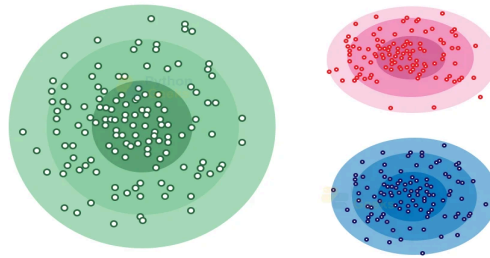


Data Mining

ENSIA 2025-2026

Lab sheet N°8: Clustering



K-Means clustering Notebook (90 minutes):

- **Notebook:** [Clustering Lab.ipynb](#)
- **Dataset:** [load_digits — scikit-learn 1.7.2 documentation](#)
- **Distance metrics:** [cdist — SciPy v1.16.2 Manual](#)
- **K-Means:** [KMeans — scikit-learn 1.7.2 documentation](#)
- **Mini-Batch K-Means:** [MiniBatchKMeans — scikit-learn 1.7.2 documentation](#)
- **K-Means vs Mini-Batch K-Means:** [Comparison of the K-Means and MiniBatchKMeans clustering algorithms](#)

Exercise 1: K-means clustering (45 minutes)

Use the K-means algorithm and Euclidean distance to cluster these 8 points into 3 clusters:

A1=(2,10), **A2**=(2,5), **A3**=(8,4), **A4**=(5,8), **A5**=(7,5), **A6**=(6,4), **A7**=(1,2), **A8**=(4,9).

The distance matrix based on the Euclidean distance is given below:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Suppose that the initial seeds (centers of each cluster) are **A1**, **A4**, and **A7**. Run the K-means algorithm for 1 epoch only. At the end of this epoch, show:

- a) The new clusters (i.e. the points belonging to each cluster)
- b) The centers of the new clusters
- c) Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.
- d) How many more iterations are needed to converge? Draw the result for each epoch.

Exercise 2: Hierarchical clustering (45 minutes)

Anime	Genre
Bleach	Action, Adventure, Fantasy
Demon Slayer	Action, Fantasy
Attack On Titan	Action, Drama, Mystery
Parasyte	Horror, Fantasy, Drama, Mystery

You are tasked with developing a recommendation system for an anime streaming application. The goal is to group anime series based on their genres using hierarchical clustering. Follow these steps to perform hierarchical clustering on the simplified dataset of anime series and their genres.

1. **Create a Distance Matrix:** Use the Jaccard distance to calculate the distance between each pair of anime series based on their genres.
 - *Hint: Jaccard Distance = 1 - Jaccard Similarity*
 - *Hint: Jaccard similarity is the count of positions where both vectors have a 1, divided by the count of positions where at least one vector has a 1.*
2. **Perform Hierarchical Clustering:** use the single linkage method to construct a hierarchical clustering tree, and update the distance matrix at each iteration.
3. **Visualize the Clustering:** Create a dendrogram to visualize the hierarchical clustering, including the splitting points.