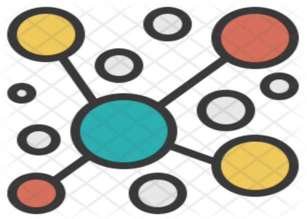
Data Mining

ENSIA 2025-2026

Lab sheet N°9: Clustering



Resources

Clustering

- **Clustering algorithms:** **2.3. Clustering — scikit-learn 1.7.2 documentation**
- **Distance metrics:** **cdist — SciPy v1.16.2 Manual**

Hierarchical clustering

- **Scipy**
  - **Hierarchical clustering (scipy.cluster.hierarchy) — SciPy v1.16.2 Manual**
  - **linkage — SciPy v1.16.2 Manual**
  - **dendrogram — SciPy v1.16.2 Manual**
- **ScikitLearn**
  - **AgglomerativeClustering — scikit-learn 1.7.2 documentation**
  - **Plot Hierarchical Clustering Dendrogram — scikit-learn 1.7.2 documentation**

DBSCAN clustering

- **ScikitLearn:** **DBSCAN — scikit-learn 1.7.2 documentation**

Clustering - Evaluation metrics

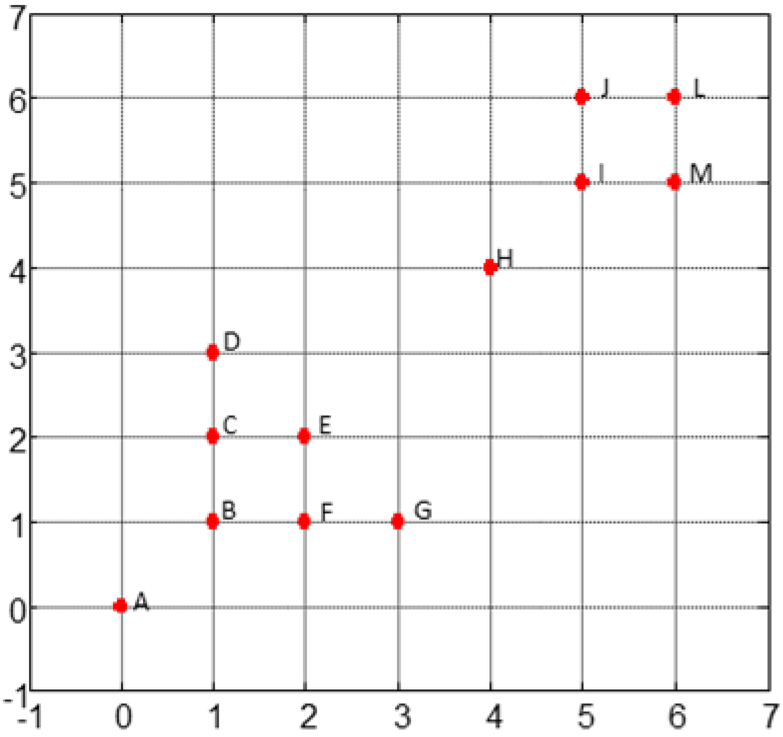
- **Clustering metrics:** **Clustering performance evaluation**
- **Silhouette Score:** **silhouette_score — scikit-learn 1.7.2 documentation**
- **Homogeneity score:** **homogeneity_score — scikit-learn 1.7.2 documentation**

Notebooks:

- **Clustering - Guided**
- **Clustering - Non-Guided**

Exercise: DBSCAN

Suppose we apply DBSCAN to cluster the following dataset using Euclidean distance.



Given that MinPts = 3 and EPS = 1, answer the following questions.

a) Compute the proximity matrix

b) Label all points as “core points,” “boundary points,” and “noise.”

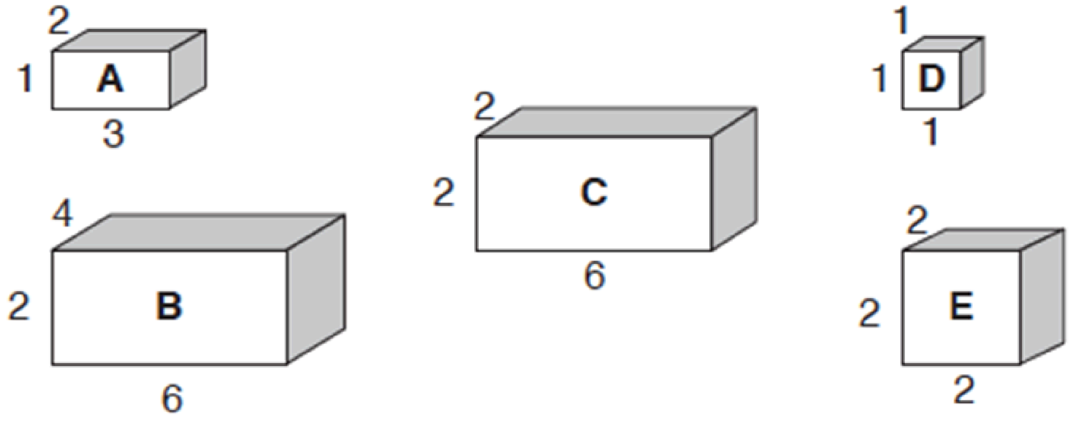
c) What is the clustering result?

d) Repeat the above two questions when epsilon = $\sqrt{10}$

e) What should be the value of MinPts and EPS to have two clusters, with no noise?

Exercise: Hierarchical clustering

Consider the five objects (A, B, C, D, and E) shown in the figure below. Each object has three features: length, width, and height. For example, the features of object A are (3,2,1).



(a) Suppose we apply the single link (MIN) algorithm to cluster the objects. Draw the dendrogram for the clusters assuming the similarity measure is Euclidean distance.

(b) Repeat the question in part (a) assuming that the similarity measure is correlation.

(c) Suppose we apply the complete link (MAX) algorithm to cluster the objects. Draw the dendrogram for the clusters assuming the similarity measure is Euclidean distance.

Exercise: Cluster Evaluation 1

The following table (confusion matrix) shows the clustering results in a land cover classification dataset that consists of many pieces of land. The number provided in the table is the number of objects (pieces of land) that are clustered into each cluster that belongs to each category. For example, the number in the forest column and cluster 1 row means that 10 forest items are clustered into cluster 1.

Table: Clustering results for land cover classification dataset.

	Forest	**Farm**	**Shrubland**	**Urban**	**Water**
Cluster 1	20	10	10	10	950
Cluster 2	400	100	400	50	50
Cluster 3	50	50	500	200	200
Cluster 4	200	250	150	200	200

Answer the following questions based on the table. No calculations are necessary.

1. Which cluster has the smallest entropy?
2. Which cluster has the biggest entropy?
3. Give a label name for each of the four clusters based on its land cover

Exercise: Cluster Evaluation 2

The following table (confusion matrix) shows the K-means clustering results for a land cover classification dataset that consists of many pieces of land. The number provided in the table is the number of objects (pieces of land) that are clustered into each cluster that belongs to each category. For example, the number in the forest column and cluster 1 row means that 10 forest items are clustered into cluster 1.

Table: K-means clustering results for land cover classification dataset

	Forest	**Farm**	**Shrubland**	**Urban**	**Water**
Cluster 1	10	100	20	10	3000
Cluster 2	3000	10	1000	10	0
Cluster 3	10	3000	500	150	200
Cluster 4	2000	2500	1500	3000	1400

Answer the following questions based on the table. No calculations are necessary.

1. Which cluster has the smallest entropy?
2. Which cluster has the biggest entropy?
3. Give a label name for each of the four clusters based on its land cover
4. Is this clustering result better than the one in the previous exercise?

Note: you can use the following Excel file: X Supervised_cluster_evaluation.xlsx