UNIVERSITÉ DE PARIS DESCARTES

Département de Mathématique | Informatique

Master AI- Machine Learning for Data Science

Apprentissage Non Supervisé

Rapport du projet
Partitionnement des données de consommation d'électricité

Wafa AYAD

December 15th, 2016

Introduction

L'évaluation des performances d'un modèle de classification est primordial. Les différentes méthodes utilisées dans ce processus ont pour objectif de donner une idée claire des performances du predicteur lors du deploiement, ainsi que de comparer plusieurs modèles candidats. Dans ce document, nous allons appliquer des algorithmes d'apprentissage non supervisé sur le jeu de données "electricite.txt" fourni, afin de le partitionner en classe homogènes, à savoir:

- La classification ascendante hierarchique (CAH)
- L'algorithme K-means
- Self-Organizing Map (SOM)

Par la suite, nous faisons une étude comparative entre les résultats obtenus par ces trois méthodes appliquées sur les données initiales et un nouveau jeu de données sans les 48 dernières heures (sans les jours du weekend). Pour en fin, conclure sur la méthode donnant de meilleurs résultats d'apprentissage sur le jeu *electricite.txt*.

1 Analyse des données *electricite*

Ces données sont représentées par une matrice de 2914 individus décrits par 168 variables de valeurs quantitatives continues. Ces individus ne sont pas labelés. Il n'y a pas des valeurs manquantes. Les données n'ont pas besoin de méthodes de prétraitement car elles sont déjà lissées. Donc, elles peuvent être utilisées directement pour le processus de classification. Dans la suite, *electricite_weekend* est le fichier des données sans les 2 jours du weekend, et *electricite* est le fichier les contenant. # Partitionnement des données en classes homogènes Dans ce qui suit nous discutons l'application des techniques de clusturisation étudiées en cours, citées auparavant.

1.1 Choix du nombre de classes

Il existe plusieurs methodes pour trouver le nombre de clusters à lequel les données sont suptibles d'être divisées. Nous avons choisi le package *NbClust* qui s'appuie sur un ensemble d'algorithmes permettant de définir le nombre de classes pour un jeu de donnees en entree, retournant le nombre approprié selon la majorité.

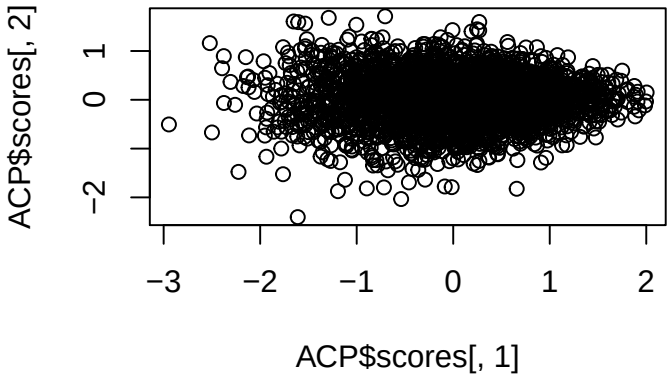
```
*******************************************************************
* Among all indices:
*  3  proposed 3 as the best number of clusters
* 10 proposed 4 as the best number of clusters
* 7  proposed 5 as the best number of clusters
* 1  proposed 6 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is 4


*******************************************************************
```

Figure 1: Nombre de classes des consomation d'electricité

Dans la suite de ce projet, nous considérons que le nombre de classes entre 2 et 6. L'analyse des profils permettera par la suite de valider ce choix et de trouver la meilleure valeur de k.

1.2 Réduction de dimension de *electricite*

- L'application d'une ACP (ou d'une autre méthode de réduction de dimension) sur nos données permet de réduire le nombre de variables de 168 en un nombre plus petit et qui portent toujours l'information contenue dans les variables initiale.



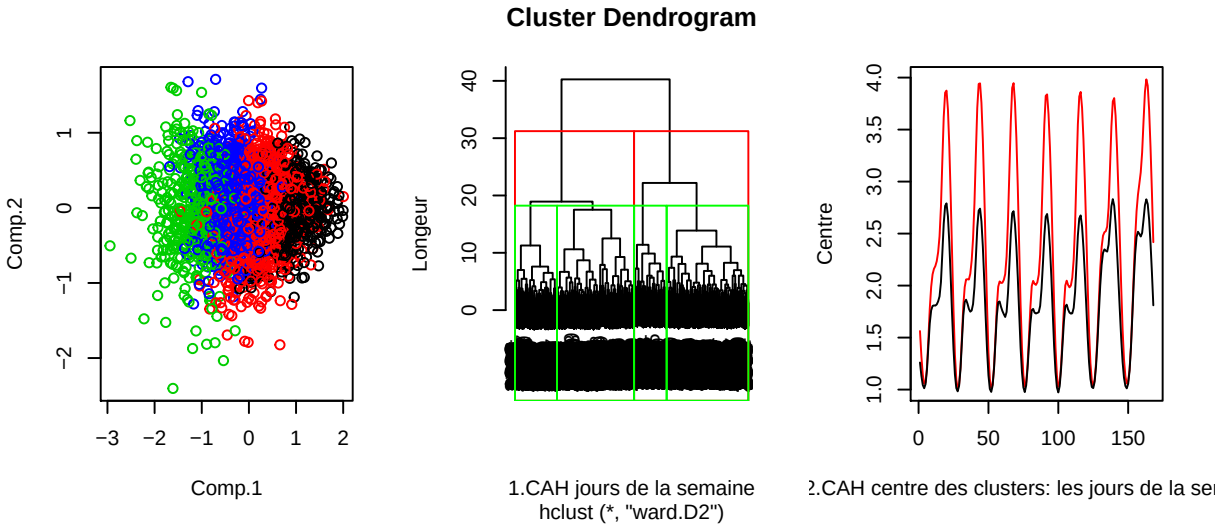
Nous remarquons bien que la grande majorité des individus sont proches l'un aux autres et que les variables sont correlées entre elles. Selon le cumul de la variance donné par l'ACP, on remarque qu'à partir de la 29ème dimension, la vriance ne se change pas d'une manière significative. Donc, pour la classification de ces données, nous n'allons prendre que les 29 premières dimensions, car 95% de l'information est contenue dans ces dimensions.

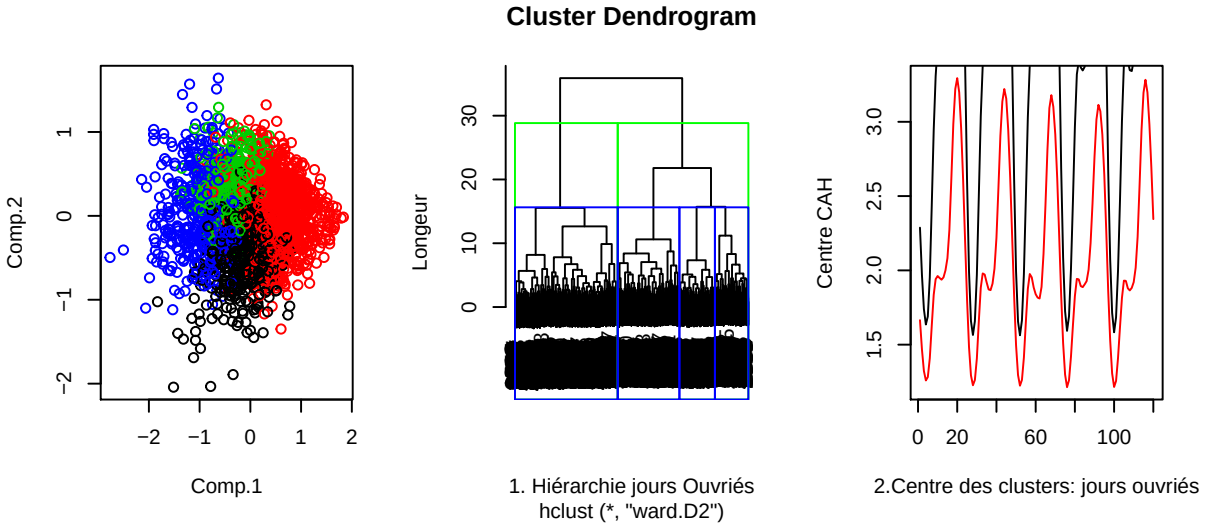
1.3 Application des méthodes non supervisées

Avant de procéder à l'application des différentes méthodes sur les données que nous avons, la compréhension des données sur lesquelles nous désirons apprendre est une étape indisponsable pour prendre le bon sense de l'étude. Dans ce cas, il s'agit des consomateurs d'électricité en Irlande, et l'échantillon des données pris concerne des heures d'une semaine (1er novembre 2010 au 7 novembre 2010). La consomation sera trés élevées dans les jours du weekend (vu que moyenement tous les membres de la familles seront à la maison), elle est moins considérable en heures de nuits (fort prbabelement entre minuit et 6h du matin) et équitable entre les heures du jours de chaque jour de semaine (hors weekend).

1.3.1 Classification Ascendente Hiérarchique (CAH)

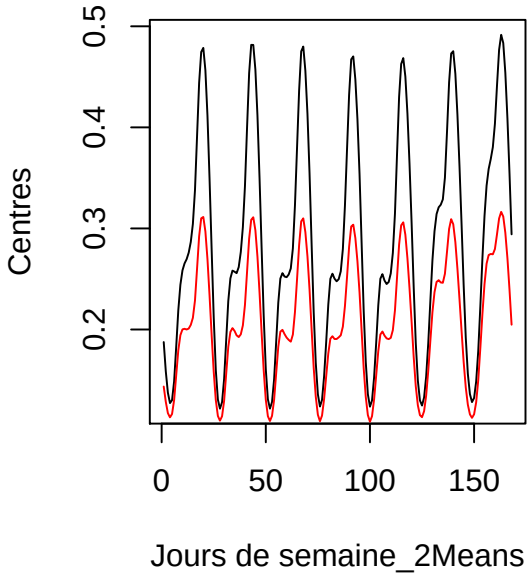
En appliquant la CAH sur les résultats des données réduites de l'ACP, sur les deux échantillons de données (avec et sans les jours du weekend).

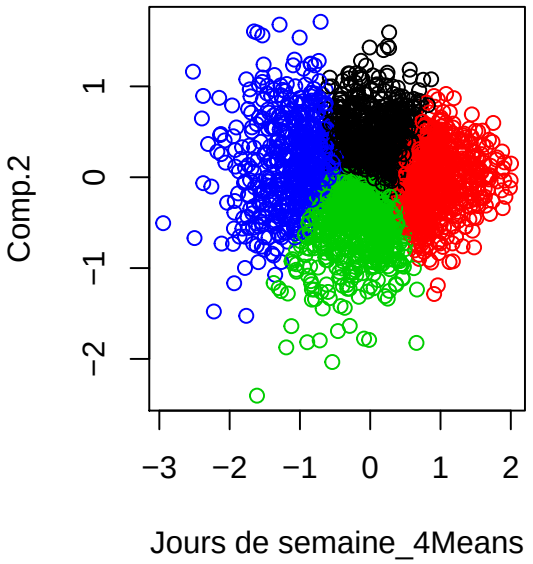

Cluster Dendrogram
Comp.2
Comp.1
Longeur
Centre
1.CAH jours de la semaine
hclust (*, "ward.D2")
2.CAH centre des clusters: les jours de la se

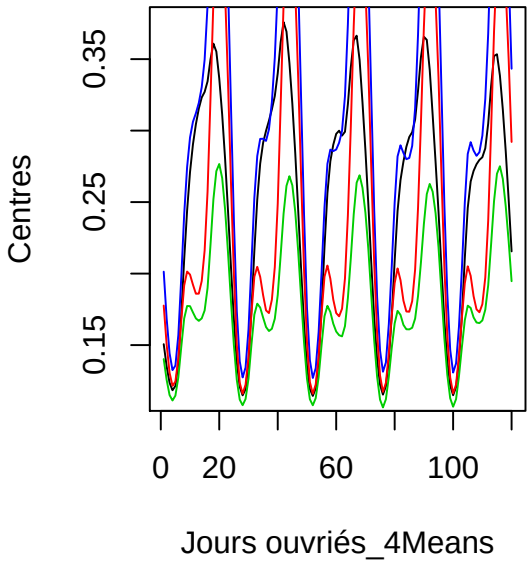

Cluster Dendrogram
Comp.2
Comp.1
Longeur
Centre CAH
1. Hiérarchie jours Ouvriés
hclust (*, "ward.D2")
2.Centre des clusters: jours ouvriés

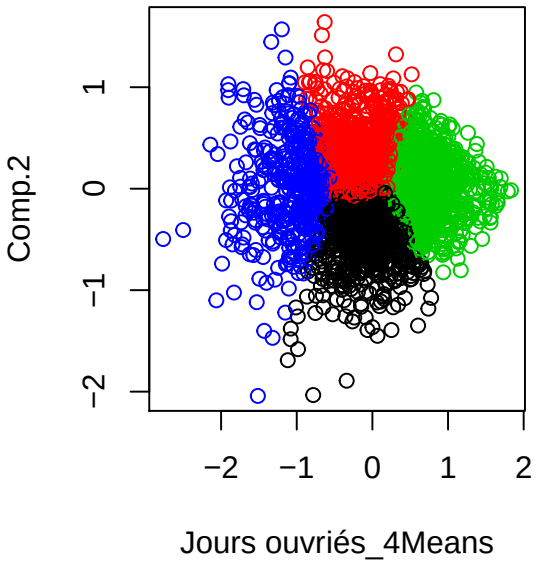
1.3.2 Kmeans

Dans cette partie, nous appliquons l'ACP suivie d'un K-means, et nous discutons les resultats obtenus.

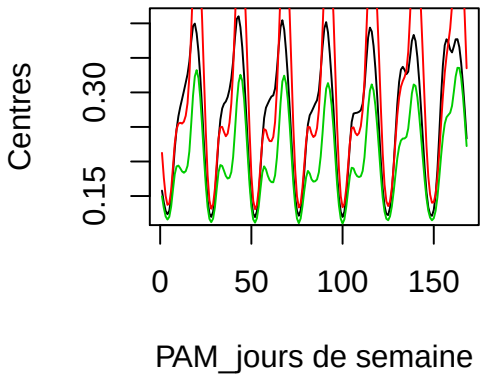


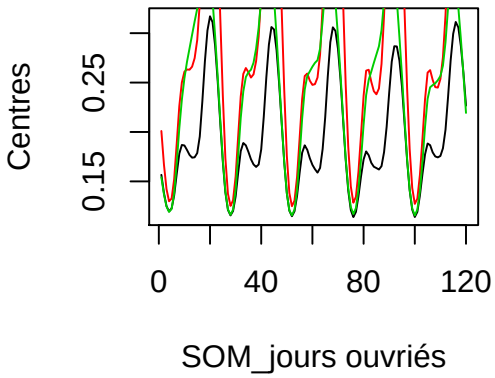




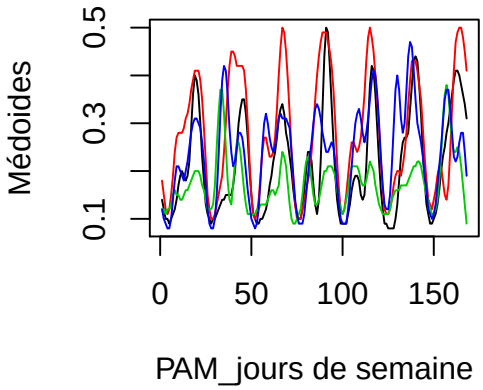


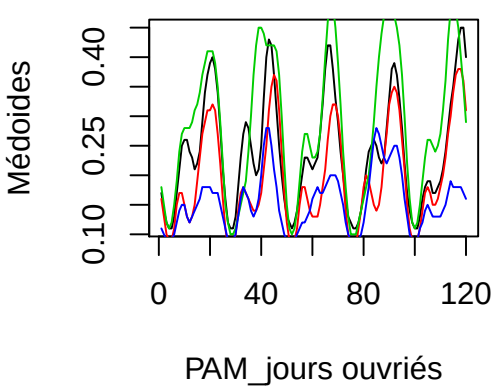
1.3.3 Self-Orgnizing Map (SOM)





1.3.4 PAM





2 Analyse des profils

2.1 Tous les jours de la semaine

D'aprés les différents résultats que nous avons eu des courbes des centres des classes trouvés par l'application des méthodes d'apprentissage non supervisé, nous remarquons que nous pouvons distinguer différents profils de consommateurs d'électricité en Irlande. La première classe (courbe noire) avec une consommation modeste toute la semaine et une petite variation le weekend, ce qui est expliqué par la présence des habitants dans leurs maison vu que c'est les jours de repos. La deuxieme classe (courbe verte) avec une consommation élevée

sur tous les autres profils des autres classes expliquées par la pique maximale au début de la soirée pendant toute la semaine. La troisième classe (en rouge) et la quatrième classe (en bleu) ont une consommation moyenne avec une différence de consommation selon les horaires. Cette dernière est visible en demi-journée et au debut de la soirée.

2.2 Sans les jours du weekend

Le résultats trouvés pour seulement les jours ouvriés indiquent -comme on l'a deviné auparavant- montre une consommation considérablement faible par rapport à celle des résultats précédents. Nous pouvons distinguer 3 classes. Dont la première représente des consomateurs avec consomations trés elevée au milieu de mla semaine par rapport aux autres expliquée par la présence des habitants dans leurs loyer plus ce que les autres jours, et moins faible en début et fin de la semaine. La deuxièmme concerne les consomations moyennes qui se varie d'une journée à une autre, ce qui explique la présence aléatoire des habitants chez eux. ET la dernière, indiquent une faible consomation.

Synthèse et conclusion

L'excustion des méthodes d'apprentissage citées auparavant a donné des résultats différents d'un modèle à un autre, cela est dû au fait que chaque algorithme a ses points forts et points faibles, car le type, la dispertion et le volume des données sont les facteurs sur lesquels les modèles s'appuient pour définir les groupes d'individus. Pour notre cas d'application, en Irlande, il existe moyenenement 3 type de consommateurs d'électricité. Les forts, les moyens et les faibles consommateurs traduisant le taux de présence / abscence des habitants et leurs type.

---

3 Annexe

Le code en r est disponible dans le fichier **projet_Electricite.r**