Chapitre 7 - Apprentissage non supervisé : Le regroupement

Neila Mezghani

(Hiver 2025)

COLLÈGE DE
BOIS-DE-BOULOGNE

Plan du cours

Techniques d'apprentissage machine

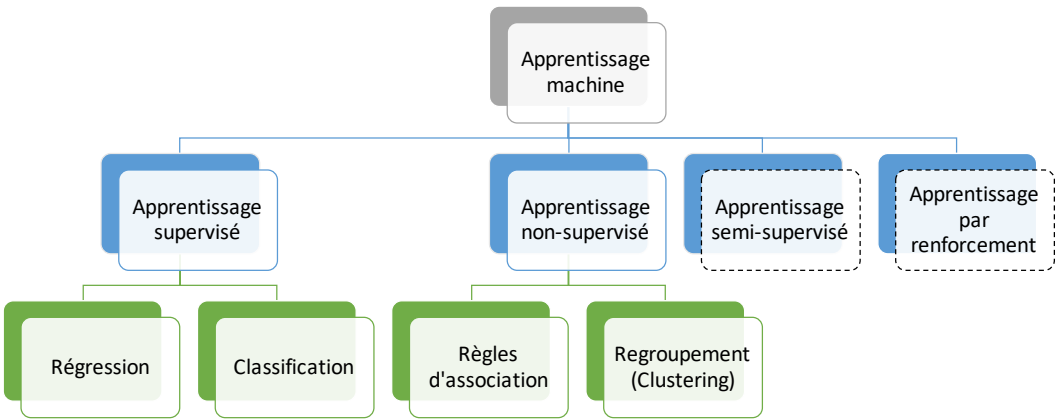
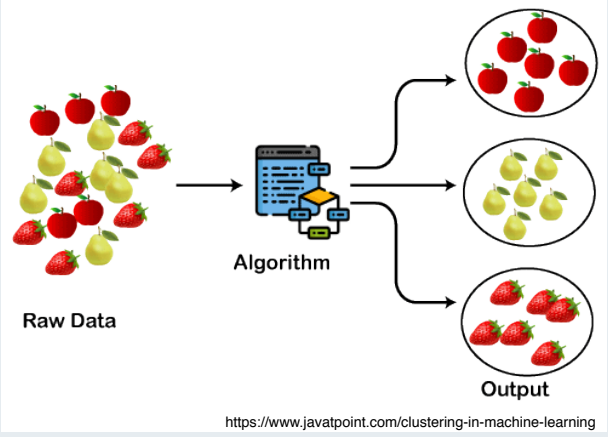


Table of Contents

Apprentissage non-supervisé (1/3)

- Contrairement à l'apprentissage supervisé, dans le cas non supervisé les données de sortie ne sont pas connues.
- On dispose de n individus décrits par leur vecteur de caractéristiques $D_n = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$.
- Le système apprend alors de lui-même à organiser les données ou à déterminer des structures dans les données.
- La tâche d'apprentissage non-supervisé la plus courante est le regroupement (*clustering*) qui consiste à regrouper les données d'entrées selon leurs caractéristiques communes.

Apprentissage non-supervisé (2/3)



https://www.javatpoint.com/clustering-in-machine-learning

Apprentissage non-supervisé (3/3)

- En épidémiologie, on essaye toujours de trouver des hypothèses explicative de l'apparition de l'épidémie.
- À partir d'un ensemble assez large de victimes de cancer du foie, on veut tenter de faire émerger des hypothèses explicatives
  - Les scientifiques de données vont déterminer les différents groupes
  - L'épidémiologiste chercherait ensuite à associer à divers facteurs explicatifs, origines géographique, génétique, habitudes ou pratiques de consommation, expositions à divers agents potentiellement ou effectivement toxiques (métaux lourds, toxines telle que l'aflatoxine, etc.).

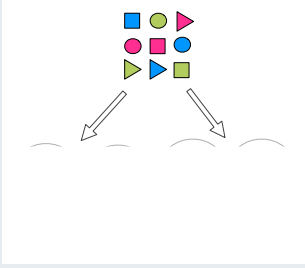
Table of Contents

Regroupement

Principe

- Le **regroupement**, aussi appelé **agrégation** (*clustering*), est une méthode statistique d'analyse et de classification non supervisée de données (*unsupervised learning*)
- Objectif : former des groupes ou agrégats (*clusters*) d'objets similaires à partir d'un ensemble hétérogène d'objets.

Principe



Principe

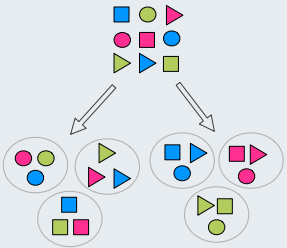


Figure – Regroupement des données en trois groupes selon la forme (à gauche) et selon la couleur (à droite).

Applications

- Marketing : segmentation du marché segmentation d'un marché en groupes de clients distincts à partir de bases de données d'achats.
- Environnement : détermination de zones terrestres similaires (pour leur utilisation) dans une base de données d'observation de la terre.
- Assurance : détermination de groupes d'assurés distincts qui ont présenté un nombre important de réclamations.

Approches de regroupement

Plusieurs approches sont utilisées. Parmi lesquelles :

- **Partitionnement :** Construire plusieurs partitions puis les évaluer selon certains critères
- **Hiérarchiques :** Créer une décomposition hiérarchique des individus selon certains critères
- **Basés sur la densité :** basés sur des notions de connectivité et de densité
- **...**

Table of Contents

Principe de l'algorithme $k-$moyenne

L'algorithme $k-$moyenne

- Les méthodes de regroupement par partition consistent à construire une partition unique en C groupes à partir des N individus à regrouper.
- Il existe de multiples méthodes de partitionnement des données :
  - la méthode $k-$moyennes
  - la méthode des $C-$medoids
  - les nuées dynamiques.
- L'algorithme $k-$moyennes ($k-means$) utilise une technique d'affinement itératif qui consiste à améliorer progressivement la qualité des groupes selon un indice de similarité entre les différents individus.

L'algorithme $k-$moyenne

Algorithme $k-$moyennes : On cherche à déterminer k regroupements à partir d'un ensemble de N individus.

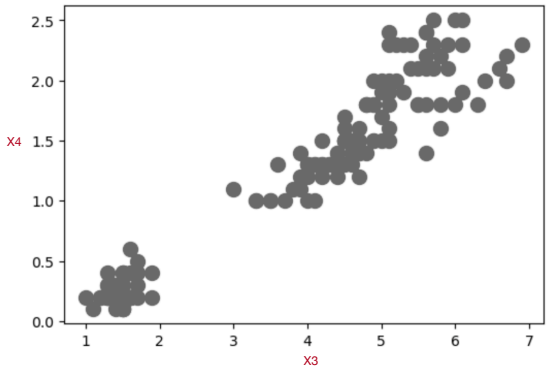
1. Choisir k centres initiaux G_i avec $i = 1, 2, ...k$.
2. Répartir chaque individu dans le groupe G_i le plus proche.
3. Recalculer les nouveaux centres G_i.
4. **Répéter** les étapes 2. et 3.
5. **jusqu'à ce que** les centres deviennent stables.

`sklearn.cluster.KMeans`

Exemple de regroupement (1/3)

	X1	X2	X3	X4
0	4.3	3.0	1.1	0.1
1	4.4	2.9	1.4	0.2
2	4.4	3.0	1.3	0.2
3	4.4	3.2	1.3	0.2
4	4.5	2.3	1.3	0.3
...	...	...	...	...
145	7.7	3.8	6.7	2.2
146	7.7	2.6	6.9	2.3
147	7.7	2.8	6.7	2.0
148	7.7	3.0	6.1	2.3
149	7.9	3.8	6.4	2.0

```
fig = plt.figure(figsize=(6, 4))
plt.scatter(df[['X3']], df[['X4']], s = 100, c = 'dimgray', label = 'Dataset')
plt.show()
```



Exemple de regroupement (2/3)

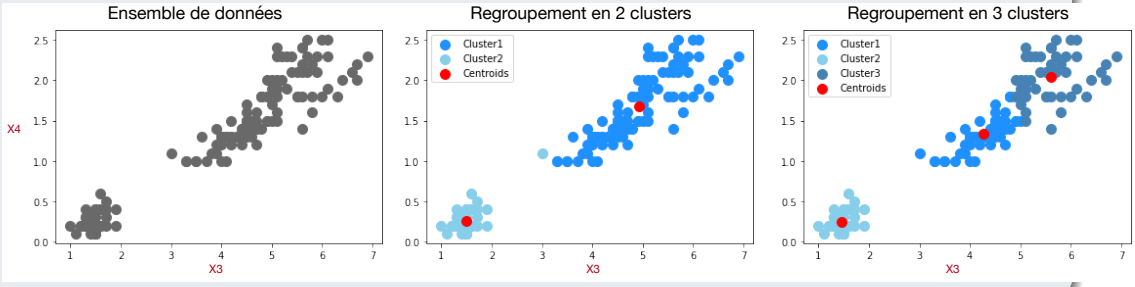
	X1	X2	X3	X4
0	4.3	3.0	1.1	0.1
1	4.4	2.9	1.4	0.2
2	4.4	3.0	1.3	0.2
3	4.4	3.2	1.3	0.2
4	4.5	2.3	1.3	0.3
...	...	...	...	...
145	7.7	3.8	6.7	2.2
146	7.7	2.6	6.9	2.3
147	7.7	2.8	6.7	2.0
148	7.7	3.0	6.1	2.3
149	7.9	3.8	6.4	2.0

-

```
X = df[['X3','X4']].values
fn = ['X3', 'X4']
```

```
KC=2
kmeans = KMeans(n_clusters = KC, init = 'random', n_init = 10, random_state = 0)
y_kmeans = kmeans.fit_predict(X)
```

Exemple de regroupement (3/3)



Avantages et inconvénients du k-means

Avantage :

- L'algorithme k-moyenne est facile à implémenter

Inconvénients :

- La valeur de k doit être choisi a priori.
- Les clusters dépendent de l'initialisation et de la distance choisie (`Exemple de simulations de l'initialisation`)

k-means pour les données qualitatives

- L'algorithme K-means est conçu pour fonctionner uniquement avec des données quantitatives (numériques).
- Il repose sur la mesure de distances, comme la distance euclidienne, entre les points de données, ce qui nécessite que les variables soient numériques.
- Cependant, il existe des variantes et des approches pour utiliser des données qualitatives ou mixtes (quantitatives et qualitatives) dans un contexte de clustering.

k-means pour les données qualitatives

Pour les données qualitatives

- K-modes : Une variante de K-means qui fonctionne avec des données catégoriques. Elle utilise une mesure de similarité adaptée, comme la distance de Hamming.
- K-prototypes : Une combinaison de K-means et K-modes, conçue pour traiter des données mixtes (quantitatives et qualitatives).

Se référer à K-modes et K-prototypes

Table of Contents

Évaluation de la qualité du regroupement

Évaluation de la qualité du regroupement

- Une bonne méthode de regroupement produit des groupes d'individus homogènes entre eux et distincts des autres groupes.
- Critères
  - Une similarité **intra-groupe** importante, pour obtenir les groupes les plus homogènes possibles.
  - Une similarité **inter-groupe** faible afin d'obtenir des sous-ensembles bien distincts.

Inertie intra-groupes (1/3)

- Soit $\mathcal{D} = \{\mathbf{x}_i\}_{i=1...N}$ un ensemble de données contenant N individus.
- Chaque individu $\mathbf{x}_i = (x_i^1, ..., x_i^K)'$ est décrit par K caractéristiques = un individu est un point de l'espace $\mathbb{R}^K$ caractérisé par un vecteur $\mathbf{x}_i = (x_i^1, ..., x_i^K)'$.
- On suppose que l'ensemble des données est réparti dans C groupes G_c, $c = 1...C$.

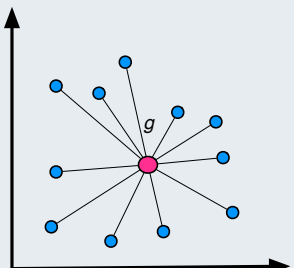
  Se référer à `kmeans.inertia_`

Inertie intra-groupes (2/3)

- Pour chaque groupe G_c, l'inertie est la variance des individus formant le groupe. Elle est déterminée selon :

$$J_c = \sum_{i \in G_c} d^2(\mathbf{x}_i, \mu_c) \qquad (1)$$

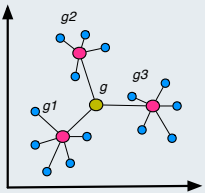
où μ_c désigne la moyenne du groupe G_c.



Inertie intra-groupes (3/3)

- L'inertie intra-groupe est donc la somme des inerties de tous les groupes :

$$J_w = \sum_{c=1}^{C} \sum_{i \in G_c} d^2(\mathbf{x}_i, \mu_c) = \sum_{i \in G_c} J_c \quad (2)$$



- L'inertie intra-groupe permet, donc, de mesurer la concentration des individus du groupe autour du centre de gravité.
- On peut ainsi conclure que plus la valeur de l'inertie est faible, plus les individus sont répartis (moins dispersés) autour du centre de gravité du groupe en question.

Inertie inter-groupes (1/2)

- Soit μ la moyenne de toutes les données indépendamment de leur groupe d'appartenance :

$$\mu = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_i \tag{3}$$

où N désigne le nombre d'individus.

- L'inertie inter-groupes :

$$J_b = \sum_{c} N_c d^2(\mu_c, \mu) \tag{4}$$

- Contrairement à l'inertie intra-groupe, il n'y a pas une méthode pour l'inertie inter-groupe.

Inertie inter-groupes (2/2)

Deux regroupements possibles sur un même ensemble de données : (a) une bonne partition et (b) une meilleure structure puisque l'inertie inter-groupe est plus grande.

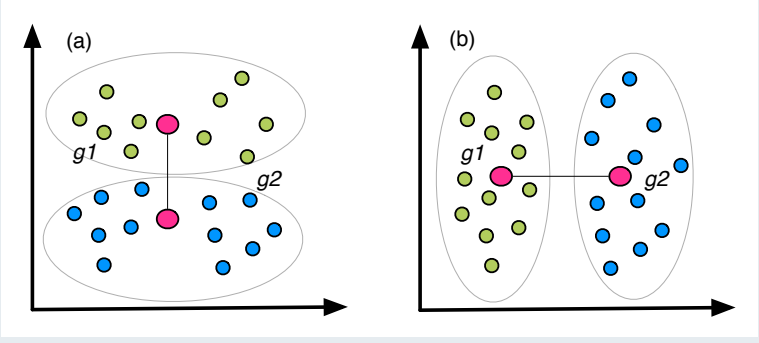


Table of Contents

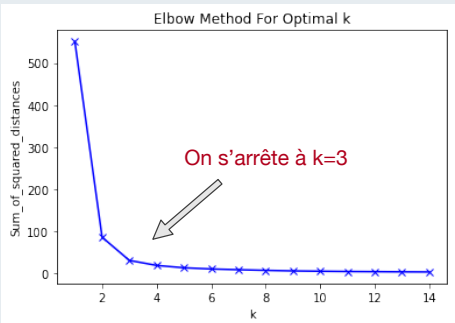
Détermination du nombre optimal de regroupement

Trois méthodes sont généralement employées :

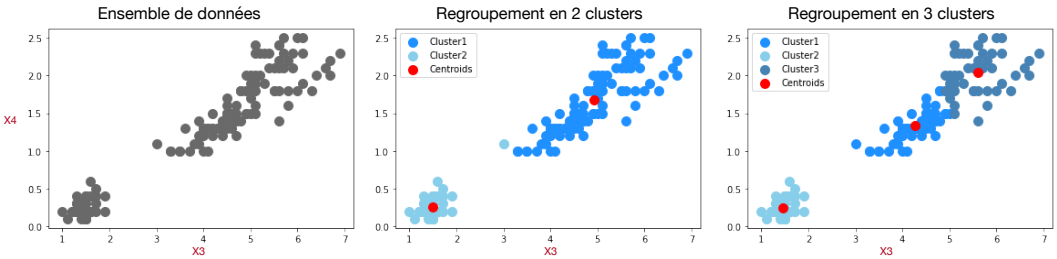
- La méthode de coude (Elbow method)
- La méthode de silhouette (silhouette average).
- The Gap Statistic

La méthode de coude

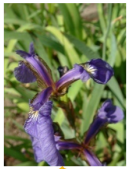
- La méthode de coude (Elbow method) : basée sur la minimisation de la somme des carrés des écarts à l'intérieur des clusters (`clusters.kmeans.inertia_`)
- Examine le pourcentage de variance expliqué en fonction du nombre de clusters
- On s'arrête lorsque l'ajout d'un cluster n'ajoute pas grand chose à la variance



Exemple de regroupement (1/3)

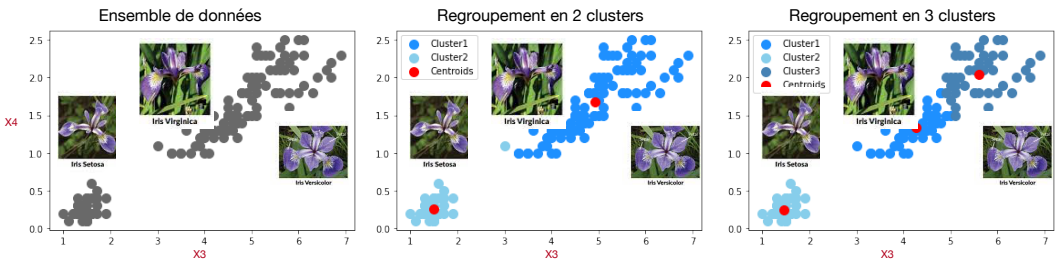


Exemple de regroupement (2/3)



	A	B	X3	X4	E
1	Sepal Length	Sepal Width	Petal Length	Petal Width	Class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5	3.4	1.5	0.2	Iris-setosa

Exemple de regroupement (3/3)



Annexe : Distance de Hamming

Soient deux vecteurs A et B de longueur n, la distance de Hamming d_H est définie par :

$$d_H(A, B) = \sum_{i=1}^{n} 1(A_i \neq B_i)$$

Où :

- A_i et B_i sont les i-ème éléments des vecteurs A et B,
- $1(A_i \neq B_i)$ est une fonction indicatrice qui vaut 1 si $A_i \neq B_i$, et 0 sinon.

Annexe : Distance de Hamming - Exemple

Si $A = [1, 0, 1, 1]$ et $B = [1, 1, 0, 1]$, alors :

1. Comparer chaque position :
   - $A_1 = B_1$ (pas de différence),
   - $A_2 \neq B_2$ (1 différence),
   - $A_3 \neq B_3$ (1 différence),
   - $A_4 = B_4$ (pas de différence).
2. Compter les différences : $d_H(A, B) = 2$.