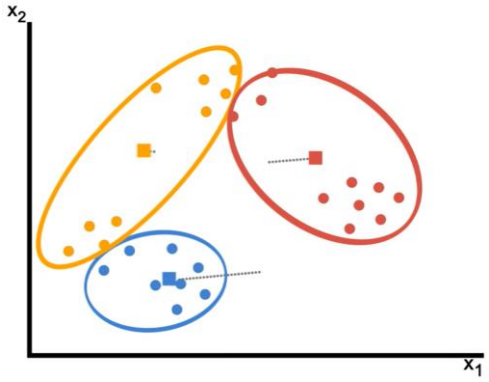
Chaque observation est assignée au cluster dont le centroïde est le plus proche (le plus ressemblant).



Etape 3 : Mise à jour

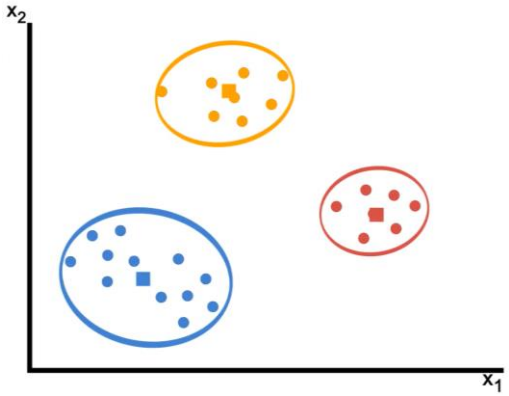
Les centroïdes sont mis à jour pour qu'ils se trouvent au centre de leur cluster (moyenne des points).

[1] Guillaume Saint-Cirgue (Machine Learnia) : https://www.youtube.com/watch?v=FTtzd31IAOw&t=355s



Etape 4 : Répétition

Le processus est itératif donc les étapes d'affectation et de mise à jour des centroïdes sont répétées jusqu'à ce que les centroïdes atteignent une position d'équilibre (donc clusters stables).



Parfois, le choix initial des centroïdes peut être problématique mais nous ne développerons pas cet aspect dans ce cours.

2- EXERCICES

2-1. Ex1

Ecrivez le diagramme d'actions correspondant à cet algorithme.

2-2. Ex2

Dans le cadre de son engagement envers la qualité de ses services, Orange envisage d'implanter trois nouvelles antennes GSM en Wallonie. L'objectif principal de cette initiative est d'optimiser la couverture réseau, en particulier pour ses 20 meilleurs clients dans la région. Pour garantir une localisation stratégique des antennes, on vous demande d'examiner les coordonnées GPS de leurs clients (coordonnées GPS schématisées) les plus importants, disponibles dans la feuille « 20clients » du fichier Excel et de proposer une implantation optimale des antennes.

2-2.1 Marche à suivre :

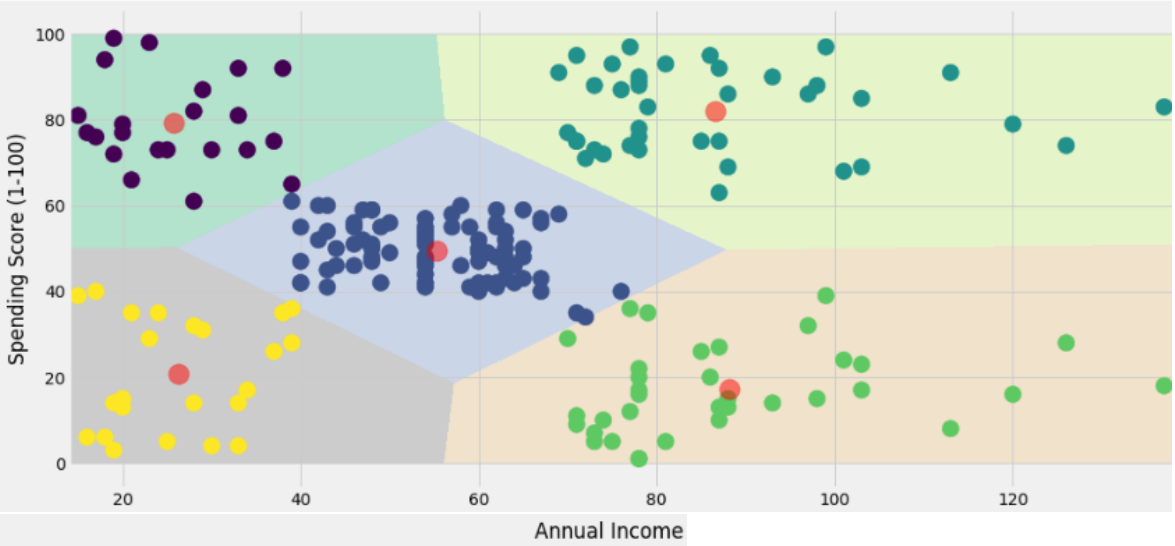
- Placez les 20 points correspondant aux coordonnées des clients (cfr feuille « 20clients » du fichier « K-Means.xlsx ») sur une feuille quadrillée ;
- Attribuez une couleur à chaque centroïde ;

- A chaque itération, colorez chaque point dans la couleur de son cluster et entourez les clusters ;
- Répétez le processus pour chaque itération.

Cette représentation visuelle permettra de mieux comprendre la convergence de l'algorithme des K-Means.

2-3. Ex3

- Orange souhaite aussi réaliser « une cartographie » de ses clients en liant leurs revenus annuels (en milliers d'euros) et leur score (score entre 1 et 100 attribué en fonction de leur comportement et de la nature de leurs dépenses). Dans la feuille « 200clients » sont répertoriées les observations concernant 200 clients. Réalisez un K-Means avec 5 clusters.
- Graphique de la solution : interprétez les clusters.



2-4. Ex 4

Feuille 200Clients+, réalisez K-Means avec 5 clusters sur les observations complètes.