



Travail encadré de recherche :

Profilage des consommateurs à l'aide des k-means++ semi-supervisés



Par :

Destin ASHUZA CIRUMANGA
Elvina GOVENDASAMY

Master 1 Mathématiques et Applications
Parcours Data Science

Sous la direction de Mme Eunice OKOME OBIANG

2019-2020

Description

La segmentation des populations est un problème rencontré dans de nombreuses disciplines. En marketing par exemple, cette technique permet aux entreprises d'identifier et de caractériser les différents profils de consommateurs, afin de leur offrir des produits et des services adaptés à leurs besoins.

Le clustering par l'algorithme des k-means est la méthode d'apprentissage non supervisé la plus utilisée pour répondre à ce type de problématique. Son objectif est de segmenter la population en k groupes disjoints sur base d'un critère de similarité. Toutefois, la qualité de ce modèle varie selon le choix des centres initiaux des clusters. L'idée des k-means++ semi supervisés est de contrôler l'initialisation des centres de clusters en intégrant des données supervisées dans le processus d'apprentissage pour ainsi améliorer les performances du clustering en termes de coût et de durée d'exécution.

Dans ce travail, nous nous intéressons dans un premier temps à la présentation de différents résultats théoriques sur les différents algorithmes des k-means et à la démonstration d'un critère de qualité associé aux k-means++ semi-supervisés. Dans un second temps, nous programmons et testons l'algorithme des k-means++ semi-supervisés sur un jeu de données clients.

Table des matières

Introduction	5
I Partie théorique	5
I.1 Les k-means standards	5
Présentation de l'algorithme	5
Convergence	6
I.2 Les k-means++	7
Présentation de l'algorithme	7
Critère de qualité	8
I.3 Les k-means++ semi-supervisés	15
Présentation de l'algorithme	15
Critère de qualité	16
I.4 Résultats annexes utilisés dans cette partie	22
Inégalité de Cauchy-Schwarz	22
Inégalité des moyennes	23
Majoration des sommes partielles de la série harmonique	23
II Partie appliquée	24
II.1 Choix de la valeur optimale pour k	25
II.2 Expérimentations	26
Références	28

Introduction

A l'ère du big data et de l'intelligence artificielle, le recours aux outils d'analyse et d'exploitation des données est de plus en plus fréquent, et ce dans tous les secteurs désormais. L'apprentissage statistique - machine learning en anglais - est la science au cœur de l'exploitation de ces données et de l'automatisation des méthodes dévolues à cette fin. Il existe différents types d'apprentissage, chacun ayant ses points forts et ses inconvénients. La distinction la plus courante se fait entre l'apprentissage supervisé et l'apprentissage non supervisé qui regroupent la plupart des problèmes. Et un mariage entre ces deux types donne l'apprentissage semi-supervisé.

Dans l'apprentissage supervisé, l'on dispose de p variables notées X mesurées sur n individus et d'une variable réponse notée Y . L'objectif est de trouver une relation entre les variables explicatives X permettant d'approcher au mieux la vraie réponse Y . Les avantages de cet apprentissage sont entre autres des résultats facilement interprétables, de meilleures performances prédictives ou encore des modèles facilement comparables. Cependant l'acquisition de l'information a un coût. En effet, l'information contenue dans la variable réponse nécessite une mobilisation des moyens pour son obtention. Nous pouvons penser à des questionnaires d'enquêtes ou l'embauche du personnel supplémentaire pour s'en occuper. De plus, dans la catégorisation des individus, il n'est pas impossible que l'humain y insère un biais pour telle ou telle autre raison.

Quant à l'apprentissage non supervisé, la variable réponse n'est plus disponible. Le but sera alors d'identifier des patterns ou des relations entre les variables, de regrouper les individus présentant des similitudes (clustering) ou encore de mettre en évidence les variables les plus significatives. L'inconvénient majeur de ce type d'apprentissage est la difficulté à interpréter les modèles ou à évaluer leurs performances. Mais ils ont pour avantages une facilité de mise en œuvre, des coûts réduits pour l'acquisition de l'information et une possible correction du biais qui serait dû aux à priori sur la catégorisation des individus.

Pour ce qui est de l'apprentissage semi-supervisé, il est obtenu comme un jumelage de deux types précédents : l'on dispose de l'information Y juste pour une partie des données. Dès lors, ses avantages sont un mélange entre les avantages du supervisé et du non supervisé : la correction du biais sur les catégories, des performances significatives par au rapport au non supervisé et augmentant avec la quantité des données supervisées, l'acquisition de l'information moins coûteuse par rapport au supervisé. Toutefois, dans certains cas, l'interprétation du modèle pose toujours problème.

L'algorithme des k-means qui fait l'objet de ce travail est un algorithme simple et très utilisé relevant de l'apprentissage non supervisé. Dans une première partie théorique de ce travail, nous présentons cet algorithme et son principe de fonctionnement ; ensuite nous présentons sa version améliorée, les k-means++, en démontrant un critère de qualité qui lui est associé ; et enfin nous étendons cette dernière version au cas où l'on dispose de certaines données qui sont labellisées, ce qui donne l'algorithme des k-means++ semi-supervisés. Dans une deuxième partie appliquée, nous implémentons ces algorithmes grâce au langage python et nous les testons sur une base réelle des données clients. Il en résulte que les k-means++ semi-supervisés sont de loin plus performants.

I Partie théorique

Cette partie est consacrée à la présentation et à l'étude théorique de différents algorithmes de k-means. Les résultats énoncés et démontrés sont essentiellement tirés de [1] et de [3]. Ils sont suffisamment détaillés pour être facilement accessibles à toute personne ayant des bases en probabilités et en statistique.

I.1 Les k-means standards

Présentation de l'algorithme

L'algorithme des k-means est utilisé pour résoudre le problème suivant : étant donné un entier naturel k fixé et un ensemble de points $X \subset \mathbb{R}^d$ (avec $d \geq 1$), trouver un ensemble de centres $C = \{c_i \in \mathbb{R}^d, i = 1, 2, \dots, k\}$ tel que

$$C = \underset{\{A \subset \mathbb{R}^d, \text{card}(A)=k\}}{\operatorname{argmin}} \sum_{x \in X} \min_{c \in A} \|x - c\|^2 \quad (1)$$

Par la suite, en utilisant ces centres, le label de chaque élément x est donné par :

$$\ell(x) = \underset{1 \leq i \leq k}{\operatorname{argmin}} \|x - c_i\|$$

Résoudre l'équation (1) de manière exacte pour trouver C est "NP-hard" ; c'est-à-dire qu'il est difficile de trouver une solution exacte avec un algorithme de complexité polynomiale. Ainsi, en pratique, l'on se contente d'approcher la solution de manière locale. L'algorithme de Lloyd est le plus utilisé pour cette fin. Il prend en entrée l'ensemble des données X , les centres initiaux C et retourne les centres finaux résolvant localement l'équation (1). Concrètement, on l'implémente comme suit :

Algorithme de Lloyd

Entrées : X (le dataset) et C (l'ensemble des k centres initiaux)

Sorties : C (l'ensemble des centres définitifs)

1 faire

2 associer chaque $x_i \in X$ au centre le plus proche $c(x_i) \in C$

3 remplacer chaque $c_j \in C$ par le point moyen des $x \in X$ tels que $c(x) = c_j$

4 jusqu'à ce que C ne change plus

5 sortir C

L'algorithme des k-means standards est alors obtenu en exécutant l'algorithme de Lloyd avec les centres initiaux qui sont choisis uniformément dans X

Convergence

Après avoir compris le fonctionnement de l'algorithme de Lloyd (et donc celui des k-means standards qui en découle), une question légitime se pose : l'algorithme s'arrête-t-il ?

Avant de répondre à cette question, introduisons d'abord quelques définitions et notations qui nous seront utiles dans la suite de ce travail.

Définition 1. (Clustering)

Un clustering C est un ensemble de centres qui sont utilisés pour déterminer le label de chaque point de notre jeu de données X .

Définition 2. (Fonction potentielle)

Étant donné un clustering C , la fonction potentielle ϕ associée à C est la fonction à valeurs dans \mathbb{R}^+ définie dans l'ensemble des parties de X par :

$$\forall A \subset X, \quad \phi(A) = \sum_{a \in A} \min_{c \in C} \|a - c\|^2$$

Par la suite, par abus de notation, l'on écrira tout simplement $\phi(X) = \phi$ et uniquement pour X .

Définition 3. (Clustering optimal et cluster optimal)

Soit C_{OPT} le clustering qui serait la solution exacte du problème des k -means tel que formulé par l'équation (1) à la page 6. C_{OPT} sera alors appelé clustering optimal et tout cluster dans C_{OPT} sera appelé cluster optimal.

Définition 4. (Poids D^2)

Étant donné un clustering C , le poids D^2 est défini par :

$$\forall x \in X, \quad D^2(x) = \phi(\{x\}) = \min_{c \in C} \|x - c\|^2$$

C'est donc le carré de la distance de x au centre dans C qui lui est le plus proche.

A la lumière de ces définitions, nous pouvons résumer le problème de k -means comme suit :

- Le problème de k -means tel que formulé par l'équation (1) à la page 6 consiste à trouver le clustering qui minimise la fonction potentielle ϕ
- Théoriquement, on sait qu'il existe une solution exacte C_{OPT} à ce problème mais en pratique il est difficile de la calculer.
- L'algorithme des k -means fait donc appel à l'algorithme de Lloyd (donné à la page 6) pour trouver un minimum local de ϕ atteint en une solution approchée donnée par un clustering C .

Nous pouvons maintenant énoncer le résultat qui nous permet d'affirmer que l'algorithme s'arrête :

Lemme 1. Soient S un ensemble de points, $c(S)$ son point moyen (son centre de gravité, son isobarycentre) et z un point quelconque.

Alors, on a :

$$\sum_{x \in S} \|x - z\|^2 - \sum_{x \in S} \|x - c(S)\|^2 = \text{card}(S) \cdot \|c(S) - z\|^2$$

où $\text{card}(S)$ désigne le cardinal de S .

Démonstration.

Nous avons :

$$\begin{aligned}
\sum_{x \in S} \|x - z\|^2 &= \sum_{x \in S} \|(x - c(S)) + (c(S) - z)\|^2 \\
&= \sum_{x \in S} \|x - c(S)\|^2 + 2 \cdot \sum_{x \in S} \langle x - c(S), c(S) - z \rangle + \sum_{x \in S} \|c(S) - z\|^2 \\
&= \sum_{x \in S} \|x - c(S)\|^2 + 2 \cdot \langle \sum_{x \in S} x - \text{card}(S) \cdot c(S), c(S) - z \rangle + \text{card}(S) \cdot \|c(S) - z\|^2 \\
&= \sum_{x \in S} \|x - c(S)\|^2 + 2 \cdot \text{card}(S) \cdot \langle c(S) - c(S), c(S) - z \rangle + \text{card}(S) \cdot \|c(S) - z\|^2 \\
&= \sum_{x \in S} \|x - c(S)\|^2 + 2 \cdot \text{card}(S) \cdot \langle 0, c(S) - z \rangle + \text{card}(S) \cdot \|c(S) - z\|^2 \\
&= \sum_{x \in S} \|x - c(S)\|^2 + \text{card}(S) \cdot \|c(S) - z\|^2
\end{aligned}$$

d'où le résultat. □

Le lemme 1 nous permet d'affirmer que remplacer le centre de chaque cluster par le point moyen de ses éléments à chaque tour de boucle de l'algorithme de Llyod permet de décroître nécessairement ϕ . Ainsi dès que ϕ atteint un minimum local, l'algorithme s'arrête.

Nous pourrions être content de savoir que l'algorithme s'arrête toujours. Malheureusement rien ne garantit que la solution locale obtenue donne un bon clustering. En effet, ϕ peut être arbitrairement grand malgré l'optimum local atteint. C'est le défaut principal de l'algorithme des k-means standards. Dans la section suivante, nous présentons les k-means++ qui résolvent ce problème.

I.2 Les k-means++

Présentation de l'algorithme

L'algorithme des k-means++ est obtenu en changeant uniquement la manière dont sont initialisés les centres initiaux qui seront utilisés dans l'algorithme de Llyod. Au lieu de les choisir uniformément, ces centres initiaux sont choisis avec une probabilité proportionnelle au poids D^2 . Ensuite le reste se passe exactement comme pour les k-means standards en faisant appel à l'algorithme de Llyod. Voici l'algorithme d'initialisation en question :

Algorithme d'initialisation des centres pour les k-means++

Entrées : X (le dataset) et k (le nombre de centres)

Sorties : C (l'ensemble des centres initiaux)

1 choisir un $x \in X$ suivant la loi uniforme

2 initialiser $C = \{x\}$

3 **Tant que** $\text{card}(C) \leq k$ **faire**

4 choisir un $x \in X$ avec probabilité proportionnelle à $D^2(x)$

5 actualiser $C = C \cup \{x\}$

6 sortir C

Critère de qualité

Dans cette section, nous présentons les résultats qui montrent que l'algorithme des k-means++ est meilleur que celui des k-means standards dans le sens où l'espérance de la fonction potentielle est majorée. Plus exactement, le théorème 1 présenté à la fin de cette section montre que la fonction potentielle obtenue à fin de l'étape de l'initialisation de l'algorithme des k-means++ est un $O(\ln k)$ par rapport au clustering optimal théorique. Et cela est suffisant car dans l'algorithme de Lloyd, cette fonction ne peut faire que baisser.

Nous présentons trois lemmes qui permettent de démontrer le théorème en question : le premier permet de contrôler ϕ pour le choix du premier centre ; le deuxième montre que ϕ reste toujours majorée lorsque l'on choisit les centres restants dans les clusters de optimaux théoriques proportionnellement au poids D^2 . Enfin le troisième lemme généralise cela pour n'importe quel choix des centres, ce qui conduit à la démonstration du théorème.

Lemme 2. *Soient A un cluster arbitraire dans C_{OPT} et $c(A)$ son point moyen. Soit C un clustering avec un unique centre choisi aléatoirement dans A et de fonction potentielle ϕ .*

Alors, on a :

$$\mathbb{E}[\phi(A)] = 2 \cdot \phi_{OPT}(A)$$

Démonstration.

Supposons que $C = \{a_0\}$ avec a_0 qui a été choisi aléatoirement dans A ; c'est-à-dire :

$$\mathbb{P}(\{\text{choisir } a_0\}) = \frac{1}{\text{card}(A)}$$

En faisant varier le choix de a_0 , nous voyons clairement que $\phi(A)$ est une variable aléatoire discrète à valeurs dans \mathbb{R} . En effet, $\text{card}(\phi(A)(\Omega)) = \text{card}(A)$ et pour chaque choix de a_0 , la valeur prise par $\phi(A)$ est donnée par :

$$\begin{aligned} \phi_{a_0}(A) &= \phi(A; C) \\ &= \phi(A; \{a_0\}) \\ &= \sum_{a \in A} \min_{c \in \{a_0\}} \|a - c\|^2 \\ &= \sum_{a \in A} \|a - a_0\|^2 \end{aligned}$$

De plus, par le lemme 1 énoncé à la page 7, nous savons que le point moyen $c(A)$ du cluster A est nécessairement le centre utilisé pour ce cluster dans le clustering optimal C_{OPT} sinon ϕ_{OPT} ne serait pas minimale. De tout ce qui précède, nous pouvons alors

écrire :

$$\begin{aligned}
\mathbb{E}[\phi(A)] &= \sum_{a_0 \in A} \phi_{a_0}(A) \cdot \mathbb{P}(\phi(A) = \phi_{a_0}(A)) \\
&= \sum_{a_0 \in A} \left(\sum_{a \in A} \|a - a_0\|^2 \cdot \frac{1}{\text{card}(A)} \right) \\
&= \frac{1}{\text{card}(A)} \cdot \sum_{a_0 \in A} \sum_{a \in A} \|a - a_0\|^2
\end{aligned}$$

En appliquant le lemme 1 avec $S = A$ et $z = a_0$, nous obtenons :

$$\begin{aligned}
\mathbb{E}[\phi(A)] &= \frac{1}{\text{card}(A)} \sum_{a_0 \in A} \left(\text{card}(A) \cdot \|c(A) - a_0\|^2 + \sum_{a \in A} \|a - c(A)\|^2 \right) \\
&= \sum_{a_0 \in A} \|c(A) - a_0\|^2 + \frac{1}{\text{card}(A)} \cdot \sum_{a_0 \in A} \sum_{a \in A} \|a - c(A)\|^2 \\
&= \sum_{a_0 \in A} \|c(A) - a_0\|^2 + \frac{1}{\text{card}(A)} \cdot \text{card}(A) \cdot \sum_{a \in A} \|a - c(A)\|^2 \\
&= 2 \cdot \sum_{a \in A} \|c(A) - a\|^2 \\
&= 2 \cdot \phi_{OPT}(A)
\end{aligned}$$

ce qui conclut la preuve. □

Lemme 3. Soient A un cluster arbitraire dans C_{OPT} et C un clustering quelconque. Supposons que l'on ajoute aléatoirement un centre à C choisi dans A avec une probabilité proportionnelle à D^2 , la distance euclidienne au carré. Soit ϕ la fonction potentielle du clustering ainsi obtenu. Alors, on a :

$$\mathbb{E}[\phi(A)] \leq 8 \cdot \phi_{OPT}(A)$$

Démonstration.

Soit $1 \leq l < k$ (ici k représente le nombre de centres utilisés pour le clustering optimal C_{OPT}).

Supposons que $C = \{c_1, c_2, \dots, c_l\}$ et considérons un cluster A dans C_{OPT} .

Rappelons que la distance D^2 associée au clustering C vérifie :

$$\forall a \in A, D^2(a) = \min_{c \in C} \|a - c\|^2$$

Soit $a_0 \in A$ choisi avec probabilité $\frac{D^2(a_0)}{\sum_{a \in A} D^2(a)}$

Notons $C(a_0) = C \cup \{a_0\}$ et ϕ_{a_0} respectivement le clustering et la fonction potentielle

obtenus pour ce choix de a_0 . Nous pouvons alors écrire :

$$\begin{aligned}
\mathbb{E}[\phi(A)] &= \sum_{a_0 \in A} \phi_{a_0}(A) \cdot \mathbb{P}(\phi(A) = \phi_{a_0}(A)) \\
&= \sum_{a_0 \in A} \sum_{a \in A} \min_{c \in C(a_0)} \|a - c\|^2 \cdot \frac{D^2(a_0)}{\sum_{a \in A} D^2(a)} \\
&= \sum_{a_0 \in A} \sum_{a \in A} \min \left(\min_{c \in C} \|a - c\|^2, \|a - a_0\|^2 \right) \cdot \frac{D^2(a_0)}{\sum_{a \in A} D^2(a)} \\
&= \sum_{a_0 \in A} \sum_{a \in A} \min(D^2(a), \|a - a_0\|^2) \cdot \frac{D^2(a_0)}{\sum_{a \in A} D^2(a)}
\end{aligned}$$

De plus pour tout $a \in A$, nous avons :

$$\begin{aligned}
D(a_0) &= \min_{c \in C} \|a_0 - c\| \\
&\leq \|a_0 - c\| = \|(a_0 - a) + (a - c)\| \quad \forall c \in C \\
&\leq \|a_0 - a\| + \|a - c\| \quad \forall c \in C \quad (\text{par inégalité triangulaire})
\end{aligned}$$

En particulier pour $c_j \in C$ tel que $D(a) = \|a - c_j\|$, nous avons :

$$D(a_0) \leq \|a_0 - a\| + D(a)$$

D'une part, en élevant au carré les membres de cette inégalité, nous obtenons :

$$D^2(a_0) \leq \|a_0 - a\|^2 + D^2(a) + 2 \cdot \|a_0 - a\| \cdot D(a)$$

D'autre part,

$$\begin{aligned}
\left(\|a_0 - a\| - D(a) \right)^2 &\geq 0 \\
\Leftrightarrow \|a_0 - a\|^2 + D^2(a) &\geq 2 \cdot \|a_0 - a\| \cdot D(a)
\end{aligned}$$

Ainsi, nous en déduisons que

$$\forall a \in A, \quad D^2(a_0) \leq 2 \cdot \|a_0 - a\|^2 + 2 \cdot D^2(a)$$

En sommant sur toutes les valeurs de a , nous obtenons alors

$$D^2(a_0) \leq \frac{2}{\text{card}(A)} \cdot \sum_{a \in A} \|a_0 - a\|^2 + \frac{2}{\text{card}(A)} \cdot \sum_{a \in A} D^2(a)$$

Nous utilisons ensuite cette inégalité pour majorer $D^2(a_0)$ dans l'expression de $\mathbb{E}[\phi(A)]$ obtenue précédemment, ce qui conduit à

$$\begin{aligned}
\mathbb{E}[\phi(A)] &\leq \frac{2}{\text{card}(A)} \cdot \sum_{a_0 \in A} \sum_{a \in A} \min(D^2(a), \|a - a_0\|^2) \cdot \frac{\sum_{a \in A} \|a_0 - a\|^2}{\sum_{a \in A} D^2(a)} \\
&\quad + \frac{2}{\text{card}(A)} \cdot \sum_{a_0 \in A} \sum_{a \in A} \min(D^2(a), \|a - a_0\|^2) \cdot \frac{\sum_{a \in A} D^2(a)}{\sum_{a \in A} D^2(a)}
\end{aligned}$$

Dans le second membre de cette inégalité, nous utilisons le fait que

$$\min(D^2(a), \|a - a_0\|^2) \leq D^2(a)$$

pour le premier bloc, et

$$\min(D^2(a), \|a - a_0\|^2) \leq \|a - a_0\|^2$$

pour le second bloc, nous obtenons :

$$\begin{aligned} \mathbb{E}[\phi(A)] &\leq \frac{2}{\text{card}(A)} \cdot \sum_{a_0 \in A} \sum_{a \in A} \|a_0 - a\|^2 \\ &\quad + \frac{2}{\text{card}(A)} \cdot \sum_{a_0 \in A} \sum_{a \in A} \|a - a_0\|^2 \end{aligned}$$

C'est-à-dire

$$\mathbb{E}[\phi(A)] \leq 4 \cdot \frac{1}{\text{card}(A)} \cdot \sum_{a_0 \in A} \sum_{a \in A} \|a_0 - a\|^2$$

L'expression $\frac{1}{\text{card}(A)} \cdot \sum_{a_0 \in A} \sum_{a \in A} \|a_0 - a\|^2$ est celle que nous avons obtenue pour l'espérance lors de la démonstration du lemme 2 à la page 8. Ainsi, en appliquant le lemme 2, nous obtenons donc

$$\mathbb{E}[\phi(A)] \leq 8 \cdot \phi_{OPT}(A)$$

ce qui est le résultat escompté. □

Lemme 4. Soient C un clustering arbitraire de fonction potentielle ϕ et $u > 0$ un entier naturel. Choisissons u clusters non couverts par C dans C_{OPT} . Notons X_u l'ensemble des points de ces clusters et $X_c = X - X_u$ son complémentaire. Supposons que nous ajoutons $t \leq u$ centres à C choisis de manière aléatoire proportionnellement au poids D^2 . Notons C' le nouveau clustering ainsi obtenu et ϕ' sa fonction potentielle associée. Alors, on a :

$$\mathbb{E}[\phi'] \leq \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_t) + \frac{u-t}{u} \cdot \phi(X_u);$$

avec $H_t = \sum_{i=1}^t \frac{1}{i}$ la $t^{\text{ème}}$ somme partielle de la série harmonique.

Démonstration. Elle se fait par une récurrence à deux niveaux en montrant que si l'inégalité est vérifiée pour les couples d'entiers $(t-1, u)$ et $(t-1, u-1)$, alors elle l'est également pour le couple (t, u) .

Regardons d'abord de plus près la fonction ϕ' et sa relation avec ϕ :
Notons c'_1, c'_2, \dots, c'_t les t centres ajoutés à C . Pour toute partie $A \subset X$, on a alors :

$$\begin{aligned} \phi'(A) &= \sum_{x \in A} \min_{c' \in C'} \|x - c'\|^2 \\ &= \sum_{x \in A} \min \left(D^2(x), \|x - c'_1\|^2, \dots, \|x - c'_t\|^2 \right) \\ &\leq \sum_{x \in A} D^2(x) = \phi(A). \end{aligned}$$

Passons à la démonstration proprement dite maintenant :

- Initialisation : il suffit de vérifier pour $t = 0, u > 0$ et $t = 1, u = 1$.

Pour $t = 0, u > 0$ on a :

$$\begin{aligned}\mathbb{E}[\phi'] &= \phi \quad (\text{car } \phi' = \phi \text{ qui est constante}) \\ &= \phi(X_c) + \phi(X_u) \\ &= \left(\phi(X_c) \right) \cdot (1 + H_0) + \frac{u-0}{u} \cdot \phi(X_u) \quad \text{avec } H_0 = 0 \\ &\leq \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_0) + \frac{u-0}{u} \cdot \phi(X_u)\end{aligned}$$

Pour $t = 1, u = 1$ on distingue deux cas : soit le nouveau centre est choisi dans le seul cluster non couvert, soit il est choisi dans l'un des clusters couverts.

Si le nouveau centre c'_1 est choisi dans le cluster non couvert, il l'est avec probabilité $\frac{\phi(X_u)}{\phi}$ et on a :

$$\begin{aligned}\phi' &= \phi'(X_c) + \phi'(X_u) \\ &\leq \phi(X_c) + \phi'(X_u)\end{aligned}$$

D'où,

$$\begin{aligned}\mathbb{E}[\phi' \mid c'_1 \in X_u] &\leq \phi(X_c) + \mathbb{E}[\phi'(X_u) \mid c'_1 \in X_u] \\ &\leq \phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \quad (\text{par le lemme 3}).\end{aligned}$$

Sinon, le nouveau centre c'_1 est choisi dans X_c avec probabilité $\frac{\phi(X_c)}{\phi}$ et on a :

$$\mathbb{E}[\phi' \mid c'_1 \in X_c] \leq \phi \quad (\text{car } \phi' \leq \phi).$$

Finalement,

$$\begin{aligned}\mathbb{E}[\phi'] &= \mathbb{E}[\phi' \mathbf{1}_{\{c'_1 \in X_u\}}] + \mathbb{E}[\phi' \mathbf{1}_{\{c'_1 \in X_c\}}] \\ &= \mathbb{P}(c'_1 \in X_u) \mathbb{E}[\phi' \mid c'_1 \in X_u] + \mathbb{P}(c'_1 \in X_c) \mathbb{E}[\phi' \mid c'_1 \in X_c] \\ &\leq \frac{\phi(X_u)}{\phi} \left(\phi(X_c) + 8\phi_{OPT}(X_u) \right) + \frac{\phi(X_c)}{\phi} \phi \\ &\leq 2 \cdot \phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \\ &\leq \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_1) + \frac{1-1}{1} \cdot \phi(X_u)\end{aligned}$$

- Hypothèse de récurrence (HR) et hérédité : soient $t \geq 1$ et $u \geq 1$.

Supposons l'inégalité vérifiée pour les couples $(t-1, u)$ et $(t-1, u-1)$ et montrons qu'elle est vraie pour le couple (t, u) également :

Comme précédemment, on considère deux cas suivant le choix du premier centre c'_1 .

1^{er} cas : choix du premier centre dans un cluster inclus dans X_c .

On a :

$$\begin{aligned}\mathbb{E}[\phi' \mathbf{1}_{\{c'_1 \in X_c\}}] &= \mathbb{P}(c'_1 \in X_c) \mathbb{E}[\phi' \mid c'_1 \in X_c] \\ &= \frac{\phi(X_c)}{\phi} \mathbb{E}[\phi' \mid c'_1 \in X_c]\end{aligned}$$

Notons ϕ'' la fonction potentielle associée au clustering $C \cup \{c'_1\}$:

$$\phi'' = \sum_{x \in X} \min \left(D^2(x), \|x - c'_1\|^2 \right) \leq \phi$$

En appliquant l'hypothèse de récurrence avec le couple $(t-1, u)$ car il reste $t-1$ centres à choisir et u clusters toujours non couverts, on obtient :

$$\begin{aligned}\mathbb{E}[\phi' \mid c'_1 \in X_c] &\leq \left(\phi''(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{u-t+1}{u} \cdot \phi''(X_u) \\ &\leq \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{u-t+1}{u} \cdot \phi(X_u)\end{aligned}$$

Ainsi,

$$\mathbb{E}[\phi' \mathbf{1}_{\{c'_1 \in X_c\}}] \leq \frac{\phi(X_c)}{\phi} \left[\left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{u-t+1}{u} \cdot \phi(X_u) \right]$$

2^{ème} cas : choix du premier centre dans un cluster $A \subset X_u$.

On a :

$$\begin{aligned}\mathbb{E}[\phi' \mathbf{1}_{\{c'_1 \in A\}}] &= \mathbb{P}(c'_1 \in A) \mathbb{E}[\phi' \mid c'_1 \in A] \\ &= \frac{\phi(A)}{\phi} \mathbb{E}[\phi' \mid c'_1 \in A]\end{aligned}$$

Notons p_a la probabilité de choisir l'élément $a \in A$ comme premier centre et $\phi_a = \phi''(A)$ où ϕ'' désigne la fonction potentielle associée au clustering $C \cup \{c'_1\}$ comme précédemment mais avec $c'_1 = a$ ici. Alors, on a :

$$\begin{aligned}\mathbb{E}[\phi' \mid c'_1 \in A] &= \sum_{a \in A} \mathbb{P}(c'_1 = a) \mathbb{E}[\phi' \mid c'_1 = a] \\ &= \sum_{a \in A} p_a \mathbb{E}[\phi' \mid c'_1 = a]\end{aligned}$$

En appliquant l'hypothèse de récurrence avec le couple $(t-1, u-1)$ car il reste $t-1$ centres à choisir et $u-1$ clusters non couverts, on obtient :

$$\mathbb{E}[\phi' \mid c'_1 \in A] \leq \sum_{a \in A} p_a \left[\left(\phi''(X_c \cup A) + 8\phi_{OPT}(X_u - A) \right) \cdot (1 + H_{t-1}) + \frac{u-t}{u-1} \cdot \phi''(X_u - A) \right]$$

Comme

$$\begin{aligned}\phi''(X_c \cup A) &= \phi''(X_c) + \phi''(A) \\ &\leq \phi(X_c) + \phi_a,\end{aligned}$$

Et

$$\phi''(X_u - A) + \phi(A) \leq \phi(X_u - A) + \phi(A) = \phi(X_u)$$

$$\Rightarrow \quad \phi''(X_u - A) \leq \phi(X_u) - \phi(A) ,$$

On aboutit à :

$$\begin{aligned} \mathbb{E}[\phi' \mid c'_1 \in A] &\leq \sum_{a \in A} p_a \left[\left(\phi(X_c) + \phi_a + 8 \cdot \phi_{OPT}(X_u) - 8 \cdot \phi_{OPT}(A) \right) \cdot (1 + H_{t-1}) \right. \\ &\quad \left. + \frac{u-t}{u-1} \cdot \left(\phi(X_u) - \phi(A) \right) \right] \\ &\leq \left(\phi(X_c) + \sum_{a \in A} p_a \phi_a + 8 \cdot \phi_{OPT}(X_u) - 8 \cdot \phi_{OPT}(A) \right) \cdot (1 + H_{t-1}) \\ &\quad + \frac{u-t}{u-1} \cdot \left(\phi(X_u) - \phi(A) \right) \end{aligned}$$

Par le lemme 3, on sait que :

$$\sum_{a \in A} p_a \phi_a \leq 8 \cdot \phi_{OPT}(A)$$

En appliquant ce résultat, on obtient :

$$\mathbb{E}[\phi' \mid c'_1 \in A] \leq \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{u-t}{u-1} \cdot \left(\phi(X_u) - \phi(A) \right).$$

D'où,

$$\begin{aligned} \mathbb{E}[\phi' \mathbb{1}_{\{c'_1 \in A\}}] &= \frac{\phi(A)}{\phi} \mathbb{E}[\phi' \mid c'_1 \in A] \\ &\leq \frac{\phi(A)}{\phi} \left[\left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{u-t}{u-1} \cdot \left(\phi(X_u) - \phi(A) \right) \right] \end{aligned}$$

Comme

$$\mathbb{E}[\phi' \mathbb{1}_{\{c'_1 \in X_u\}}] = \sum_{A \subset X_u} \mathbb{E}[\phi' \mathbb{1}_{\{c'_1 \in A\}}] ,$$

on obtient :

$$\mathbb{E}[\phi' \mathbb{1}_{\{c'_1 \in X_u\}}] \leq \frac{\phi(X_u)}{\phi} \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{1}{\phi} \frac{u-t}{u-1} \cdot \left(\phi(X_u)^2 - \sum_{A \subset X_u} \phi(A)^2 \right)$$

Or, par l'inégalité des moyennes¹ (moyenne arithmétique et moyenne quadratique), on a :

$$\sum_{A \subset X_u} \phi(A)^2 \geq \frac{1}{u} \left(\sum_{A \subset X_u} \phi(A) \right)^2 = \frac{1}{u} \phi(X_u)^2$$

1. L'énoncé et la démonstration de cette inégalité sont présentés en annexe 2 à la page 23

$$\Rightarrow - \sum_{A \subset X_u} \phi(A)^2 \leq -\frac{1}{u} \phi(X_u)^2$$

Ainsi,

$$\begin{aligned} \mathbb{E} [\phi' \mathbf{1}_{\{c'_1 \in X_u\}}] &\leq \frac{\phi(X_u)}{\phi} \left(\phi(X_c) + 8\phi_{OPT}(X_u) \right) (1 + H_{t-1}) + \frac{1}{\phi} \frac{u-t}{u-1} \left(\phi(X_u)^2 - \frac{1}{u} \phi(X_u)^2 \right) \\ &\leq \frac{\phi(X_u)}{\phi} \left[\left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{u-t}{u} \cdot \phi(X_u) \right] \end{aligned}$$

Finalement, en regroupant les deux cas, on obtient :

$$\begin{aligned} \mathbb{E} [\phi'] &= \mathbb{E} [\phi' \mathbf{1}_{\{c'_1 \in X_c\}}] + \mathbb{E} [\phi' \mathbf{1}_{\{c'_1 \in X_u\}}] \\ &\leq \frac{\phi(X_c)}{\phi} \left[\left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{u-t+1}{u} \cdot \phi(X_u) \right] \\ &\quad + \frac{\phi(X_u)}{\phi} \left[\left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{u-t}{u} \cdot \phi(X_u) \right] \\ &\leq \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{u-t}{u} \cdot \phi(X_u) \\ &\quad + \frac{1}{u} \frac{\phi(X_u)}{u} \phi(X_c) \\ &\leq \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_{t-1}) + \frac{u-t}{u} \cdot \phi(X_u) \\ &\quad + \frac{1}{u} \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \\ &\leq \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot \left(1 + H_{t-1} + \frac{1}{u} \right) + \frac{u-t}{u} \cdot \phi(X_u) \\ &\leq \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_t) + \frac{u-t}{u} \cdot \phi(X_u) \quad \left(\text{car } t \leq u \Rightarrow \frac{1}{u} \leq \frac{1}{t} \right). \end{aligned}$$

L'inégalité est ainsi établie pour le couple (t, u) . \square

Théorème 1. Soient C' un clustering obtenu lors de l'initialisation de l'algorithme des k -means++ et ϕ' sa fonction potentielle associée. Alors on a :

$$\mathbb{E}[\phi'] \leq 8 \cdot (\ln k + 2) \cdot \phi_{OPT}$$

Démonstration.

Considérons le clustering C avec un seul centre obtenu après avoir choisi le premier centre de manière uniforme lors de la première étape de l'algorithme des k -means++ et notons ϕ sa fonction potentielle associée. Soit A le cluster de C_{OPT} dans lequel ce premier centre a été choisi. Le clustering C' de fonction potentielle ϕ' est obtenu alors en ajoutant à C les $k-1$ centres restants. Ainsi, en appliquant le lemme 3 énoncé à la page 9 $t = u = k-1$ et $X_c = A$ (donc $X_u = X - A$), nous obtenons :

$$\mathbb{E}[\phi'] \leq \left(\phi(A) + 8 \cdot \phi_{OPT}(X - A) \right) \cdot (1 + H_{k-1}) + \frac{(k-1) - (k-1)}{k-1} \cdot \phi(X - A)$$

Comme

$$\phi_{OPT}(X - A) = \phi_{OPT}(X) - \phi_{OPT}(A) = \phi_{OPT} - \phi_{OPT}(A) ,$$

l'inégalité équivaut à

$$\mathbb{E}[\phi'] \leq \left(\phi(A) + 8 \cdot \phi_{OPT} - 8 \cdot \phi_{OPT}(A) \right) \cdot (1 + H_{k-1})$$

En passant à l'espérance dans les deux membres, l'inégalité est préservée grâce à la propriété de positivité et nous obtenons :

$$\mathbb{E} \left[\mathbb{E}[\phi'] \right] \leq \mathbb{E} \left[\left(\phi(A) + 8 \cdot \phi_{OPT} - 8 \cdot \phi_{OPT}(A) \right) \cdot (1 + H_{k-1}) \right]$$

Remarquons que $\mathbb{E}[\phi']$ est une constante. De même, dans le second membre de l'inégalité ci-dessus, tous les termes sont toujours constants sauf $\phi(A)$. Ainsi, par les propriétés de l'espérance, cette inégalité revient à :

$$\mathbb{E}[\phi'] \leq \left(\mathbb{E}[\phi(A)] + 8 \cdot \phi_{OPT} - 8 \cdot \phi_{OPT}(A) \right) \cdot (1 + H_{k-1})$$

Par le lemme 2, nous savons que

$$\mathbb{E}[\phi(A)] = 2 \cdot \phi_{OPT}(A)$$

En utilisant ce résultat, nous aboutissons donc à :

$$\begin{aligned} \mathbb{E}[\phi'] &\leq \left(8 \cdot \phi_{OPT} - 6 \cdot \phi_{OPT}(A) \right) \cdot (1 + H_{k-1}) \\ &\leq 8 \cdot \phi_{OPT} \cdot (1 + H_{k-1}) \end{aligned}$$

Finalement, en utilisant l'inégalité² $H_{k-1} < H_k \leq 1 + \ln k$, nous obtenons

$$\mathbb{E}[\phi'] \leq 8 \cdot \phi_{OPT} \cdot (2 + \ln k)$$

ce qui est le résultat attendu. □

I.3 Les k-means++ semi-supervisés

Dans cette section, nous nous intéressons au cas semi supervisé où certaines observations sont déjà labellisées. Nous supposons, par exemple, que les données supervisées sont obtenues de la manière suivante :

1. on choisit un cluster i aléatoirement de manière uniforme
2. on tire g_i observations du cluster i de manière uniforme
3. on affecte à ces g_i observations le label i

Il est possible de répéter ces trois étapes plusieurs fois de façon à avoir plus de données supervisées et plus de clusters concernés par la supervision.

Nous introduisons deux nouvelles notations : $X^{(u)}$ et $X^{(s)}$ désignant respectivement les données non labellisées et les données supervisées (labellisées).

2. La démonstration de cette inégalité entre la série harmonique et le logarithme est rappelée en annexe 3 à la page 23

Présentation de l'algorithme

Ici aussi, seule l'initialisation change pour tenir compte des données supervisées. Le reste de l'algorithme reste le même en recourant à l'algorithme de Lloyd. Voici l'algorithme d'initialisation pour ce cas :

Algorithme d'initialisation des centres pour les k-means++

Entrées : $X^{(u)}$ (données non labellisées), $X^{(s)}$ (données non labellisées) et k (le nombre de centres)

Sorties : C (l'ensemble des centres initiaux)

1 poser n_ℓ le nombre d'observations labellisées avec le label ℓ

2 initialiser $C = \emptyset$

3 **Pour** $\ell = 1, 2, \dots, k$ **faire**

4 **Si** $n_\ell > 0$ **alors**

5 poser c_ℓ égal au point moyen des données ayant ℓ pour label

6 actualiser $C = C \cup \{c_\ell\}$

7 **Tant que** $\text{card}(C) \leq k$ **faire**

8 choisir un $X^{(u)}$ avec probabilité proportionnelle à $D^2(x)$

9 actualiser $C = C \cup \{x\}$

10 **sortir** C

Critère de qualité

En suivant le même cheminement que pour les k-means++, nous démontrons que cette nouvelle manière d'initialiser les centres initiaux lorsqu'il existe des données supervisées permet d'améliorer la borne obtenue pour les k-means++. Ce résultat est énoncé dans le théorème 2 qui vient à la page 21. Les lemmes présentés dans cette section ont pour but de démontrer ledit théorème.

Lemme 5. Soient $d \geq 1$ un entier, $A = \{x_i \in \mathbb{R}^d, i = 1, 2, \dots, n\}$ un ensemble de points et $S \subset A$ un sous ensemble de A de cardinal g choisi aléatoirement de manière uniforme dans l'ensemble des parties de A à g éléments.

Alors, on a :

$$\mathbb{E} \left[\sum_{x_i \in S} x_i \right] = \frac{g}{n} \sum_{x_i \in A} x_i$$

Démonstration.

Posons $X_i = \mathbb{1}_{\{x_i \in S\}}$. C'est une variable aléatoire définie sur A et suivant une loi de Bernoulli de paramètre $\frac{g}{n}$ car le tirage est uniforme. S étant de cardinal g , nous pouvons alors écrire :

$$\begin{aligned} \mathbb{E} \left[\sum_{x_i \in S} x_i \right] &= \mathbb{E} \left[\sum_{i=1}^n x_i \cdot X_i \mid \sum_{i=1}^n X_i = g \right] \\ &= \sum_{i=1}^n x_i \mathbb{E} \left[X_i \mid \sum_{i=1}^n X_i = g \right] \quad (\text{par linéarité de l'espérance}) \end{aligned}$$

Pour tout $1 \leq i \leq n$, la variable $X_i \mid \sum_{i=1}^n X_i = g$ est une variable de Bernoulli également. Calculer son espérance revient donc à déterminer la probabilité qu'elle prenne la valeur 1. Ainsi, nous avons :

$$\begin{aligned}
\mathbb{E} \left[X_i \mid \sum_{i=1}^n X_i = g \right] &= \mathbb{P} \left(X_i = 1 \mid \sum_{i=1}^n X_i = g \right) \\
&= \frac{\mathbb{P} \left(X_i = 1, \sum_{i=1}^n X_i = g \right)}{\mathbb{P} \left(\sum_{i=1}^n X_i = g \right)} \\
&= \frac{\mathbb{P} \left(X_i = 1, \sum_{j=1, j \neq i}^n X_j = g - 1 \right)}{\mathbb{P} \left(\sum_{i=1}^n X_i = g \right)} \\
&= \frac{\mathbb{P}(X_i = 1) \cdot \mathbb{P} \left(\sum_{j=1, j \neq i}^n X_j = g - 1 \right)}{\mathbb{P} \left(\sum_{i=1}^n X_i = g \right)} \quad (\text{par indépendance}) \\
&= \frac{\frac{g}{n} \cdot \binom{n-1}{g-1} \cdot \left(\frac{g}{n}\right)^{g-1} \cdot \left(1 - \frac{g}{n}\right)^{n-g}}{\binom{n}{g} \cdot \left(\frac{g}{n}\right)^g \cdot \left(1 - \frac{g}{n}\right)^{n-g}} \\
&= \frac{\binom{n-1}{g-1}}{\binom{n}{g}} \\
&= \frac{g}{n}
\end{aligned}$$

Il en résulte alors

$$\mathbb{E} \left[\sum_{x_i \in S} x_i \right] = \sum_{i=1}^n x_i \cdot \frac{g}{n}$$

ce qui conclut la preuve. □

Lemme 6. Soient $d \geq 1$ un entier, $A = \{x_i \in \mathbb{R}^d, i = 1, 2, \dots, n\}$ un ensemble de points et $S \subset A$ un sous ensemble de A de cardinal g choisi aléatoirement de manière uniforme dans l'ensemble des parties de A à g éléments.

Alors, on a :

$$\mathbb{E} \left[\left(\sum_{x_i \in S} x_i \right)^\top \left(\sum_{x_i \in S} x_i \right) \right] = \frac{g(g-1)}{n(n-1)} \sum_{i \neq j} x_i^\top x_j + \frac{g}{n} \sum_{i=1}^n x_i^\top x_i$$

Démonstration.

Les hypothèses étant les mêmes que celles du lemme 5, la démonstration est du même type et utilise les mêmes arguments.

Posons $X_i = \mathbb{1}_{\{x_i \in S\}}$. C'est une variable aléatoire définie sur A et suivant une loi de Bernoulli de paramètre $\frac{g}{n}$ car le tirage est uniforme. S étant de cardinal g , nous pouvons alors écrire :

$$\begin{aligned} \mathbb{E} \left[\left(\sum_{x_i \in S} x_i \right)^\top \left(\sum_{x_i \in S} x_i \right) \right] &= \mathbb{E} \left[\left(\sum_{i=1}^n x_i \cdot X_i \right)^\top \left(\sum_{j=1}^n x_j \cdot X_j \right) \mid \sum_{l=1}^n X_l = g \right] \\ &= \mathbb{E} \left[\sum_{1 \leq i, j \leq n} x_i^\top x_j X_i X_j \mid \sum_{l=1}^n X_l = g \right] \\ &= \sum_{1 \leq i, j \leq n} x_i^\top x_j \mathbb{E} \left[X_i X_j \mid \sum_{l=1}^n X_l = g \right] \end{aligned}$$

Pour tous $1 \leq i, j \leq n$, la variable $X_i X_j \mid \sum_{l=1}^n X_l = g$ est une variable de Bernoulli également. Calculer son espérance revient donc à déterminer la probabilité qu'elle prenne la valeur 1. Nous distinguons deux cas : le cas où $i = j$ et le cas où $i \neq j$.

Si $i = j$, comme $X_i^2 = X_i$, nous nous retrouvons avec la même espérance qui a été calculée dans la démonstration du lemme 5 et qui vaut donc $\frac{g}{n}$

Si $i \neq j$, nous avons :

$$\begin{aligned}
\mathbb{E} \left[X_i X_j \mid \sum_{l=1}^n X_l = g \right] &= \mathbb{P} \left(X_i X_j = 1 \mid \sum_{l=1}^n X_l = g \right) \\
&= \frac{\mathbb{P} \left(X_i = 1, X_j = 1, \sum_{l=1}^n X_l = g \right)}{\mathbb{P} \left(\sum_{l=1}^n X_l = g \right)} \\
&= \frac{\mathbb{P} \left(X_i = 1, X_j = 1, \sum_{l=1, l \neq i, j}^n X_l = g - 2 \right)}{\mathbb{P} \left(\sum_{l=1}^n X_l = g \right)} \\
&= \frac{\mathbb{P}(X_i = 1) \mathbb{P}(X_j = 1) \cdot \mathbb{P} \left(\sum_{l=1, l \neq i, j}^n X_l = g - 2 \right)}{\mathbb{P} \left(\sum_{l=1}^n X_l = g \right)} \\
&= \frac{\left(\frac{g}{n}\right)^2 \cdot \binom{n-2}{g-2} \cdot \left(\frac{g}{n}\right)^{g-2} \cdot \left(1 - \frac{g}{n}\right)^{n-g}}{\binom{n}{g} \cdot \left(\frac{g}{n}\right)^g \cdot \left(1 - \frac{g}{n}\right)^{n-g}} \\
&= \frac{\binom{n-2}{g-2}}{\binom{n}{g}} \\
&= \frac{g(g-1)}{n(n-1)}
\end{aligned}$$

En combinant les deux cas nous obtenons alors

$$\begin{aligned}
\mathbb{E} \left[\left(\sum_{x_i \in S} x_i \right)^\top \left(\sum_{x_i \in S} x_i \right) \right] &= \sum_{i \neq j} x_i^\top x_j \mathbb{E} \left[X_i X_j \mid \sum_{l=1}^n X_l = g \right] + \sum_{i=1}^n x_i^\top x_i \mathbb{E} \left[X_i X_j \mid \sum_{l=1}^n X_l = g \right] \\
&= \sum_{i \neq j} x_i^\top x_j \frac{g(g-1)}{n(n-1)} + \sum_{i=1}^n x_i^\top x_i \frac{g}{n}
\end{aligned}$$

ce qui donne le résultat attendu. □

Lemme 7. Soient $d \geq 1$ un entier, $A = \{x_i \in \mathbb{R}^d, i = 1, 2, \dots, n\}$ un cluster dans C_{OPT} et $S \subset A$ un sous ensemble de A de cardinal g choisi aléatoirement de manière uniforme dans l'ensemble des parties de A à g éléments.

Alors, on a :

$$\mathbb{E} [\phi(A; \bar{a}_S)] = \left(1 + \frac{n-g}{g(n-1)}\right) \phi_{OPT}(A)$$

avec

$$\phi_{OPT}(A) = \sum_{a_i \in A} \|a_i - c(A)\|^2 \quad ,$$

$c(A)$ le point moyen de A et \bar{a}_S le point moyen de S

Démonstration.

Nous avons :

$$\begin{aligned} \mathbb{E} [\phi(A; \bar{a}_S)] &= \mathbb{E} \left[\sum_{a \in A} \|a - \bar{a}_S\|^2 \right] \\ &= \mathbb{E} \left[\sum_{a \in A} \|a - c(A)\|^2 + n \|c(A) - \bar{a}_S\|^2 \right] \quad (\text{par le lemme 1}) \\ &= \sum_{a \in A} \|a - c(A)\|^2 + n \mathbb{E} [\|\bar{a}_S - c(A)\|^2] \\ &= \phi_{OPT}(A) + n \mathbb{E} [\|\bar{a}_S - c(A)\|^2] \end{aligned}$$

Calculons donc $\mathbb{E} [\|\bar{a}_S - c(A)\|^2]$:

Nous avons

$$\begin{aligned} \|\bar{a}_S - c(A)\|^2 &= \langle \bar{a}_S - c(A), \bar{a}_S - c(A) \rangle \\ &= \langle \bar{a}_S, \bar{a}_S \rangle - 2 \langle c(A), \bar{a}_S \rangle + \langle c(A), c(A) \rangle \\ &= \bar{a}_S^\top \bar{a}_S - 2 \langle c(A), \frac{1}{g} \sum_{a \in S} a \rangle + c(A)^\top c(A) \\ &= \bar{a}_S^\top \bar{a}_S - \frac{2}{g} c(A)^\top \sum_{a \in S} a + c(A)^\top c(A) \end{aligned}$$

En passant à l'espérance, nous obtenons :

$$\begin{aligned}
\mathbb{E} [\|\bar{a}_S - c(A)\|^2] &= \mathbb{E} [\bar{a}_S^\top \bar{a}_S] - \frac{2}{g} c(A)^\top \mathbb{E} \left[\sum_{a \in S} a \right] + c(A)^\top c(A) \\
&= \mathbb{E} [\bar{a}_S^\top \bar{a}_S] - \frac{2}{g} c(A)^\top \frac{g}{n} \sum_{a \in A} a + c(A)^\top c(A) \quad (\text{par le lemme 5}) \\
&= \mathbb{E} [\bar{a}_S^\top \bar{a}_S] - 2c(A)^\top \frac{1}{n} \sum_{a \in A} a + c(A)^\top c(A) \\
&= \mathbb{E} [\bar{a}_S^\top \bar{a}_S] - 2c(A)^\top c(A) + c(A)^\top c(A) \\
&= \mathbb{E} \left[\frac{1}{g} \left(\sum_{a \in S} a \right)^\top \frac{1}{g} \left(\sum_{a \in S} a \right) \right] - c(A)^\top c(A) \\
&= \frac{1}{g^2} \left(\frac{g(g-1)}{n(n-1)} \sum_{i \neq j} a_i^\top a_j + \frac{g}{n} \sum_{i=1}^n a_i^\top a_i \right) - c(A)^\top c(A) \quad (\text{par le lemme 6}) \\
&= \frac{g-1}{gn(n-1)} \sum_{i \neq j} a_i^\top a_j + \frac{1}{gn} \sum_{i=1}^n a_i^\top a_i - c(A)^\top c(A) \\
&= \frac{g-1}{gn(n-1)} \sum_{i,j} a_i^\top a_j - \frac{g-1}{gn(n-1)} \sum_{i=j} a_i^\top a_j + \frac{(n-1)}{gn(n-1)} \sum_{i=1}^n a_i^\top a_i - c(A)^\top c(A) \\
&= \frac{g-1}{gn(n-1)} \left(\sum_{a_i \in A} a_i \right)^\top \left(\sum_{a_j \in A} a_j \right) + \frac{n-g}{gn(n-1)} \sum_{i=1}^n a_i^\top a_i - c(A)^\top c(A) \\
&= n^2 \frac{g-1}{gn(n-1)} c(A)^\top c(A) + \frac{n-g}{gn(n-1)} \sum_{i=1}^n a_i^\top a_i - c(A)^\top c(A) \\
&= -\frac{n-g}{g(n-1)} c(A)^\top c(A) + \frac{n-g}{gn(n-1)} \sum_{i=1}^n a_i^\top a_i \\
&= \frac{n-g}{g(n-1)} \left(\frac{1}{n} \sum_{i=1}^n a_i^\top a_i - c(A)^\top c(A) \right) \\
&= \frac{n-g}{gn(n-1)} \left(\sum_{i=1}^n \|a_i\|^2 - n \|c(A)\|^2 \right) \\
&= \frac{n-g}{gn(n-1)} \left(\sum_{i=1}^n \|a_i - c(A)\|^2 \right) \quad (\text{en appliquant le lemme 1 avec } z=0) \\
&= \frac{n-g}{gn(n-1)} \phi_{OPT}(A)
\end{aligned}$$

et le résultat est démontré ! □

Énonçons maintenant la version du lemme 4 pour le cas semi - supervisé et qui est utilisé dans la démonstration du théorème 2 à la page 21 sur la majoration de l'espérance de la fonction potentielle lorsqu'il y a des données labellisées :

Lemme 8. Soient C un clustering arbitraire et ϕ sa fonction potentielle associée. Supposons qu'il y a u clusters non couverts par C dans C_{OPT} (avec $u \geq 0$ entier). Notons X_u l'ensemble des points de ces clusters et $X_c = X - X_u$ (son complémentaire) l'ensemble des points dans les clusters couverts. Supposons que nous ajoutons $t \leq u$ nouveaux centres à C (en excluant les données labellisées) choisis de manière aléatoire proportionnellement au poids D^2 . Notons C' le nouveau clustering ainsi obtenu et ϕ' sa fonction potentielle associée. Alors, on a :

$$\mathbb{E}[\phi'] \leq \left(\phi(X_c) + 8 \cdot \phi_{OPT}(X_u) \right) \cdot (1 + H_t) + \frac{u-t}{u} \cdot \phi(X_u);$$

avec $H_t = \sum_{i=1}^t \frac{1}{i}$ la $t^{\text{ème}}$ somme partielle de la série harmonique.

Démonstration. Elle est similaire à celle du lemme 4 présentée à la page 11. Il suffit juste d'utiliser les probabilités tenant compte de nouvelles hypothèses ajoutées pour le cas semi-supervisé : étant donné un ensemble A , la probabilité de choisir un nouveau centre a dans cet ensemble proportionnellement au poids D^2 tout en veillant à ne pas choisir a parmi les données labellisées est alors donnée par :

$$\frac{\phi(A \cap X^{(u)})}{\phi(X^{(u)})}$$

où $X^{(u)}$ désigne l'ensemble des données non labellisées.

De plus, en remarquant que toute donnée labellisée est considérée comme faisant partie d'un cluster couvert, nous avons alors

$$X^{(s)} \cap X_u = \emptyset$$

où $X^{(s)}$ désigne l'ensemble des données labellisées (supervisées).

Ainsi, en tenant compte de ces remarques et en suivant le cheminement de la démonstration présentée pour le lemme 4, nous obtenons le résultat annoncé. \square

Finalement, voici le résultat final de cette partie :

Théorème 2. Supposons que la supervision a été faite suivant le procédé décrit au début de cette section. Posons $C = \emptyset$. Pour chaque label ℓ pour lequel il y a des observations supervisées, nous ajoutons à C le point moyen c_ℓ de toutes ces observations. A la fin de ces ajouts, supposons que $\text{card}(C) = G$. Notons n_{ℓ_j} le nombre d'observations affectées du label ℓ_j et n_j le cardinal du cluster associé au label ℓ_j dans C_{OPT} pour $j = 1, 2, \dots, G$. Il reste donc $u = k - G$ clusters non couverts par C dans C_{OPT} . Ajoutons alors $t = u$ nouveaux centres à C pour former C' en utilisant le poids D^2 et en ne prenant pas aucune observation labellisée.

La fonction potentielle ϕ' associée au clustering C' ainsi obtenu vérifie alors

$$\mathbb{E}[\phi'] \leq 8 \cdot (2 + \ln(k - G)) \cdot \phi_{OPT}$$

Démonstration.

Nous allons appliquer le lemme 8 avec $u = t = k - G$. Pour cela, remarquons d'abord que

$$X_c = \bigcup_{j=1}^G X_{\ell_j}$$

où X_{ℓ_j} désigne de le cluster de C_{OPT} correspondant au label ℓ_j . Ainsi,

$$\phi(X_c) = \sum_{j=1}^G \phi(X_{\ell_j})$$

et

$$\begin{aligned} \phi_{OPT}(X_u) &= \phi_{OPT}(X - X_c) \\ &= \phi_{OPT}(X) - \sum_{j=1}^G \phi_{OPT}(X_{\ell_j}) \\ &= \phi_{OPT} - \sum_{j=1}^G \phi_{OPT}(X_{\ell_j}) \end{aligned}$$

Compte tenu de ces constants, le résultat du lemme 8 s'écrit alors :

$$\mathbb{E}[\phi'] \leq \left(\sum_{j=1}^G \left(\phi(X_{\ell_j}) - 8\phi_{OPT}(X_{\ell_j}) \right) + 8\phi_{OPT} \right) \cdot (1 + H_{k-G})$$

Et en passant à l'espérance dans le deuxième membre tout en gardant à l'esprit que tous les termes sont des constantes sauf $\phi(X_{\ell_j})$ et en appliquant le lemme 7 énoncé à la page 19 avec $A = X_{\ell_j}$ (S correspondant ici à l'ensemble des observations de X_{ℓ_j} qui sont labellisées avec $\bar{a}_S = c_{\ell_j}$) pour tout X_{ℓ_j} , nous obtenons :

$$\begin{aligned} \mathbb{E}[\phi'] &\leq \left(\sum_{i=1}^G \left(\left(1 + \frac{n_j - n_{\ell_j}}{n_{\ell_j}(n_j - 1)} \right) \phi_{OPT}(X_{\ell_j}) - 8\phi_{OPT}(X_{\ell_j}) \right) + 8\phi_{OPT} \right) \cdot (1 + H_{k-G}) \\ &\leq \left(\sum_{i=1}^G \left(\left(-7 + \frac{n_j - n_{\ell_j}}{n_{\ell_j}(n_j - 1)} \right) \phi_{OPT}(X_{\ell_j}) \right) + 8\phi_{OPT} \right) \cdot (1 + H_{k-G}) \\ &\leq 8\phi_{OPT} \cdot (1 + H_{k-G}) \end{aligned}$$

la dernière inégalité étant due au fait que pour tout $1 \leq j \leq G$, nous avons

$$n_j - n_{\ell_j} \leq n_j - 1$$

ce qui implique que la somme sur les j intervenant dans le second membre de l'avant dernière inégalité est toujours négative.

Pour conclure, nous utilisons l'inégalité³ $H_{k-G} \leq 1 + \ln(k - G)$ et nous obtenons

$$\mathbb{E}[\phi'] \leq 8 \cdot \phi_{OPT} \cdot \left(2 + \ln(k - G) \right)$$

ce qui est le résultat escompté. □

3. Voir l'annexe 3 à la page 23 pour la démonstration de cette inégalité

L'amélioration de la borne ne semble pas très importante comparée à la borne obtenue pour les k-means++. Cependant, plus il y a des données supervisées, plus cet algorithme sera plus performant que celui des k-means++.

I.4 Résultats annexes utilisés dans cette partie

Dans cette section, nous présentons (pour information) trois résultats qui ont été utilisés indirectement ou directement dans les démonstrations effectuées précédemment.

Inégalité de Cauchy-Schwarz

Annexe 1. (Inégalité de Cauchy-Schwarz dans \mathbb{R}^n)

On considère l'espace euclidien \mathbb{R}^n . Alors, on a :

$$\forall x \in \mathbb{R}^n, \forall y \in \mathbb{R}^n, \langle x, y \rangle \leq \|x\| \|y\|.$$

Démonstration. Soient $x, y \in \mathbb{R}^n$ et $t \in \mathbb{R}$. On a :

$$0 \leq \|x + ty\|^2 = \|y\|^2 t^2 + 2\langle x, y \rangle t + \|x\|^2$$

Le discriminant de ce polynôme du second degré est donné par :

$$\Delta = 4\langle x, y \rangle^2 - 4\|x\|^2 \|y\|^2$$

Comme ce polynôme est de signe constant, il admet au plus une racine réelle et donc son discriminant est négatif ou nul :

$$\begin{aligned} \Delta \leq 0 &\Leftrightarrow \langle x, y \rangle^2 - \|x\|^2 \|y\|^2 \leq 0 \\ &\Leftrightarrow |\langle x, y \rangle| \leq \|x\| \|y\| ; \end{aligned}$$

d'où le résultat escompté. □

Inégalité des moyennes

Annexe 2. (Inégalité des moyennes : arithmétique et quadratique)

Soient $u \geq 1$ un entier et a_1, a_2, \dots, a_u des réels. Alors :

$$\frac{1}{u} \sum_{i=1}^u a_i \leq \sqrt{\frac{1}{u} \sum_{i=1}^u a_i^2}.$$

Démonstration. C'est une application immédiate de l'inégalité de Cauchy-Schwarz dans \mathbb{R}^u (démontrée en annexe 1) en prenant $x = (a_1, a_2, \dots, a_u)$ et $y = (\frac{1}{u}, \frac{1}{u}, \dots, \frac{1}{u})$. □

Majoration des sommes partielles de la série harmonique

Annexe 3. (Majoration de la série harmonique par le logarithme)

Pour tout $k \geq 1$ entier, on pose $H_k = \sum_{i=1}^k \frac{1}{i}$ la $k^{\text{ème}}$ somme partielle de la série harmonique. Alors, on a :

$$\forall k \geq 1, H_k \leq 1 + \ln k$$

Démonstration.

Pour $k = 1$, l'inégalité est trivialement vérifiée.

Supposons alors $k \geq 2$. Pour tout $2 \leq i \leq k$, nous nous intéressons alors à la fonction $x \mapsto \frac{1}{x}$ restreinte à l'intervalle $[i-1, i]$. Cette fonction est continue et strictement décroissante. Ainsi,

$$\forall x \in [i-1, i], \quad \frac{1}{i} \leq \frac{1}{x}$$

Et par positivité de l'intégrale,

$$\frac{1}{i} \leq \int_{i-1}^i \frac{1}{x} dx$$

Et en sommant sur toutes les valeurs de i et en utilisant la relation de Chasles, nous obtenons :

$$1 + \sum_{i=2}^k \frac{1}{i} \leq 1 + \int_1^k \frac{1}{x} dx$$

Soit

$$H_k \leq 1 + \ln k$$

Ainsi, l'inégalité est bien établie. □

II Partie appliquée

Dans cette section, nous allons tester et comparer les différents algorithmes de Kmeans, Kmeans++ et Kmeans++ semi-supervisées sur un jeu de données clients. Nos mesures de performances sont le coût tel qu'estimé par la fonction potentiel, ainsi que la durée d'exécution de l'algorithme.

La Base de Données

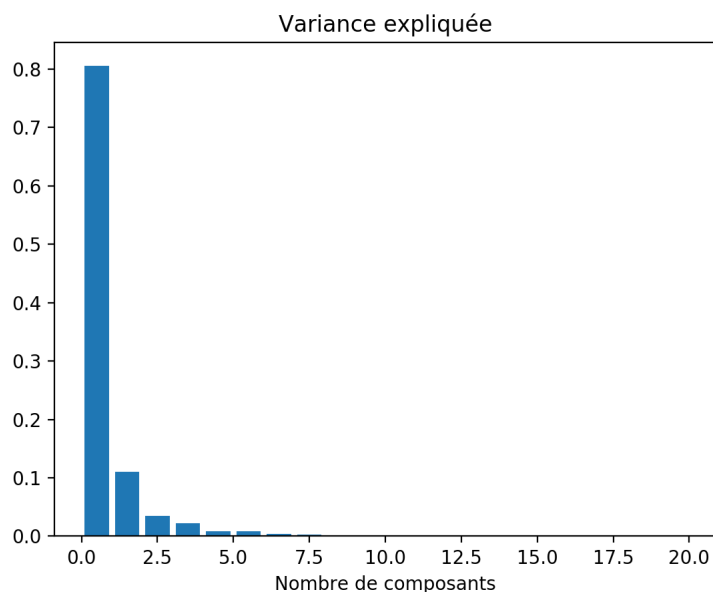
La base de données contient 60366 observations (clients) et 39 variables quantitatives et qualitatives. Tous les clients sont anonymisés se qui limitera l'interprétation de l'analyse. Nous commençons par analyser la base de données. Ci-dessous, nous affichons la statistique d'une partie de nos données.

Index	_1_distance_to_shc	_2_distance_to_shc	_3_distance_to_shc	roducts_purchase	e_products_purch	mount_purchase	avg_purchase	avg_price	shops_used	distance_shop_1	distance_shop_2	distance_shop_3	distance_shop_4	distance_shop_5
count	60366	60366	60366	60366	60366	60366	60366	60366	60366	60366	60366	60366	60366	60366
mean	2030.23	1396.43	2942.67	1778.7	330.668	4235.46	8.53208	3.66651	2.38005	2496.63	2488.24	1924.97	2882.8	2020.9
std	1119.9	1048.41	1327.52	2185.04	236.002	5006.74	10.3154	9.13317	1.01218	1281.53	1417.36	1157.75	1730.09	1260.85
min	6.64247	6.64247	6.64247	1	1	0.212	0.212	0.212	1	93.2834	11.1904	17.8443	6.64247	25.4607
25%	1250.24	573.611	2117.22	227	127	653.118	4.53291	2.53375	2	1609.58	1501.9	1253.84	1532.68	1173.41
50%	1936.88	1184.94	2869.2	925	304	2355.14	6.86278	2.94426	2	2288.67	2355.4	1746.17	2704.32	1852.93
75%	2569.71	1962.38	3580.67	2551.75	500	6054.46	10.4295	3.55811	3	3144.06	3370.72	2323.56	4056	2466.56
max	9004.16	9004.16	9267.7	22131	1465	51588.7	787.569	787.569	5	8019.92	9004.16	7395.25	9273.69	7465.81

Nous voyons qu'il existe une seule variable qualitative, la variable 'shops used', les autres étant toutes des variables quantitatives continues. Nous remarquons aussi que toutes les variables quantitatives affichent de grandes dispersions. Par exemple, en regardant le

nombre de produits achetés, nous voyons que certains clients achètent très peu, soit 1 seul article, tandis que d'autres achètent jusqu'à 22 131 articles, la moyenne étant 2552.75. Ce type d'observation nous permettra de mieux interpréter les clusters proposés par les différentes versions des k-means.

Une Analyse en Composantes Principales est implémentée afin de mieux regrouper et comprendre les variables quantitatives. Le choix du nombre de composantes principales est basé sur le pourcentage de la variance expliquée ainsi que par la méthode du coude (une baisse soudaine de la variance expliquée entre deux composantes).



Nous voyons ici que les premières composantes principales expliquent une bonne partie de l'inertie, soit plus de 80% de la variance, et ainsi nous nous baserons uniquement sur l'analyse en composantes principales sur 2 et 3 dimensions.

Sur un plan 1-2 nous obtenons le graphique tel qu'affiché ci-dessous :

AFFICHER ACP SUR LE PLAN 1-2 (DIMENSION 2)

AFFICHER ACP SUR 3 DIMENSIONS

II.1 Choix de la valeur optimale pour k

La méthode K-means est une méthode d'apprentissage non-supervisée. Lors de son application les données sont séparées en plusieurs classes prédéterminés de façon que les individus ayant le plus de similarité. C'est ainsi qu'une des tâches clefs est de trouver le nombre approprié de classes, k . Il existe plusieurs techniques pour déterminer le nombre de classes. Nous discuterons que des cas les plus connus :

1. Méthode du pouce :

Cette méthode est une méthode approximative où le nombre de classes, k est déterminé par : $k \approx \sqrt{n/2}$

2. L'indice de qualité :

Afin d'évaluer la qualité de la classification, les indices inertiels, soient l'inertie intra-classes et l'inertie l'inter-classes sont souvent utilisés. L'inertie intra-classes

‘mesure le degré d’homogénéité entre les objets appartenant à la même classe’ tandis que l’inertie inter-classes ‘mesure le degré d’hétérogénéité entre les classes.’ Il existe plusieurs indices de qualité, par exemple l’indice de Dunn, l’indice de Calinski et Harabasz (CH) ou encore l’indice de Silhouette. Le premier calcule la distance minimale inter-classes et ainsi plus cette distance est grande, meilleur sera la classification.

Introduit par Kauffman et Rousseeuw, l’indice de Silhouette nous donne une représentation visuelle de la distance entre un point d’une classe avec les points des classes voisines. Plus le coefficient ainsi calculé est proche de 1, plus la distance avec les classes voisines (inertie inter) est grande. Ceci représente le nombre de classes optimale. À l’inverse, un coefficient proche de -1 nous indique une mauvaise classification de l’observation.

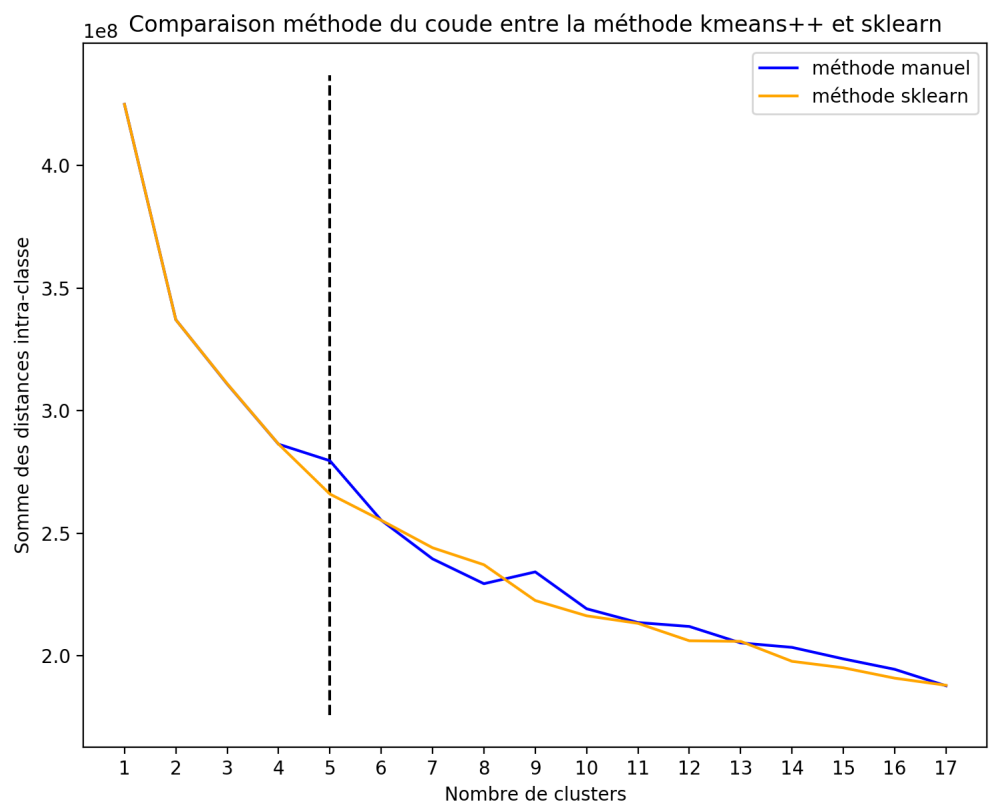
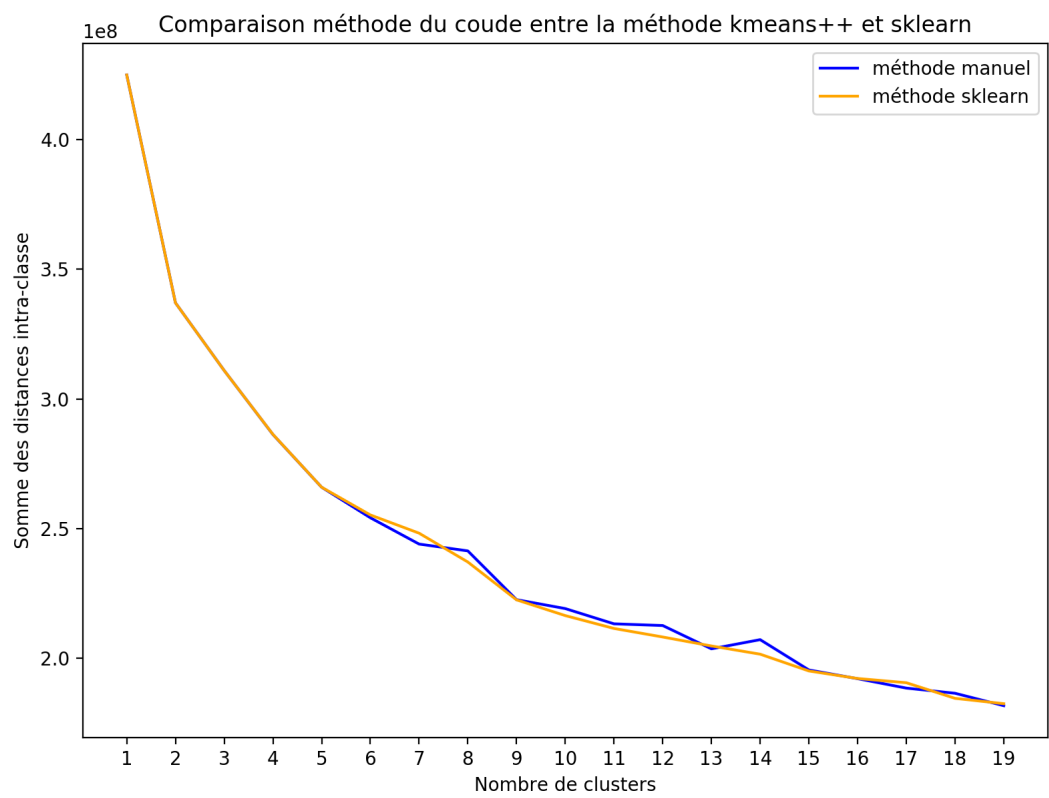
3. Méthode du coude :

La méthode du coude est une technique visuelle très connue. L’idée derrière cette technique est d’implémenter la méthode K-means en parcourant k valeurs. À chacune des k valeurs, la somme des erreurs au carré est calculée et est affichée sur un graphique, nous permettant à mieux visualiser les résultats. L’objectif est de choisir la valeur k (qui sera le nombre de classes) créant un effet de ‘coude’, c’est-à-dire provoquant une baisse plus conséquente, plus soudaine de la somme des erreurs au carré. Nous disons ceci en gardant en tête que la somme des erreurs aura toujours tendance à baisser, plus la valeur de k est grande.

4. La validation croisée :

La validation croisée regarde la stabilité des classes. Les données sont séparées en au moins deux parties. La première est utilisée pour former les classes tandis que la deuxième sert de validation. Lorsque nous parlons de stabilité, nous parlons de la fréquence à laquelle des classes similaires sont formées lorsque plusieurs itérations sont effectuées. Ainsi une plus grosse fréquence de l’apparition de mêmes classes équivaut à une plus grosse stabilité de ces classes.

Dans le cadre de notre étude, nous choisissons de travailler avec la méthode du coude, une méthode que nous appliquons sur la méthode de `kmeans++`. Cette méthode est d’ailleurs comparée avec celle de `sklearn` afin de déterminer l’exactitude de l’algorithme utilisé. Le graphique ci-dessous affiche la répartition des sommes des erreurs au carré à la fois pour notre méthode, dite la méthode manuelle, ainsi que la méthode proposée par `sklearn`. Nous voyons une baisse plus soudaine lorsque nous avons 5 clusters. Ce sera ainsi le choix du nombre de clusters utilisé lors de l’application des algorithmes de clustering.

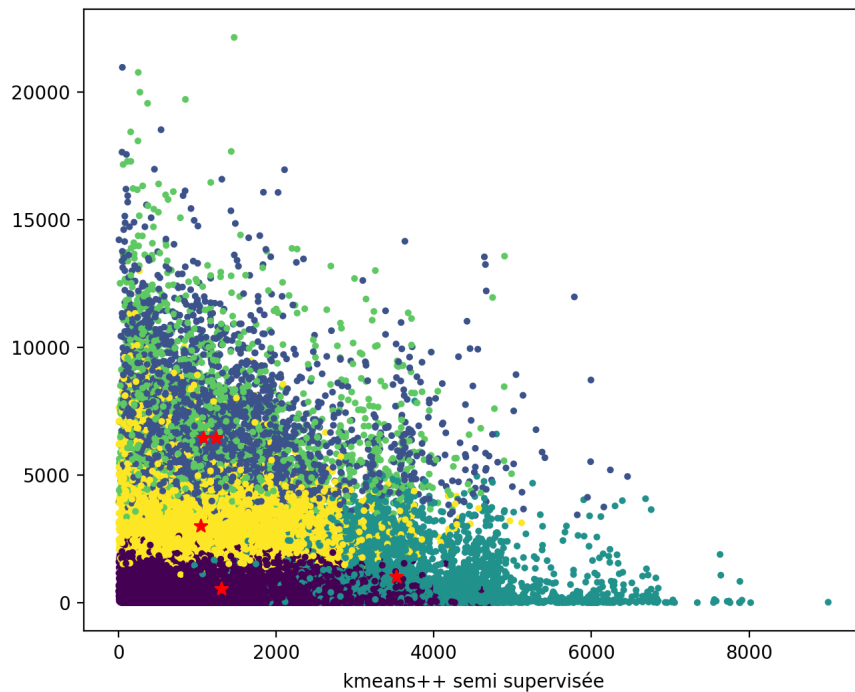


II.2 Expérimentations

Observations et Résultats

Dans cette section nous cherchons à présenter les résultats obtenus par les algorithmes kmeans, kmeans++ et kmeans++ semi supervisées. De plus ce dernier a été analysé plus en détails afin d'afficher l'impact des différents pourcentages de données labellisés. Comme pour l'ACP la variable qualitative a été écartée de notre analyse. 6 essais ont été effectués afin d'évaluer le temps d'exécution des algorithmes (en secondes). La moyenne des distances intra-classes a été obtenu en se basant sur 100 simulations. Le nombre de clusters, déterminé en utilisant la méthode du coude, est 5.

Le graphique ci-dessous affiche les 5 clusters ainsi que leurs centres, tel que proposé par l'algorithme kmeans++ semi-supervisée.



Les tableaux ci-dessous affichent les résultats obtenus pour les différents algorithmes.

Tel que prévu la durée d'exécution moyenne de l'algorithme de kmeans est plus longue que celle du kmeans++ et du kmeans++ semi-supervisée (avec 60% de données labellisées). Cette dernière est d'ailleurs en moyenne > 3 fois plus rapide que le kmeans ou le kmeans++. Au niveau du coût de la fonction nous voyons que le kmeans ++ semi supervisée affiche à nouveau le meilleur résultat (avec la plus petite distance moyenne intra-classes sur 100 simulations). Il est toutefois étonnant de voir que la distance intra-classe moyenne lorsque kmeans est utilisé est plus petite que celle du kmeans++.

TABLE 1 – Résultats en utilisant kmeans

Temps d'exécution Minimum	Temps d'exécution Maximum	Temps d'exécution Moyen	Moyenne
0.867459	4.278867	1.867284	

TABLE 2 – Résultats en utilisant kmeans++

Temps d'exécution Minimum	Temps d'exécution Maximum	Temps d'exécution Moyen	Moyenne
0.952380	1.963426	1.603216	

TABLE 3 – Résultats en utilisant kmeans++ semi-supervisées avec 60% de données labellisées

Temps d'exécution Minimum	Temps d'exécution Maximum	Temps d'exécution Moyen	Moyenne
0.386259	0.616933	0.481936	

- 1.
- 2.
- 3.
- 4.

Références

1. David Arthur, Sergei Vassilvitskii, *K-means++ : The Advantages of Careful Seeding*, 2007
2. Xin Wang, Chaofei Wang, Junyi Shen *Semi-supervised K-Means Clustering by Optimisation Initial Cluster Centers*, 2011
3. Jordan Yoder, Carey E. Priebe, *Semi-supervised K-means++*, 2016
4. Oumaima Alaoui Ismaïli, Vincent Lemaire, Antoine Cornuèjols, *Une méthode supervisée pour initialiser les centres des K-moyennes*, 2016
5. Fabien Panloup, *Apprentissage statistique en grande dimension*, Notes de cours du Master 2 Data Science 2017 - 2018