*Aymeric Chazottes, Djallel Dilmi, Cécile Mallet, Sylvie Thiria*

# TP mise en î uvre : Algorithme des k-moyennes (kmeans)

## I - Les Objectifs

Le but de ce TP est de prendre en main l'algorithme des k moyennes (kmeans) qui est un algorithme de classification non supervisée (abrégée classification dans ce document). Il est Organisé en plusieurs parties.

ŏ La première partie consiste à optimiser la classification non supervisée sur des ensembles simulés en utilisant l'algorithme des kmeans. On étudiera :

- L'importance / impact du nombre de clusters considérés.
- Des techniques pour l'évaluation quantitative d'une classification non supervisée.
- Les frontières de décision définies par l'algorithme des k-moyennes.

ŏ Dans la seconde partie on comparera les deux classifications non supervisées que sont les kmeans et la CAH avec une application sur les iris de Fisher. On étudiera ainsi :

- L'importance / impact du choix de l'algorithme de classification non supervisée.

ŏ Dans la troisième partie on traitera le problème de validation d'une classification non supervisée avec une application sur la classification des séries temporelles de précipitations. On étudiera aussi :

- L'importance / impact du choix codage utilisé (variables) sur la classification résultante.
- La confrontation de la classification résultante à l'expertise métier.

## II - Eléments pour la réalisation du TP

Pour la réalisation de ce TP, il est fortement recommandé d'utiliser les bibliothèques standards de python et de ne rien coder dans la mesure du possible. Toutes les méthodes nécessaires se trouvent dans la suite **scipy (pandas, numpy, scikit-learn, matplotlib,… )**

## III - 1ère Partie du TP :

Soit un ensemble de données d'apprentissage non labellisées en dimension 2 contenues dans le fichier **DATA1.txt** On dispose pour un ensemble d'exemples à classer de deux caractéristiques (variables). Cet ensemble d'apprentissage comprend 132 exemples qui ont été simulées selon 3 gaussiennes.

1°) **Expliquer le principe de fonctionnement, la nature et les différentes étapes d'un algorithme de k-moyennes**, puis retrouver si possible les différentes étapes spécifiquement associées à cet algorithme dans la librairie scikit-learn[1].

2°) **Exécuter les scripts pour effectuer une classification non supervisée** des données par un algorithme des k-moyennes. Faire varier le nombre de groupements réalisés par l'algorithme (k=2, 3, 5, 10, 15, 20), étudier les variations de « l'inertie intra ». **Que dire des autres indicateurs ? Que remarque-t-on ? Qu'en penser ?** (Il faudra clairement discuter ces aspects en se basant sur des valeurs quantitatives et des figures.)

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). **Déterminer les formes fortes pour quelques valeurs de k** (k=2, 3, 4, 5 par exemple). (On pourra faire 25 essais à chaque fois.)
**Présenter les formes fortes sur une figure et discuter les résultats.**

4°) Frontière de décision (question Facultative) :
On vous demande de représenter les frontières de décision dans le cas qui présente le moins de formes fortes de la question 3° précédente. Pour cela, on pourra s'inspirer du code fourni pour la partie kppv pour créer le maillage.
Pour cela on aura besoin d'affecter les éléments du maillage à une classe. On pourra coder cette affectation en utilisant un kppv avec les centroïdes comme ensemble d'apprentissage, Par souci d'efficacité, on pourra sinon utiliser la méthode predict associé à votre objet de clustering.
Si le temps le permet, on pourra également, par curiosité, essayer des valeurs de k plus élevées.

[1] De même cette étape ne doit pas vous prendre plus de 10 minutes, sinon passez à la question suivante.

# *Partie II : k moyennes appliquées aux iris de Fisher*

Les Iris de Fisher est un exemple classique de classification. Pour rappel, le jeu de données (vu dans l'U.E. ACP et le TP kppv ) contient 150 exemples d'iris décrites par 4 variables qui sont : la hauteur et la largeur du sépale (HS et LS), la hauteur et la largeur du pétale (HP et LP). L'expert (Mr Fisher) a identifié 3 classes pour ces iris dénommées Sétosa, Versicolor et Virginica. Ces classes seront utilisées à la fin. (Les classes sont repérées par les indices 1, 2 ou 3.)

1) **Faire une étude préliminaire** du jeu de données (statistiques unidimensionnelles et bidimensionnelles). Représenter les données en 2D en utilisant l'ACP et la T-SNE. **Présenter les données.**

2) Pour la suite de l'exercice, on utilise l'ensemble des données de description en non supervisé.
**Exécuter les scripts pour effectuer une classification non supervisée** des données par un algorithme des k-moyennes. Faire varier le nombre de groupements réalisés par l'algorithme (k=2, 3, 4, 5, 6, 7, 10, par exemple), étudier les variations de « l'inertie intra ». **Que dire des autres indicateurs (dont les formes fortes) ? Que remarque-t-on ? Qu'en penser ?** (Il faudra clairement discuter ces aspects en se basant sur des valeurs quantitatives et des figures.)

3) **(facultatif) On pourra reprendre la question 2) en se limitant aux deux dimensions** issues de l'une des représentations 2D obtenues à la première question.

4) **L'algorithme des K-moyennes permet d'obtenir une partition du jeu de données appris. Dans le cas des iris de Fisher, on a accès aux classes des données.** Pour chacune des iris on a donc une espèce. Ici, on veut comparer ces classes originales avec celles obtenues à l'aide de la partition optimale (dans le cas non supervisé de la question 2), Pour cela, on pourra utiliser l'indice de Rand. **Il faudra aussi discuter le résultat obtenu.**
Le nombre optimal de clusters est-t-il celui connu dans la littérature.
**Discuter les similarités/dissimilarités entre les classes** définies par Fisher et celles des k-moyennes. (Présenter une figure de votre choix ensemble qui fait ressortir la différence, i.e. les points mal classés si on peut le dire.)

5) **Comparer les performances** (inerties, temps d'exécution,…) de la meilleure classification (kmeans) **avec d'autres méthodes de classification.**
Les algorithmes de classification proposés pour la comparaison sont :

- Une classifiction CAH.
- Une classification K-médoides (facultative).

**Argumenter la pertinence de ces deux modèles** en vue de ce que vous avez vue avant.

# *Partie III : algorithme des k moyennes appliqué à la classification des séries temporelles de précipitations*

L'algorithme des k-moyennes est un algorithme de classification non supervisée. Souvent utilisé pour une tache d'exploration, la difficulté majeure réside dans la validation des résultats. Cette partie propose un exemple concret de cette problématique.

## 1- Présentation du domaine d'application

De façon grossière on distingue les pluies dites stratiformes qui se caractérisent par des précipitations de faible ou moyenne intensités et pouvant durer plusieurs heures. A l'opposé, on distingue les pluies convectives, beaucoup plus violentes, mais généralement de plus courte durée (pluie d'orage).
On se propose d'analyser les données d'un spectro-pluviomètre (figure 1.a) qui peut être utilisé pour fournir des mesures à chaque pas de temps (une minute) du taux de pluie sur une petite surface qui peut être assimilée à du ponctuel spatial.
Un épisode de pluie (appelé événement de pluie) engendre une suite de mesures par pas de temps du cumul d'eau de pluie et définit une série temporelle de taux de pluie. La figure 1.b présente la série temporelle correspondante à l'événement de pluie du 22 décembre 2012 – l'axe des abscisses représente l'axe du temps, en ordonnée la quantité d'eau par pas de temps en RR[mm/h].

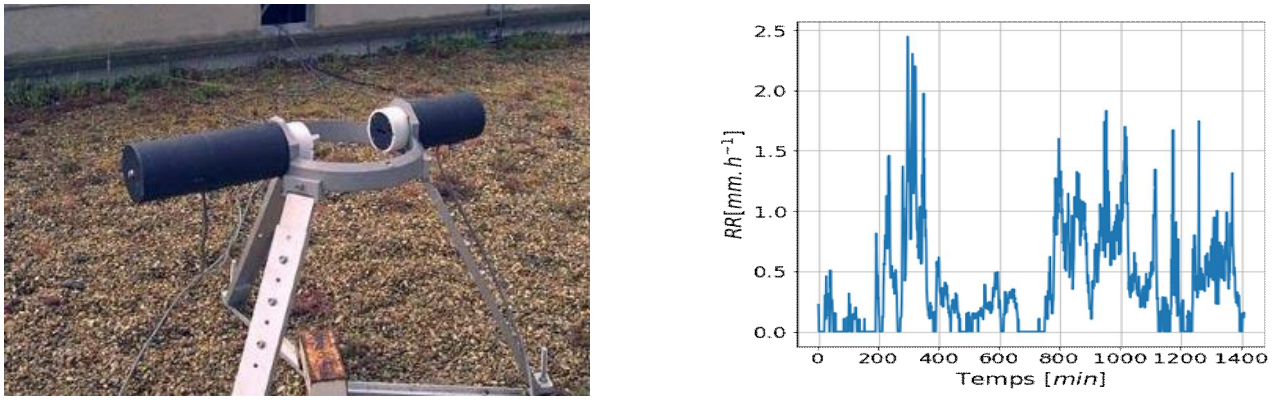



*Figure 1.(a) : à gauche le spectropluviomètre bi-faisceaux DBS (b) à droite séries temporelles représentant l'événement long du 22 décembre 2012 mesuré avec un pas de temps T=1min*

Ces séries temporelles sont décomposées en évènements de pluie qui sont décrits par des variables (cf II et III).
On veut Analyser la variabilité des précipitations à l'aide des k-moyennes i.e. une classification des événements de pluie. On vous demande donc :

## 2- Les Données

**x.txt** : base de données des événements de pluie mesurés par le spectropluviomètre installé sur le site de Palaiseau entre le 1 janvier 2012 et le 31 décembre 2013. Elle est composée de 234 événements décrits par 23 variables, dans une matrice 234x23. Cela signifie que chaque événement (série temporelle) a été décrit par un vecteur de dimension 23 qui, à son tour, correspond à une ligne de la matrice du fichier **x.txt**.

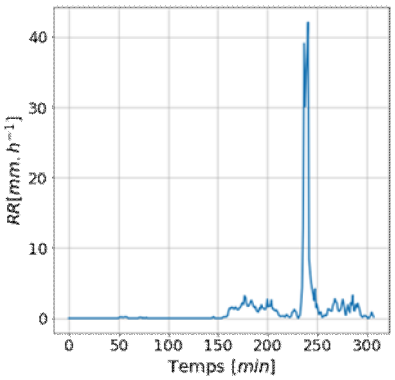
**xn.txt** : base précédente après normalisation des variables (le type de normalisation ne fait pas l'objet de ce TP). Elle est composée donc des 234 événements décrits par 23 variables normalisées, dans une matrice 234x23.

Pour des raisons de temps, on trouvera le tableau de données dans un fichier séparé **xn5.txt** qui présente les 234 événements décrits par ces cinq variables normalisées (une matrice 234x5). L'ordre des variables est l'ordre du listing ci-dessus.

**Visualisation des données brutes (les séries temporelles):**

- utiliser la fonction **plot_event** du fichier **support.py** pour afficher une série temporelle (événement) de précipitations.
- Par exemple pour voir le deuxième événement en face :

```
>>plot_event(1)
  # car les indices en python commence par 0
>> plot_event([1,40])
  # pour afficher les deux événements 2 et 41
```




## 3- Les différentes variables utilisées

La méthode des k-moyennes utilise des distances. Le calcul des distances entre les évènements de pluie (dans notre cas des séries temporelles TS) est parfois complexe pour des raisons diverses. Pour réduire cette complexité d'une part et ne conserver que l'information nécessaire pour la classification, il est d'usage de décrire ces évènements de pluie (TS) par des variables pour simplifier le calcul des distances.

Les variables utilisées pour décrire nos événements de pluie sont dans le tableau :

| Number (#) | Feature name | symbol | Formula |
|---|---|---|---|
| 1 | Durée de l'événement | $D_e$ | $D_e = T_{end} - T_{begin} + 1$ [min]<br>With $T_{begin}$: Event start time and $T_{end}$: Event end time |
| 2 | Moyenne du taux de pluie | $R_m$ | $R_m = \frac{1}{D_e}\sum_{t=T_{begin}}^{t=T_{end}} RR_t$ [mm h$^{-1}$] |
| 3 | Durée non pluvieuse intra-événement | $D_d$ | $D_d = \sum_{t=T_{begin}}^{t=T_{end}} I_t$ [min]<br>With $I_t = \begin{cases} 1 & if\ RR_t = 0\ [\text{mm } h^{-1}] \\ 0 & else \end{cases}$ |
| 4 | premier quartile | $Q_1$ | The 25th percentile [mm h$^{-1}$] |
| 5 | Médiane | $Q_2$ | the 50th percentile [mm h$^{-1}$] |
| 6 | Throisième quartile | $Q_3$ | The 75th percentile [mm h$^{-1}$] |
| 7 | Temps inter-événement précédent | $IET_p$ | $IET_p = T_{begin}(current\ event) - T_{end}(previous\ event) + 1$ [min] |
| 8 | Moyenne du taux de pluie / pluie | $R_{m,r}$ | $R_{m,r} = \frac{1}{(D_e - D_d)}\sum_{t=T_{begin}}^{t=T_{end}} RR_t$ [mm h$^{-1}$] |
| 9 | Écart-type du taux de pluie s/ événement | $\sigma_R$ | $\sigma_R = \sqrt{\frac{1}{D_e}\sum_{t=T_{begin}}^{t=T_{end}} (RR_t - R_m)^2}$ [mm h$^{-1}$] |
| 10 | Mode | $M_0$ | $M_0$ =the most frequent $RR_t$ |
| 11 | Maximum du taux de pluie | $R_{max}$ | $R_{max} = \max(RR_t)$ |
| 12 | Pourcentage de non pluie intra évènement | $D_{d\%e}$ | $D_{d\%e} = \frac{D_d}{D_e}$ |
| 13 | cumul d'eau | $R_d$ | $R_d = R_m * D_e / 60$ [mm] |
| 14 | Coefficient de variation du taux de pluie d'ordre c | $P_{c1}$ | $P_{c_i} = \sum_{t=T_{begin}}^{t=T_{end}-1} \lvert RR_{t+1} - RR_t \rvert^{c_i}$<br>For $c_i = 0.5, 1, 2$ |
| 15 | | $P_{c2}$ | |
| 16 | | $P_{c3}$ | |
| 17 | Coefficient de variation normalisé du taux de pluie d'ordre $c_i$ | $P_{C_{N1}}$ | $P_{C_{Ni}} = \frac{P_{c_i}}{D_e}$<br>For $i = 1..3$ |
| 18 | | $P_{C_{N2}}$ | |
| 19 | | $P_{C_{N3}}$ | |
| 20 | Coefficient de variation du taux de pluie d'ordre C seuillé à S | $P_{S,C}$ | $P_{S,C} = \sum_{t=T_{begin}}^{t=T_{end}-1} \lvert \max[(RR_{t+1} - S), 0] - \max[(RR_t - S), 0] \rvert^C$<br>With $s = 0.3\ and\ c = 2$ |
| 21 | Paramètre de convectivité $\beta_L$ | $\beta_{L1}$ | $\beta_{L_i} = \frac{\sum_{i=T_{begin}}^{T_{end}} RR_t\,\theta(RR_t - L_i)}{\sum_{i=T_{begin}}^{T_{end}} RR_t}$<br>For $L_i = 0.3, 1, 3\ mm\ h^{-1}$<br>With $\theta(RR_t - L_i)$ is the Heaviside function defined as<br>$\theta(RR_t - L_i) = 1\ if\ RR_t \geq L_i$<br>$\theta(RR_t - L_i) = 0\ if\ RR_t < L_i$ |
| 22 | | $\beta_{L2}$ | |
| 23 | | $\beta_{L3}$ | |

## **Prise en main des données** (visualisation- statistique descriptive)

1) Faire une étude préliminaire du jeu de données.
    a. Après avoir lu le descriptif des différentes variables, afficher les deux événements 2 (i=1) et 121 (i=120) et analyser les deux descriptions (x[1, :] et x[120, :]). Choisir deux ou trois des variables et expliquer ce qu'elles permettent de décrire.
    b. Faire une synthèse des caractéristiques statistiques des 23 variables (moyenne, max, ..) Calculer et commenter la matrice de corrélation.
2) Etude des 23 variables descriptives des événements de pluie à l'aide d'une analyse en composantes principales (ACP). En particulier :
    a. L'apport de l4ACP sur l'étude des corrélations linéaires entre les différentes variables.
    b. L'insuffisance de l'ACP dans le cas des corrélations non linéaires.

## **Classification des événements de pluie** (Application des K-moyennes)

3) Trouver une classification optimisée des données complètes par kmeans. Commenter/ interpréter les résultats.

4) Pour éliminer la redondance de l'information, on reprend en se limitant au cinq variables : (durée de l'événement (#1), écart-type du taux de pluie (#9), maximum du taux de pluie (#11), Coefficient de variation du taux de pluie d'ordre 0.5 (#14), cumul d'eau (#13) ).

    a. Dans l'introduction du domaine, il est écrit : « De façon grossière on distingue les pluies dites **stratiformes** qui se caractérisent par des **précipitations de faible ou moyenne intensités (écart-types et maximum faibles)** et pouvant durer plusieurs heures (**durées longues**). A l'opposé, on distingue **les pluies convectives,** beaucoup plus violentes (écart-types et **maximum fortes**), mais généralement de **plus courte durée** (pluie d'orage) (durées courtes).. »
    Il est demandé d'afficher l'ensemble des points sur les différents plans des cinq variables de description (par paire).
    Les différentes figures sur les cinq variables sont-elles en adéquation avec cette description des deux types ?
    b. Si la réponse est oui, une classification k-moyennes en deux classes permettrait-elle de retrouver ces deux grandes familles de précipitations ?
    c. En se basant d'une part sur la description des deux types de pluie et d'autre part sur les caractéristiques générales des deux groupes, proposer un label pour chaque classe.