Clustering Summary

Renato R. Maaliw III, *DIT*
College of Engineering
Southern Luzon State University
Lucban, Quezon, Philippines

K-Means Clustering

When to use:

a. You have a relatively **large dataset**
b. The clusters are **spherical** or well-separated
c. You need a **quick**, efficient clustering method
d. The number of clusters, k, can be **reasonably** estimated
e. Your dataset is low-dimensional

K-Means Clustering

PROS:

a. Simple and fast
b. Works well on large dataset
c. Easy to interpret

CONS:

a. Requires **pre-determining** the number of clusters (k)
b. Sensitive to **outliers** and **initialization**
c. Can struggle with non-spherical or uneven clusters

DBSCAN

When to use:

a. You don’t know the number of clusters in advance
b. The clusters are **arbitrary shapes** or **non-spherical**
c. The dataset has **outliers** or **noise** that should be identified or excluded
d. The clusters have **varying densities**

DBSCAN

PROS:

a. Handles clusters of arbitrary shapes
b. Identifies and filters out noise (outliers)
c. No need to specify the number of clusters

CONS:

a. Struggles with varying densities and high-dimensional data
b. Sensitive to the choice of parameters (epsilon and min points)

Hierarchical Clustering

When to use:

a. You need a **dendrogram** for **hierarchical relationships** between clusters
b. You have a **small** to **medium-sized** dataset.
c. You don’t know the exact number of clusters in advance, but you want to **explore possible numbers**.
d. The data is **not too large** (hierarchical clustering is computationally expensive).

Hierarchical Clustering

PROS:

a. No need to specify the number of clusters in advance
b. Produces a **hierarchy** of clusters that can be visualized
c. Can use different **linkage criteria** (single, complete, average, ward)

CONS:

a. Computationally **inefficient** for large datasets
b. Noisy data can significantly affect the result
c. Choosing the **right cutoff** to determine clusters can be **subjective**

Gaussian Mixture Models

When to use:

a. You want to model clusters with soft assignments (e.g., a point can belong to multiple clusters with probabilities)
b. The data fits the assumption of **normally distributed** clusters (i.e., Gaussian components)
c. The clusters may be **overlapping**, and you want a probabilistic interpretation.

Gaussian Mixture Model

PROS:

a. Provides **probabilistic** clustering.
b. More **flexible** than K-Means as it can handle **elliptical clusters**.
c. Suitable for **soft clustering** scenarios.

CONS:

a. Requires specifying the number of clusters in advance.
b. Prone to overfitting, especially for high-dimensional data.
c. Computationally more **expensive** than K-Means.

Other Clustering Techniques

A. Mean Shift Clustering

- finds density peaks, for unknown clusters but computationally expensive

B. Affinity Propagation

- identifies clusters with exemplars without specifying k, but is memory intensive.

Other Clustering Techniques

C. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

- memory-efficient for **large datasets**; used for hierarchical clustering.

D. OPTICS (Ordering Points to Identify Clustering Structure)

- similar to DBSCAN but better handling varying densities

Other Clustering Techniques

D. **Spectral Clustering**
- handles complex shapes, suitable for graph-like structures

E. **Self-Organizing Maps (SOM)**
- uses neural networks to visualize high-dimensional data

F. **Fuzzy C-Means Clustering**
- soft clustering with probabilistic cluster assignment

Thank you very much for listening.