Master TRIED,

prepaTP-30-10: Rapport TPEX23

Sujet:

Algorithme des k-means

Réalisé par:

FOUTSE YUEHGOH

Année universitaire : 2018/2019

Soit un ensemble de données d'apprentissage non labellisées en dimension 2 contenues dans le fichier **Data1.mat** On dispose pour un ensemble d'exemples à classer de deux caractéristiques. Cet ensemble d'apprentissage comprend 132 exemples qui ont été simulées selon 3 gaussiennes.

1°) Il est question ici d'expliquer le principe de fonctionnement, la nature et les différentes étapes d'un algorithme de k-moyennes, puis retrouver les différentes étapes spécifiquement associées à cet algorithme dans le programme kmoys.m fourni.

Kmeans est un algorithme non supervisé de clustering non hiérarchique. Il permet de regrouper en cluster distinct les observations d'un data set. Pour faire se regroupement, l'algorithme a besoin d'un moyen de comparer le degré de similarité entre les différentes observations. Ainsi, deux données d'un même groupe auront une distance de dis-similarité réduite comparé a c'elle des autres groupe.

L'algorithme K-means est composé des trois étapes suivantes :

- **Initialisation :** On initialise les centres des classes (μ (0) 1, . . ., μ (0) K) (à votre choix) pour donner le pas de départ de l'algorithme (par exemple on choisissant aléatoirement des centres "virtuels", ou K données parmi les données à traiter). Il s'agit donc de démarrer à l'itération t = 0 avec des valeurs initiales pour les paramètres du modèle (μ (0) 1, . . ., μ (0) K).
- **Etape d'affectation (classification)** : Chaque donnée est assignée à la classe du centre dont elle est la plus proche
- **Etape de recalage des centres :** le centre μ de chaque classe k est recalculé comme étant la moyenne arithmétique de toutes les données appartement à cette classe (suite à l'étape d'affectation précédente). La convergence peut être considérée comme atteinte si un nombre maximum d'itérations préfixé a été atteint.

Algorithme
Entrée:

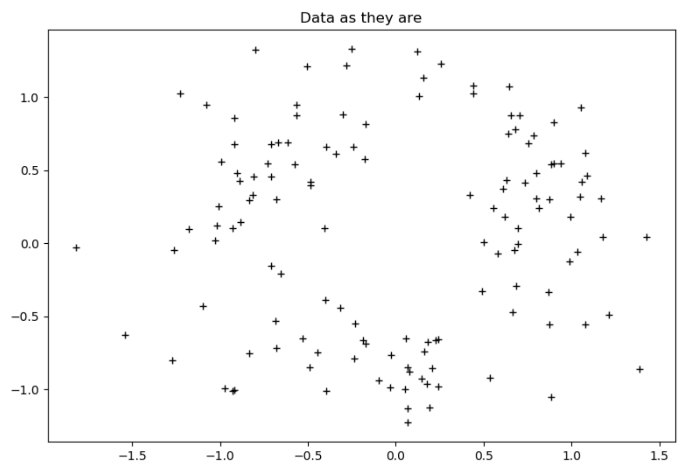
- K le nombre de cluster à former
- Le Training Set (matrice de données)

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 centroïde
JUSQU'A CONVERGENCE (la différence entre la moyenne présidente est nul)
OU (stabilisation de l'**inertie totale** de la population)
FIN ALGORITHME

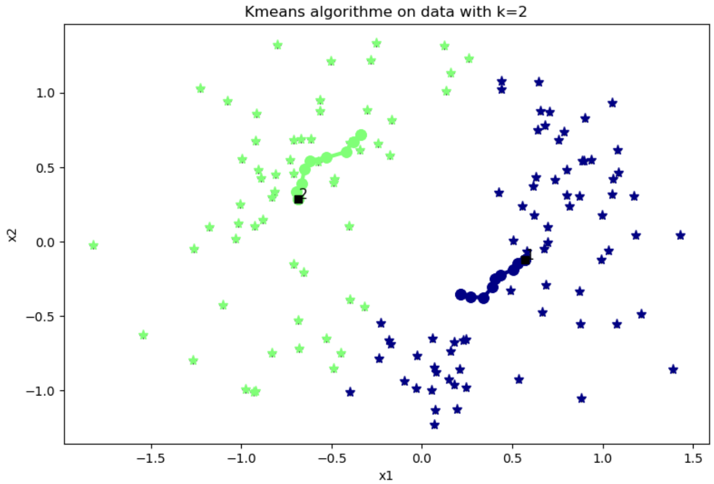
2°) Il nous est demandé d'exécuter le programme demokmeans.m pour effectuer une classification non supervisée des données par un algorithme des k-moyennes. En faisant varier le nombre de groupements réalisés par l'algorithme (k=2, 3, 5, 10, 15, 20), nous allons étudier les variations de « l'inertie intra ». On fera une figure pour chaque cas et un tableau pour justifier les résultats, on indiquera la cardinalité de chaque groupe.

Nous commençons par présenter les données avant l'application de l'algorithme.



On peut voir que les donneés forme 3 groupe vue la forme des nuage de point.

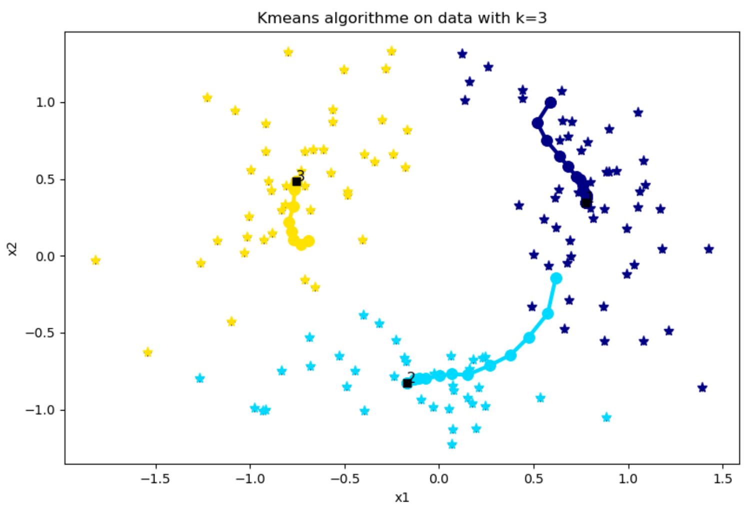
Pour les différents valeurs de k on a les figures suivantes qui présentent les groupements faits par l'algorithme des k-means:



Classe	Cardinalité	Inertie associé
1	72	24.828
2	60	16.408

inertie intra total : 41.2360

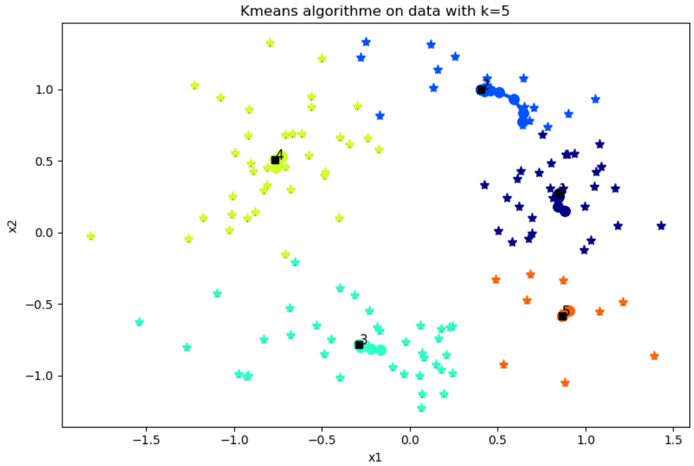
Le tableau des cardinalité et l'inertie intra.



Classe	Cardinalité	Inertie associé
1	51	6.8352
2	37	2.5213
3	44	4.6862

inertie intra total : 14.0427

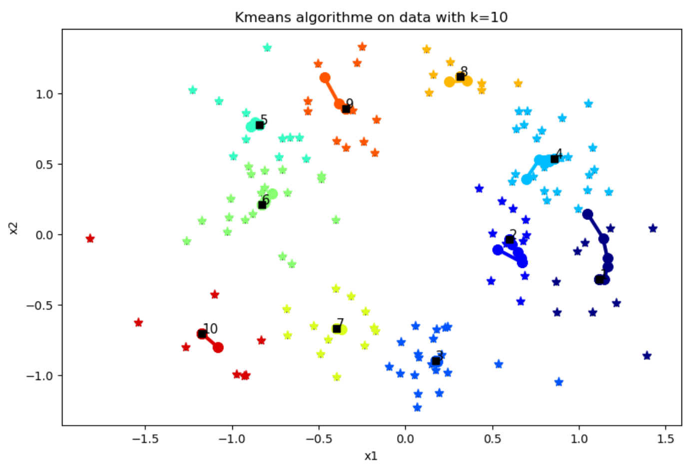
Le tableau des cardinalité et l'inertie intra.



Classe	Cardinalité	Inertie associé
1	29	0.6609
2	17	0.4174
3	38	2.9638
4	38	2.4634
5	10	0.1069

Inertie intra total: 6.612390

Le tableau des cardinalité et l'inertie intra.



Classe	Cardinalité	Inertie associé
1	9	0.0773
2	11	0.0593
3	20	0.2144
4	24	0.3235
5	11	0.0846
6	19	0.251
7	12	0.0623
8	7	0.0166
9	11	0.0752
10	8	0.1007

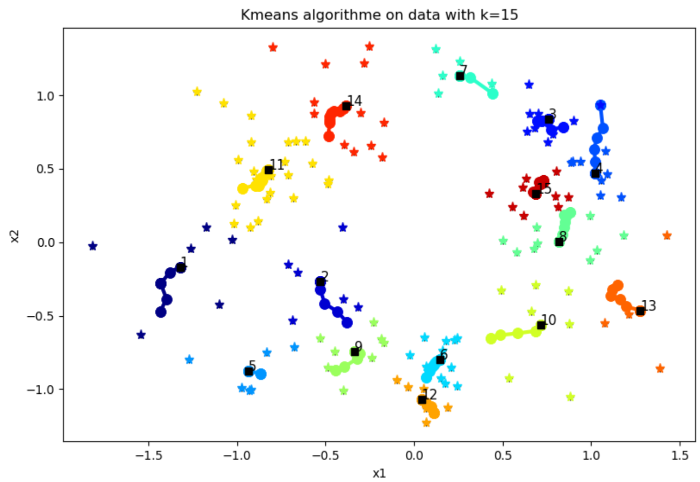
inertie intra total : 1.2649

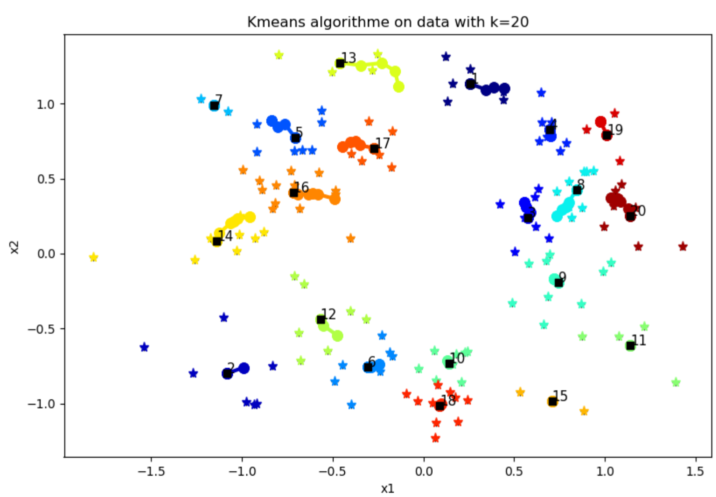
Le tableau des cardinalité et l'inertie intra.

Classe	Cardinalité	Inertie associé
1	6	0.0397
2	6	0.019
3	9	0.0182
4	8	0.01
5	6	0.0129
6	12	0.0225
7	6	0.0083
8	9	0.0351
9	8	0.0175
10	7	0.0382
11	23	0.3756
12	6	0.005
13	4	0.0153
14	12	0.1149
15	10	0.0198

inertie intra total : 0.752

Le tableau des cardinalité et l'inertie intra.



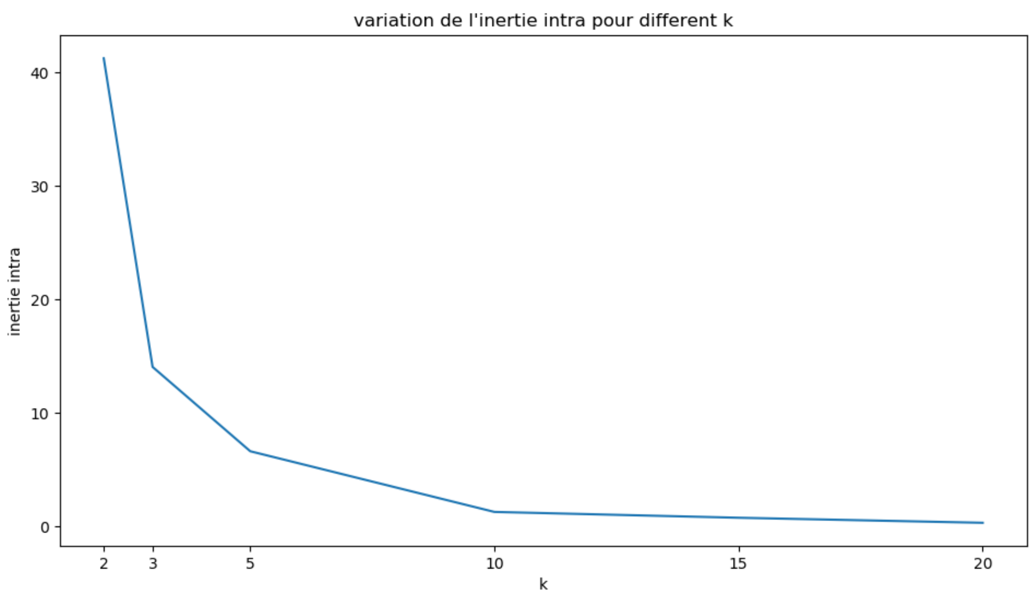


Classe	Cardinalité	Inertie associé
1	6	0.0083
2	7	0.0351
3	7	0.01
4	7	0.0065
5	7	0.0118
6	7	0.0126
7	2	0.0002
8	8	0.0084
9	9	0.0335
10	8	0.0069
11	4	0.0068
12	7	0.022
13	4	0.0062
14	8	0.0425
15	2	0.001
16	13	0.0585
17	6	0.0052
18	10	0.0157
19	3	0.0016
20	7	0.0156

inertie intra total : 0.3084

On remarque que pour de grande valeurs de k l'inertie total diminue ce qui est normale puisque quand on a plus de groupe ça implique que ceux qui sont dans le même groupe se ressemble plus.

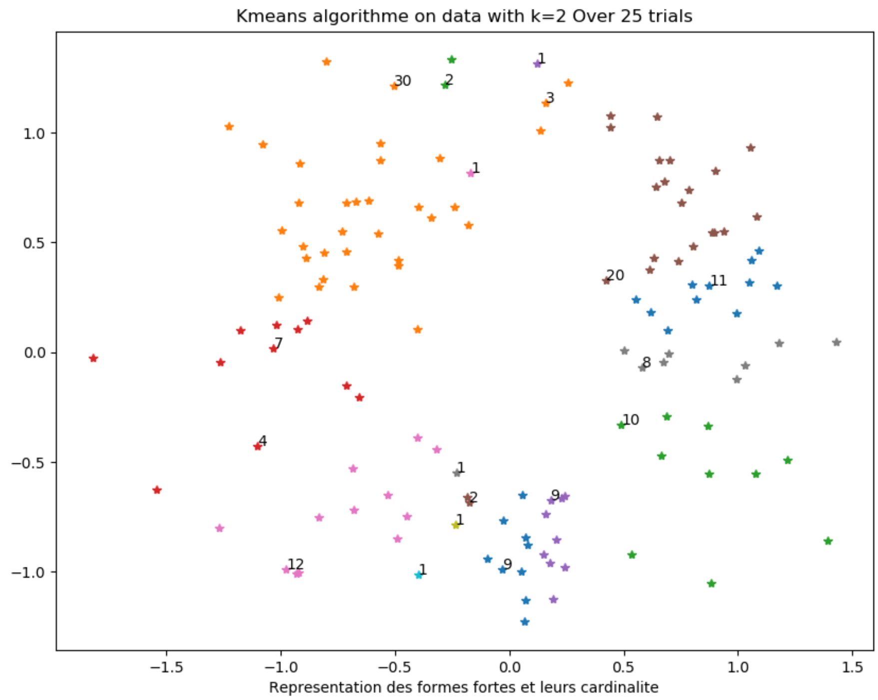
Choisir le nombre de cluster k n'est pas évident car quand nous n'avons pas un a priori ou des hypothèses sur les données, pour un jeu de données très grand, un nombre grand pour k peut conduire à un partitionnement trop fragmenté des données et peut ainsi empêcher de découvrir des patterns pertinents dans les données. Par contre, un nombre de cluster trop petit conduira à avoir des clusters trop généralistes et on n'aura pas de patterns fins à découvrir. La difficulté réside donc à trouver un k qui nous permet de découvrir des patterns intéressant dans notre jeu de données.
Les figures plus hautes et les tableaux montrant leurs cardinalité et l'inertie intra pourrons nous aider a choisi un k optimal. Considérons le plot ci-dessous des inerties intra classe contre les différentes valeurs de k.



Dans les K-means, on a des clusters avec chaqu'un sont centre de gravité. On remarque que quand le nombre de cluster augmente l'inertie décroît. Puis quand on fait un plot de ces résultats de l'inertie intra pour différentes valeurs de k, on remarque une grande chute pour certaines valeurs de k (comme de 2 à 3) puis après elle devient bien plus lente. Alors on pourrait trouver grâce à sa le nombre optimal de cluster en choisissant k entre 3 et 5.

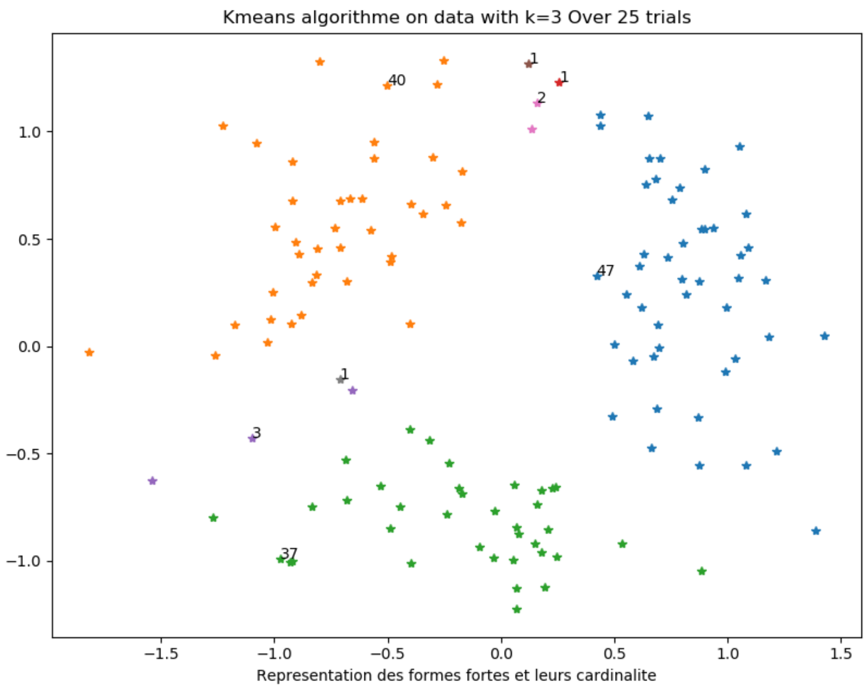
3°) On appelle forme forte les éléments de l'ensemble d'apprentissage qui ont toujours été classés ensemble au cours de plusieurs classifications (initialisations différentes). Il nous est demandé d'écrire un script qui permet de trouver le nombre de formes fortes et la cardinalité de ces formes. Nous avons déterminé les formes fortes pour quelques valeurs de k (k=2, 3, 4, 5), et faire 25 essais à chaque fois.

Les figures ci-dessous nous présentent les formes fortes et leurs cardinalité.

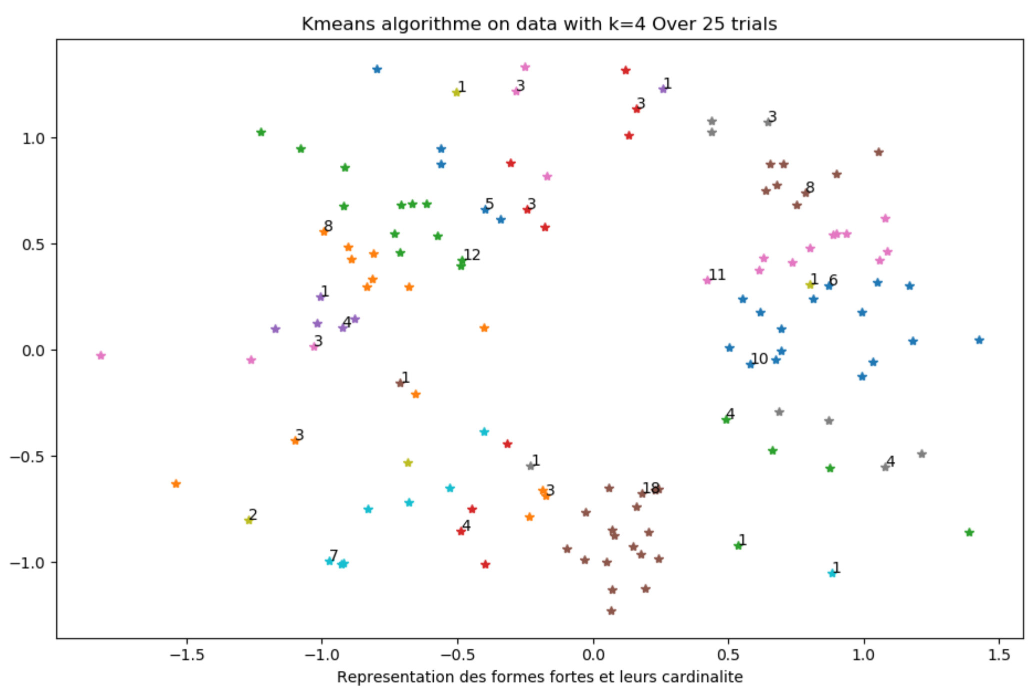


Dans cette figure, on remarque qu'il y'a 5 point qui ont constament changé de classe pour tous les 25 essaie, ils sont characterisé par leurs cardinallité qui vaux 1. On remarque egalement 2 avec une cardinalité de 2 qui nous indique qu'ils sont deux à avoir toujours été classé dans une même classe pour tous les 25 essaie, nous retrouvons aussi des petits groupe de cardinalité 3 et 4 qui représenté également des points qui ont été dans la même classe l'or du classement. On retrouve également des grand groupes qui on cheminer ensemble avec des cardinalités de 30, 20, et des groupes moyen de cardinalité 12, 11, 10, 9, 8,7.

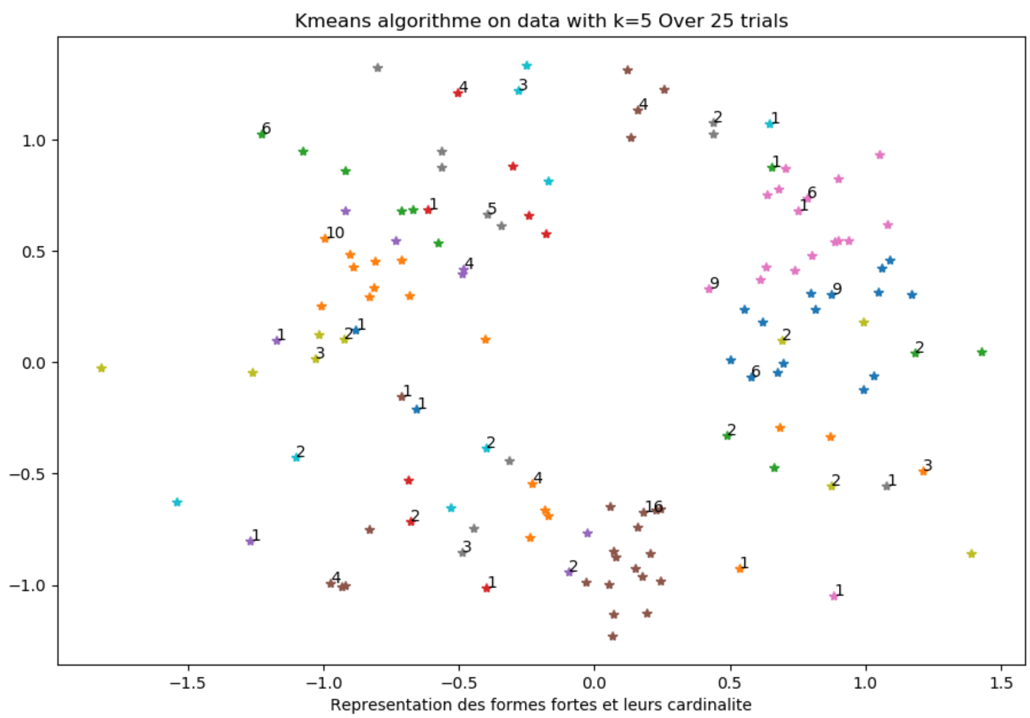
Nous pouvons constater que nous avons 5 point qui ne sont pas de forme forte (selon la définition de forme forte donné plus haut) vue qu'ils ont été seule.



Dans cette figure, on remarque 3 grand groupe qui ont toujours été ensemble, caractérisé par des cardinalités de 47, 40, 37 (des forme forte). Nous avons 1 faible groupe de cardinalité 3, 1 faible groupe de cardinalité 2. 3 point qui a toujours changé de classe (de cardinalité 1). De la figures précédente, on pourrait dire que on a un regroupement plus fiable pour k =3 car on identifie 3 grand groupe qui ont été dans la même classe durant tous les 25 essaie avec différentes initialisation.



Dans cette figure, on remarque beaucoup de petit groupe et pas de grand groupe. Les groupes les plus grands sons de cardinalité 18, 12, 10 et 8. On remarque 7 points qui ont constamment changé de classe, ils ont une cardinalité de 1. Ce regroupement est peut-être peut fiable car on a 7 point qui ont constamment changé de classe.



Dans cette figure, nous avons jusqu'à 12 points qui ont constamment changé de classe. En plus nous avons beaucoup de très petits groupes de formes forte, ce qui pourrait ne pas nous données d'information pertinente sur l'ensemble des données car elle pourrait peut-être être trop général.

Conclusion : En considérant toutes les figures, la figure pour k = 3 est celle qui a donné un regroupement avec 3 groupe pertinent qui pourrais peut-être nous donne des infos pertinente sur l'ensemble des données en se basant sur le fait que nous avons réalisé 25 essaie et que a k =2 qui étais plus petit on n'a pas eu des grand groupe de forme forte. Ceci pourrais soutenir l'hypothèse de la question 2 qui nous a montré a l'aide du graph des inerties que le k optimal pouvais être 3. Mais ceci reste dépendent des données et de l'objectif de l'étude à effectuer.