McMaster University ENGINEERING

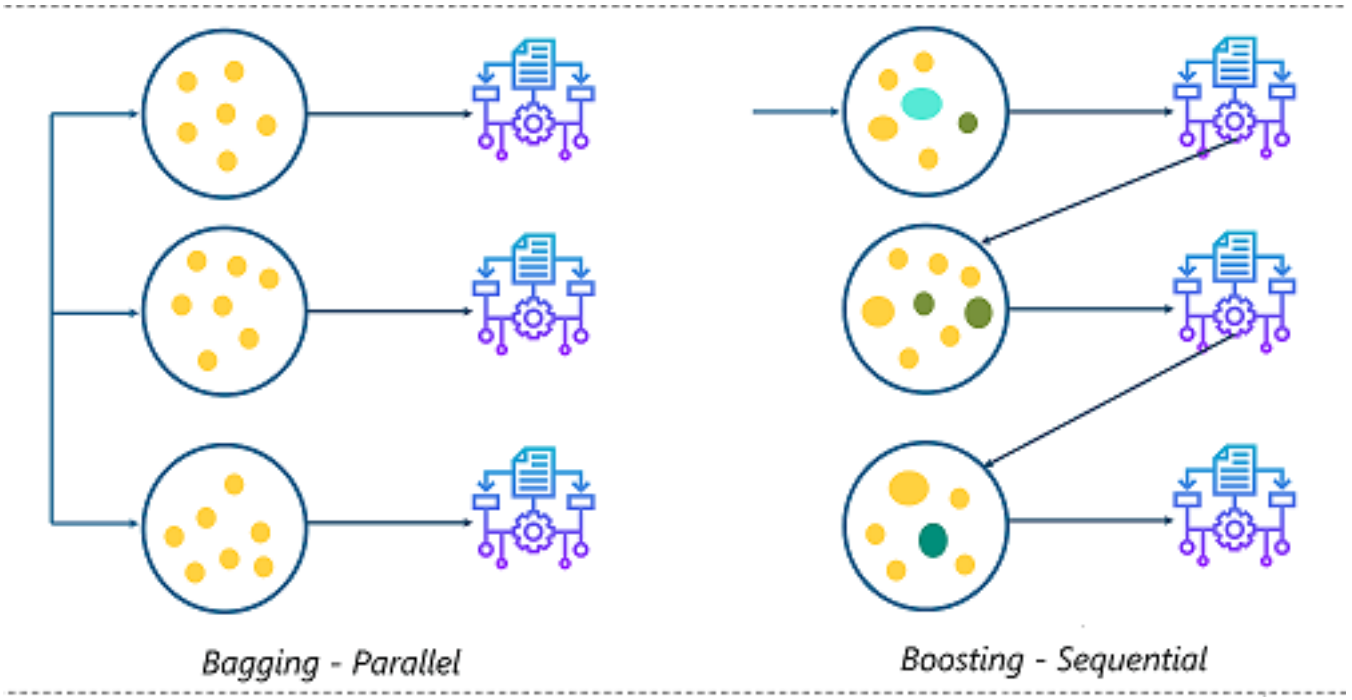
W BOOTH SCHOOL OF ENGINEERING PRACTICE AND TECHNOLOGY

SEP 785: Machine Learning

Lecture 10: Clustering

Instructor: Dr. Dalia Mahmoud, PhD
(Mechanical Engineering, McMaster University)
Email: mahmoudd@mcmaster.ca

Recap



Intended Learning Outcome

- Explain the unsupervised paradigm.
- Explain the motivation and potential applications of clustering.
- Introduction to K-means clustering.
- Broadly explain how DBSCAN works.
- Explain the idea of hierarchical clustering.
- Recognize the impact of distance measure and representation in clustering methods.

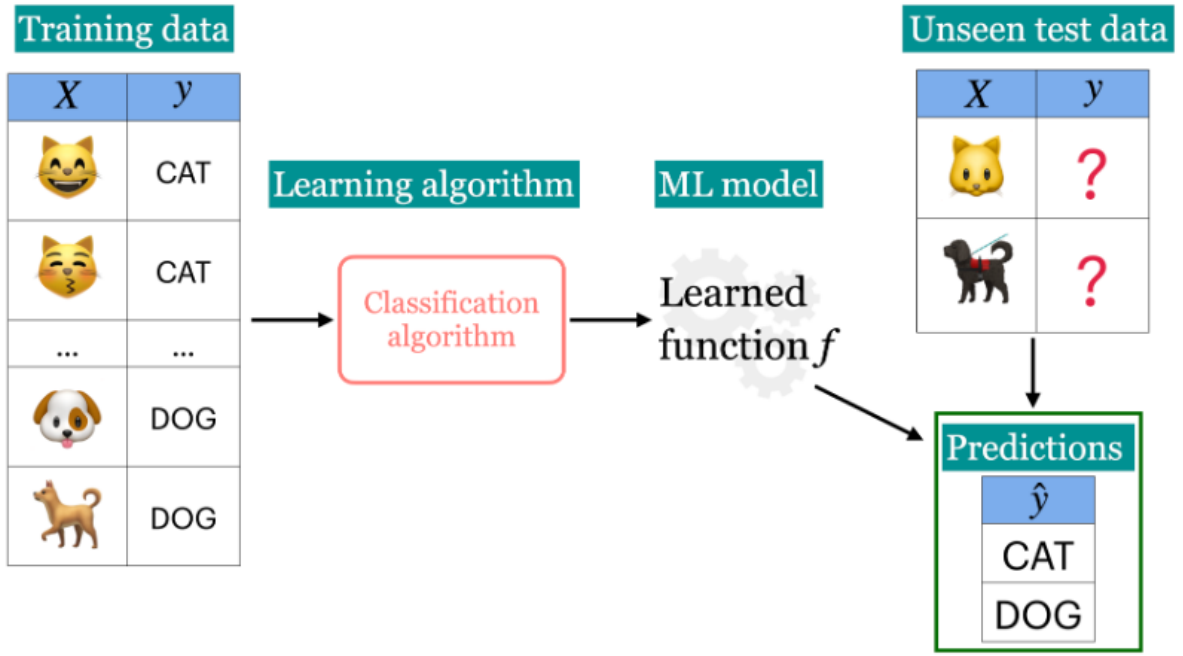
Contents

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Introduction
- K-Means Clustering
- DBSCAN
- Hierarchical Clustering
- Summary

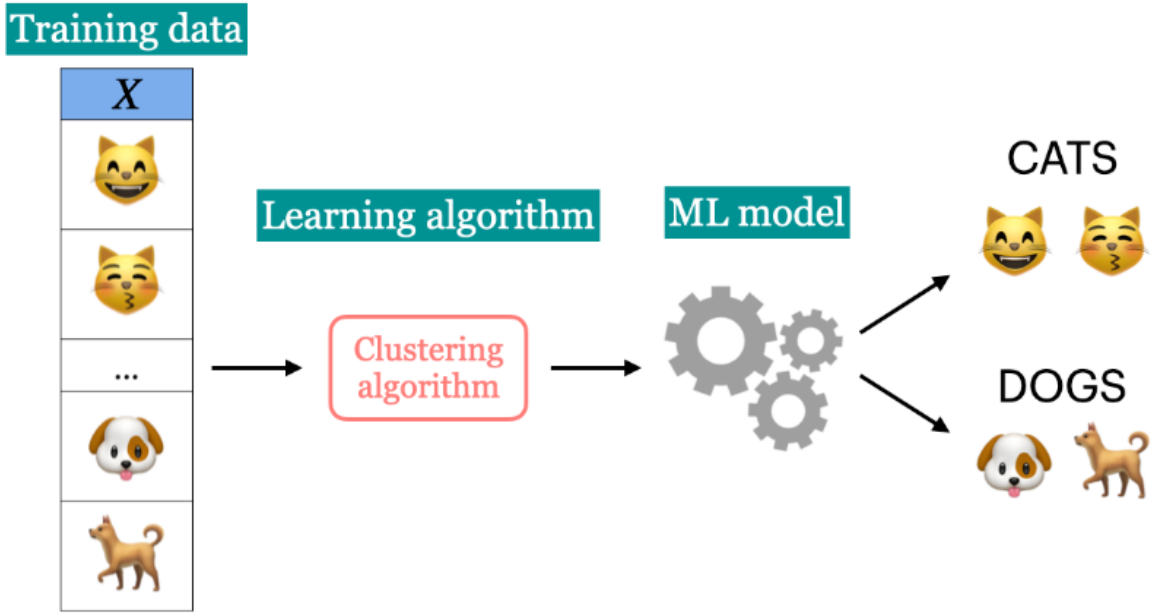
Supervised Learning

- Training data comprises a set of observations (X) and their corresponding targets (y).
- We wish to find a model function f that relates X to y.
- Then use that model function to predict the targets of new examples.



Unsupervised Learning

- Training data consists of observations (X) without any corresponding targets.
- Unsupervised learning could be used to group similar things together in X or to find underlying structure in the data.



When to use Unsupervised Learning

- **No Labels Available** → You don't have labeled data but need to find patterns.
- **Discover Hidden Groups** → Example: Customer segmentation in marketing.
- **Reduce Data Complexity** → Example: PCA for high-dimensional datasets.
- **Find Relationships in Data** → Example: Market basket analysis (Amazon, Netflix).
- **Detect Anomalies** → Example: Fraud detection in banking.
- **Preprocess Data for Supervised Learning** → Example: Feature selection before training a model.

Types of Unsupervised Learning

- **Clustering** → Groups similar data points
  K-Means (Customer Segmentation)
  DBSCAN (Anomaly Detection)
  Hierarchical Clustering (Taxonomy Classification)
- **Dimensionality Reduction** → Reduces data complexity
  PCA (High-Dimensional Data Visualization)
  t-SNE (Word Embeddings)
- **Association Rule Learning** → Identifies relationships between items
  Apriori (Market Basket Analysis)
  FP-Growth (Product Recommendation)
- **Anomaly Detection** → Detects unusual patterns
  Isolation Forest (Fraud Detection)
  One-Class SVM (Cybersecurity Intrusion)

Types of Unsupervised Learning

Scenario	Unsupervised Learning Technique	Example
Grouping similar data	Clustering (K-Means, DBSCAN)	Customer Segmentation
Reducing data size	PCA, t-SNE	Image compression
Finding hidden patterns	Association Rules	Market Basket Analysis
Detecting outliers	Isolation Forest	Fraud Detection

Contents

- Introduction
- K-Means Clustering
- DBSCAN
- Hierarchical Clustering
- Summary

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

What is clustering?

- Clustering is the task of partitioning the dataset into groups called clusters based on their similarities.
- The goal of clustering is to discover underlying groups in a given dataset such that:
  - examples in the same group are as similar as possible;
  - examples in different groups are as different as possible.

What is clustering?

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Which of the following grouping of emoticons is the "correct" grouping?

Data

Categorization 1

Group 1 (cats)

Group 2 (humans)

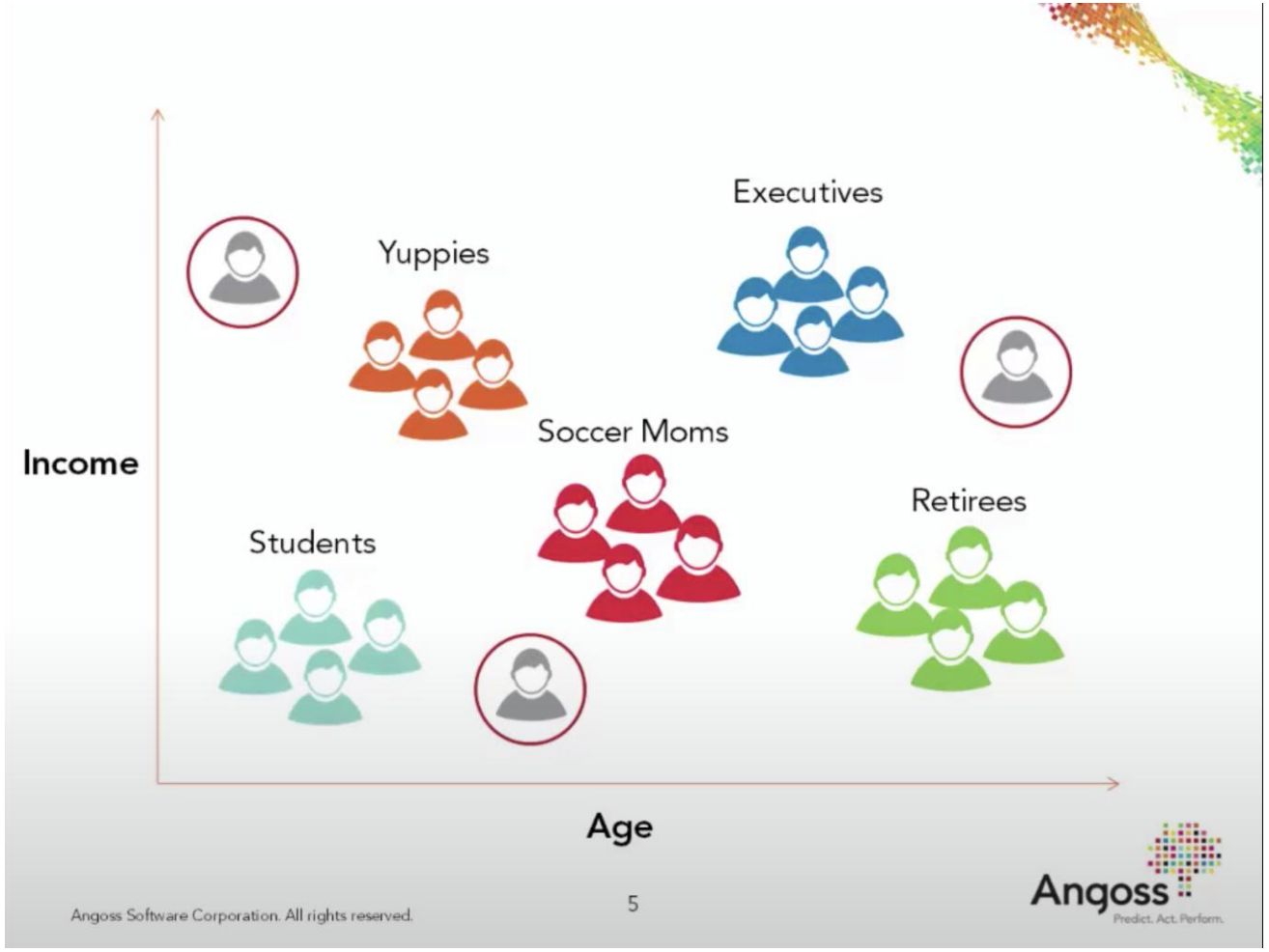
Categorization 2

Group 1 (happy)

Group 2 (unhappy)

Application: Customer Segmentation

Understand landscape of the market in businesses and craft targeted business or marketing strategies tailored for each group.



Application: Document clustering

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Grouping articles on different topics from different news sources. For example

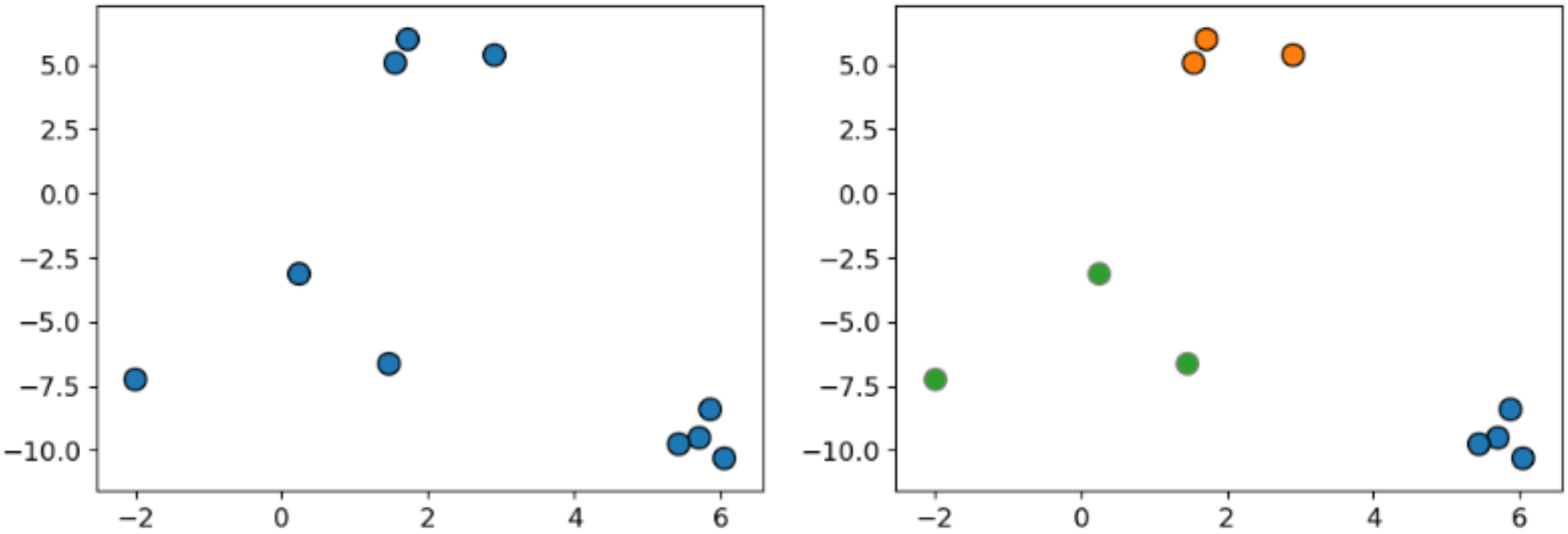
Armed man who broke into Trudeau residence charged with threatening to kill or injure PM

The Guardian · 1 hour ago

- **Corey Hurren, alleged Rideau Hall intruder, threatened Trudeau: RCMP officer**
  Global News · 4 hours ago
- **Corey Hurren had multiple firearms, uttered threat against Trudeau, court documents allege**
  CBC.ca · 2 hours ago
- **Man arrested near Rideau Hall had several weapons, threatened PM Trudeau: RCMP**
  CTV News · 22 minutes ago

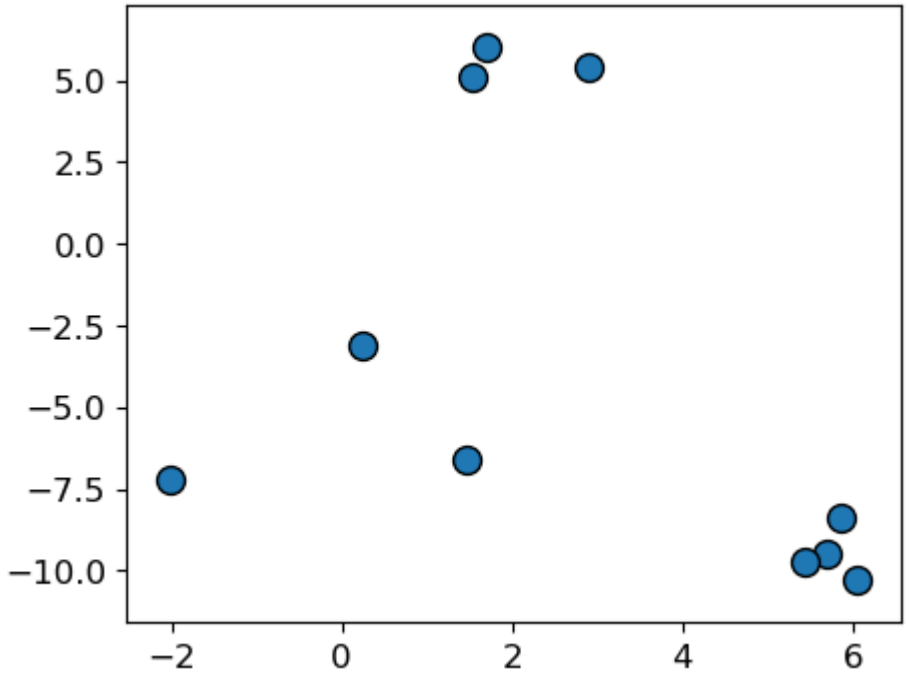
Clustering: Input and (possible) output

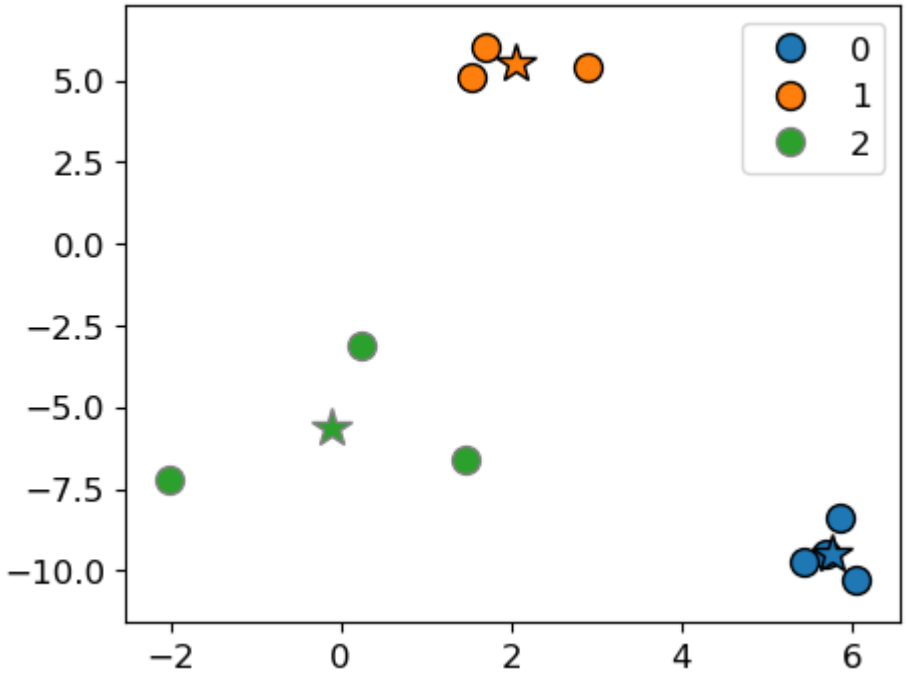

```
X, y = make_blobs(n_samples=10, centers=3, n_features=2, random_state=10)
fig, axes = plt.subplots(1, 2, figsize=(12, 4))
mglearn.discrete_scatter(X[:, 0], X[:, 1], ax = axes[0]);
mglearn.discrete_scatter(X[:, 0], X[:, 1], y=y, markers='o', ax = axes[1]);
```



K-Means algorithm: Main idea

- An unsupervised learning algorithm that partitions data into K clusters.
- Each cluster has a centroid, and points are assigned based on proximity.
- Iterative optimization: Reassign points & update centroids until convergence.





K-Means algorithm: Main idea

Input: Data points X and the number of clusters K

Initialization: K initial centers for the clusters

Iterative process:

repeat

- Assign each example to the closest center.
- Estimate new centers as *average* of observations in a cluster.

until **centers stop changing** or **maximum iterations have reached**.

K-Means algorithm: Input


```
n_examples = X.shape[0]
print("Number of examples: ", n_examples)
X
```

```
Number of examples:  10
```

```
array([[  5.69192445,  -9.47641249],
       [  1.70789903,   6.00435173],
       [  0.23621041,  -3.11909976],
       [  2.90159483,   5.42121526],
       [  5.85943906,  -8.38192364],
       [  6.04774884, -10.30504657],
       [ -2.00758803,  -7.24743939],
       [  1.45467725,  -6.58387198],
       [  1.53636249,   5.11121453],
       [  5.4307043 ,  -9.75956122]])
```

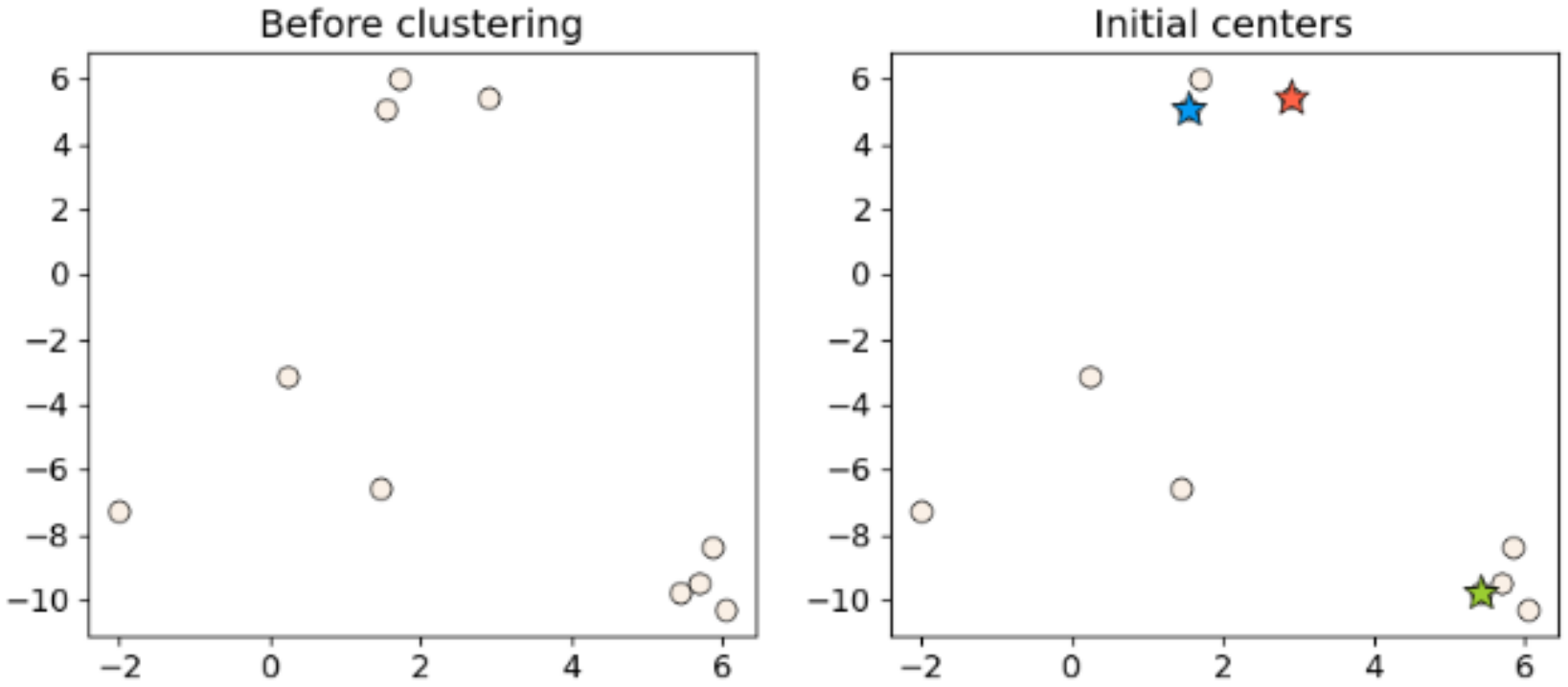
- Let K (number of clusters) be 3.

```
k = 3
```

K-Means algorithm: Initialization

Random initialization for K initial centers of the clusters.

```
np.random.seed(seed=3)
centers_idx = np.random.choice(range(0, n_examples), size=k)
centers = X[centers_idx]
plot_km_initialization(X, centers)
```



K-Means algorithm: Iterative Process

repeat

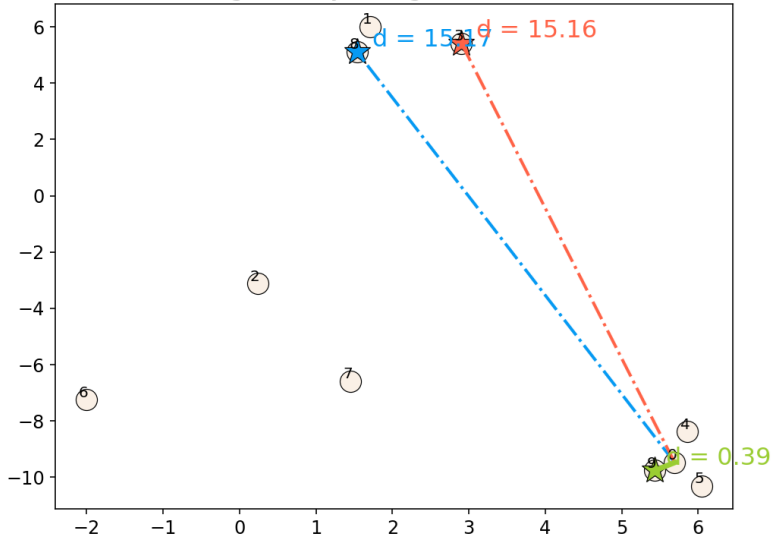
- Assign each example to the closest center. (update_Z)
- Estimate new centers as average of observations in a cluster. (update_centers)

until centers stop changing or maximum iterations have reached.

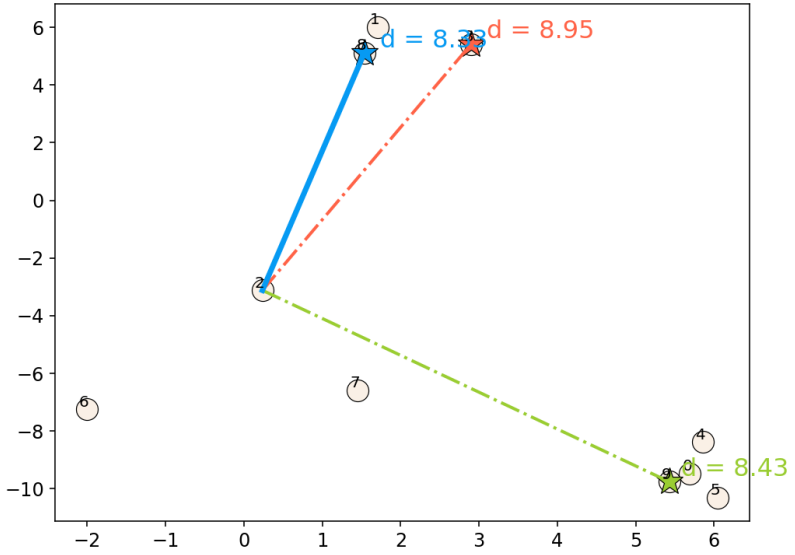
K-Means algorithm: Iterative Process

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

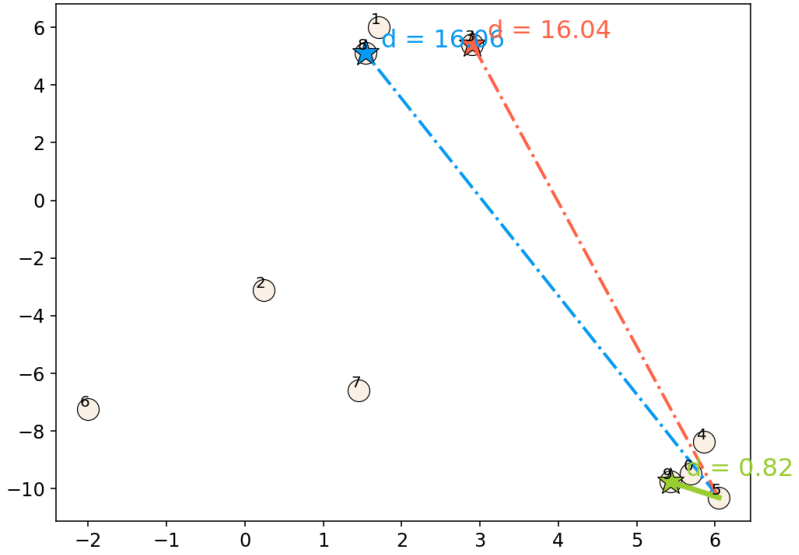
Point 0 will be assigned to yellowgreen cluster (min dist = 0.39)



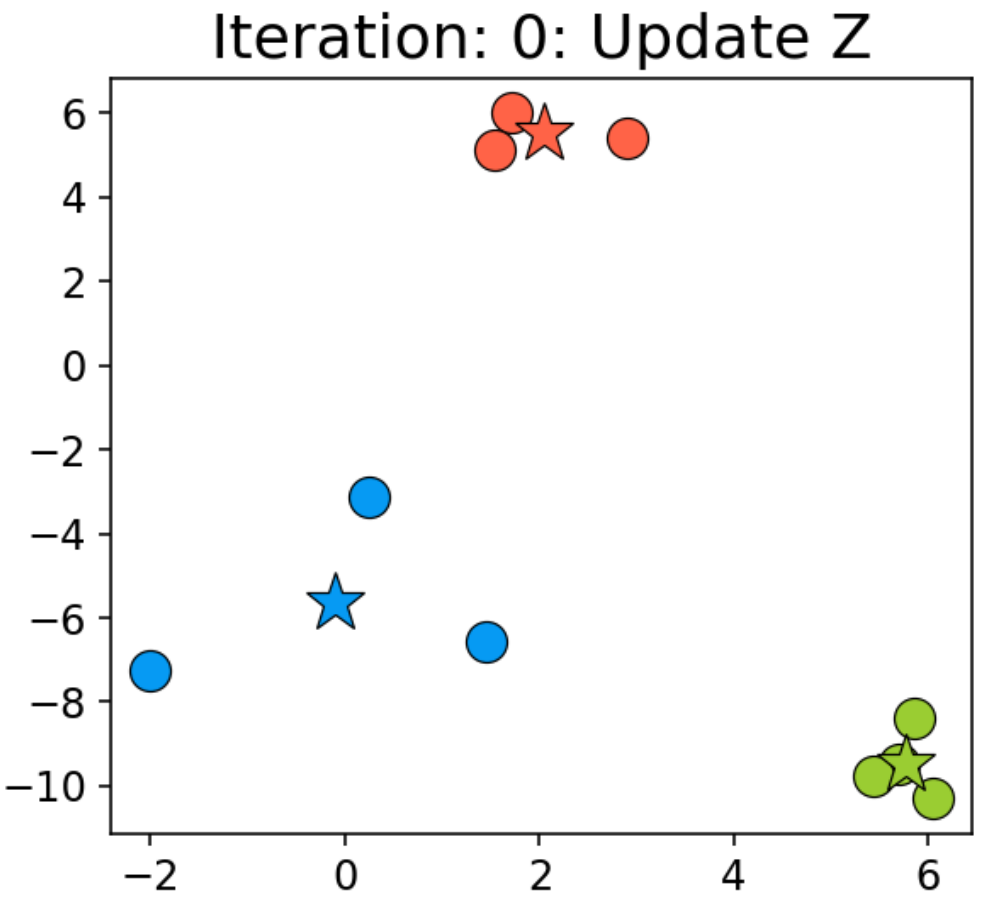
Point 2 will be assigned to xkcd:azure cluster (min dist = 8.33)

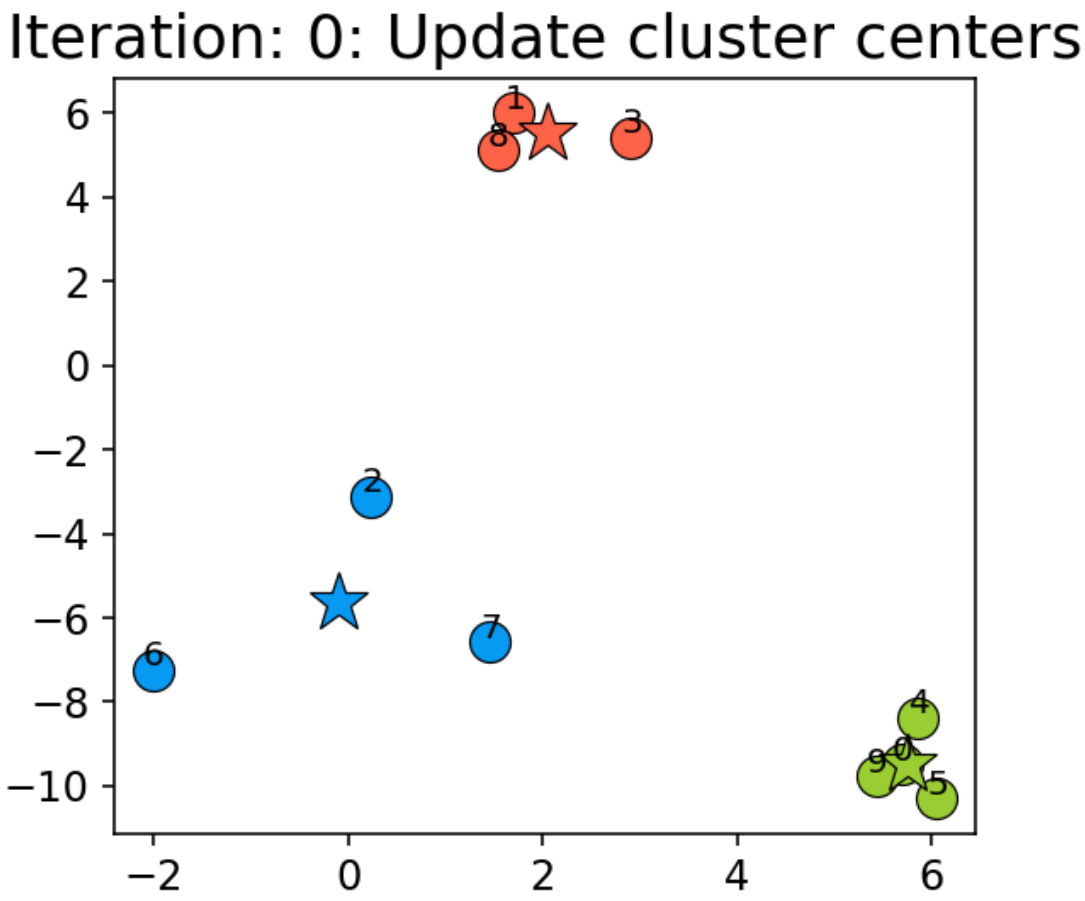


Point 5 will be assigned to yellowgreen cluster (min dist = 0.82)

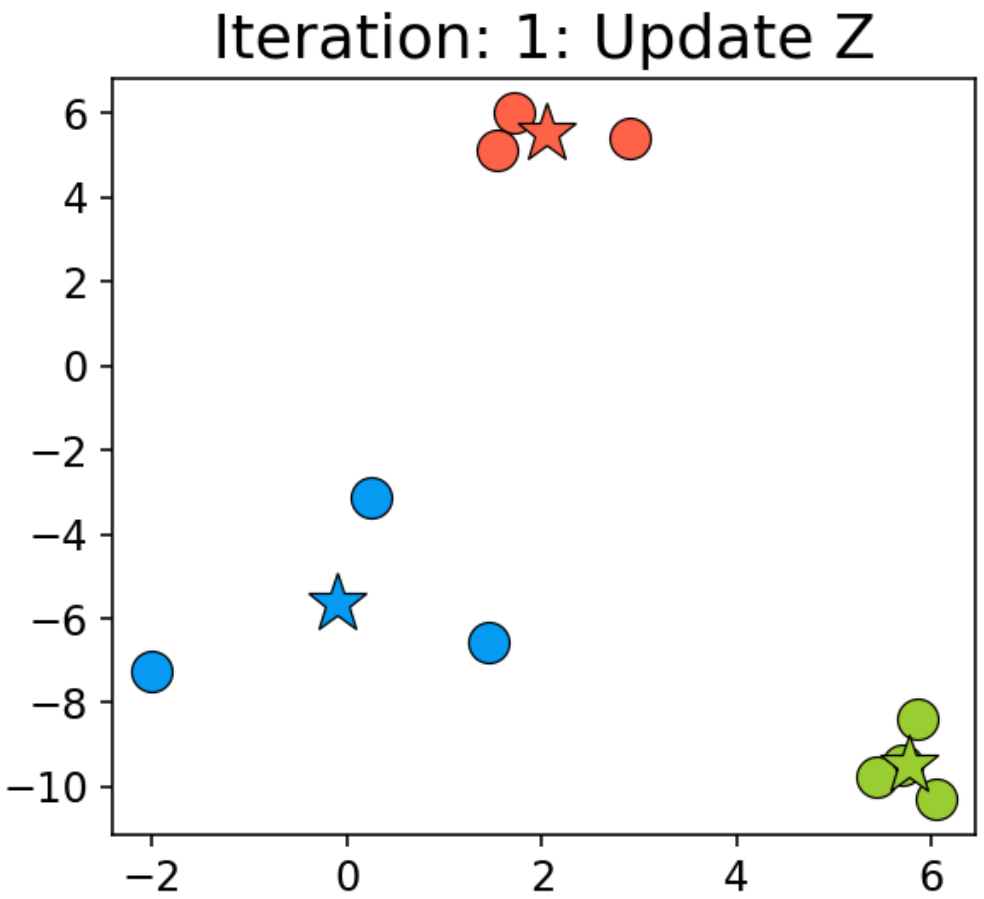


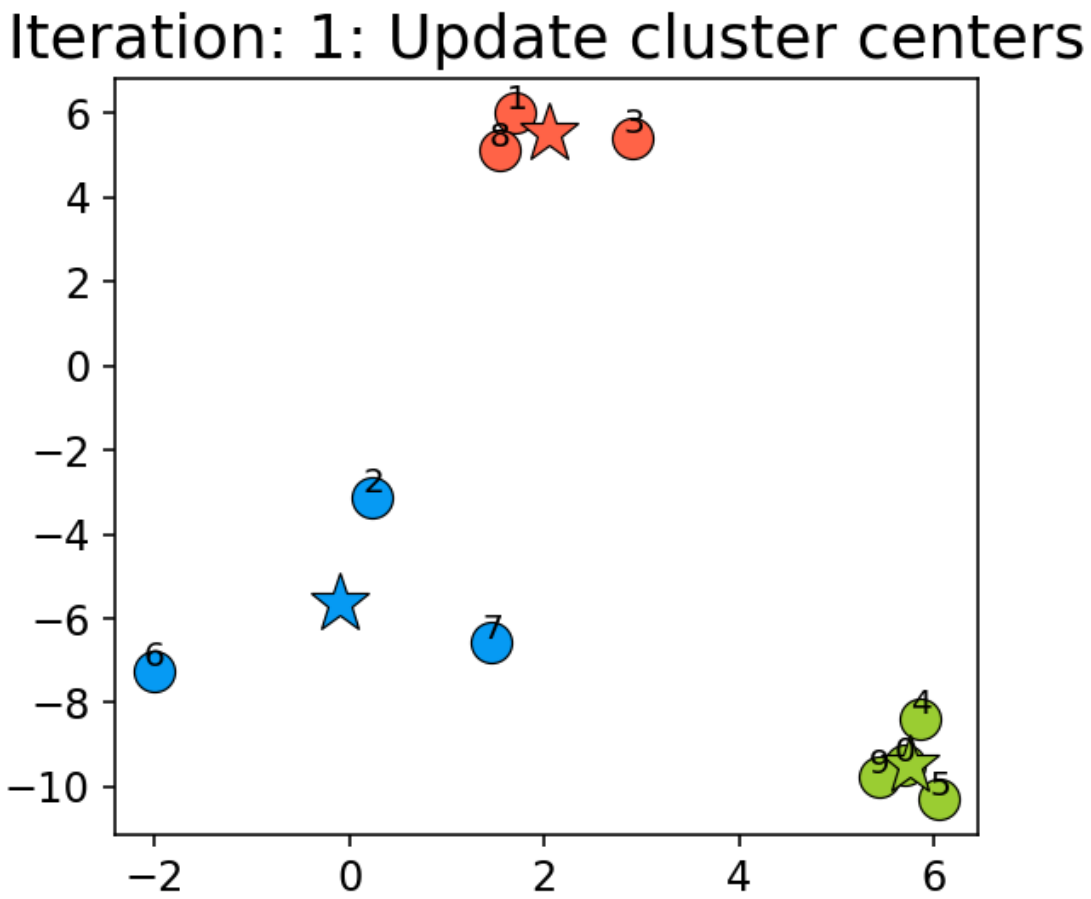
K-Means algorithm: Iterative Process



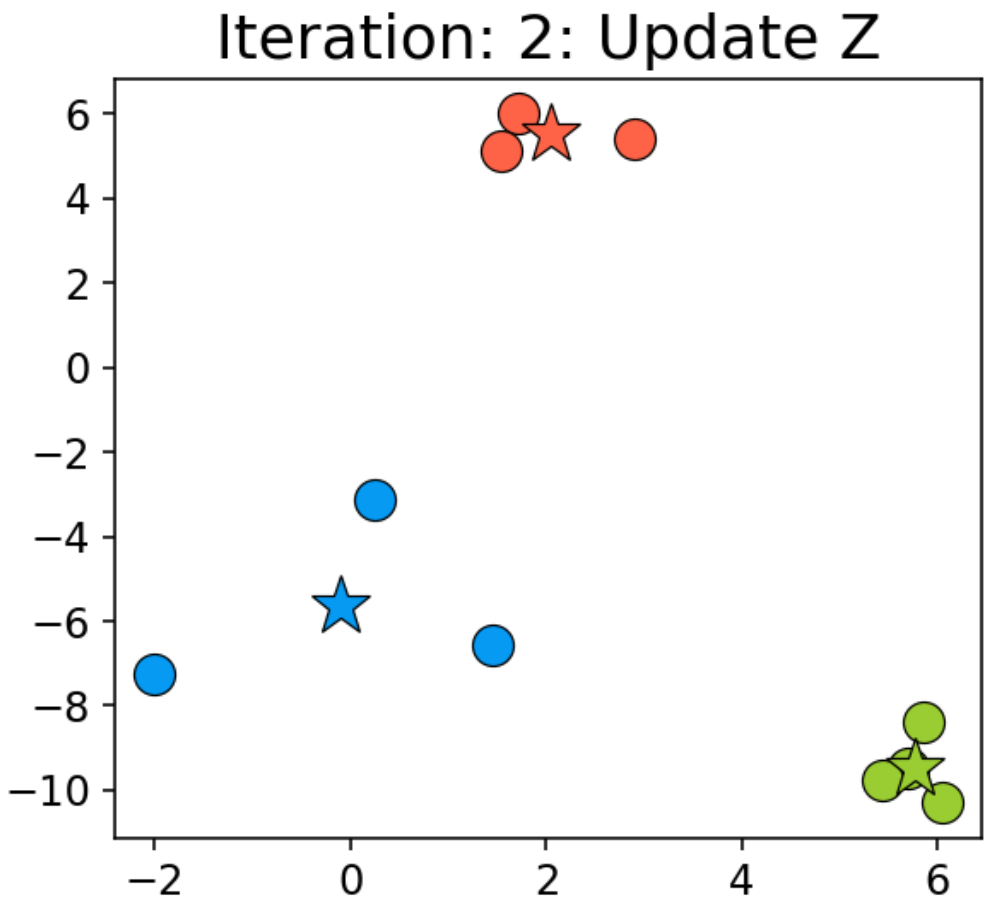


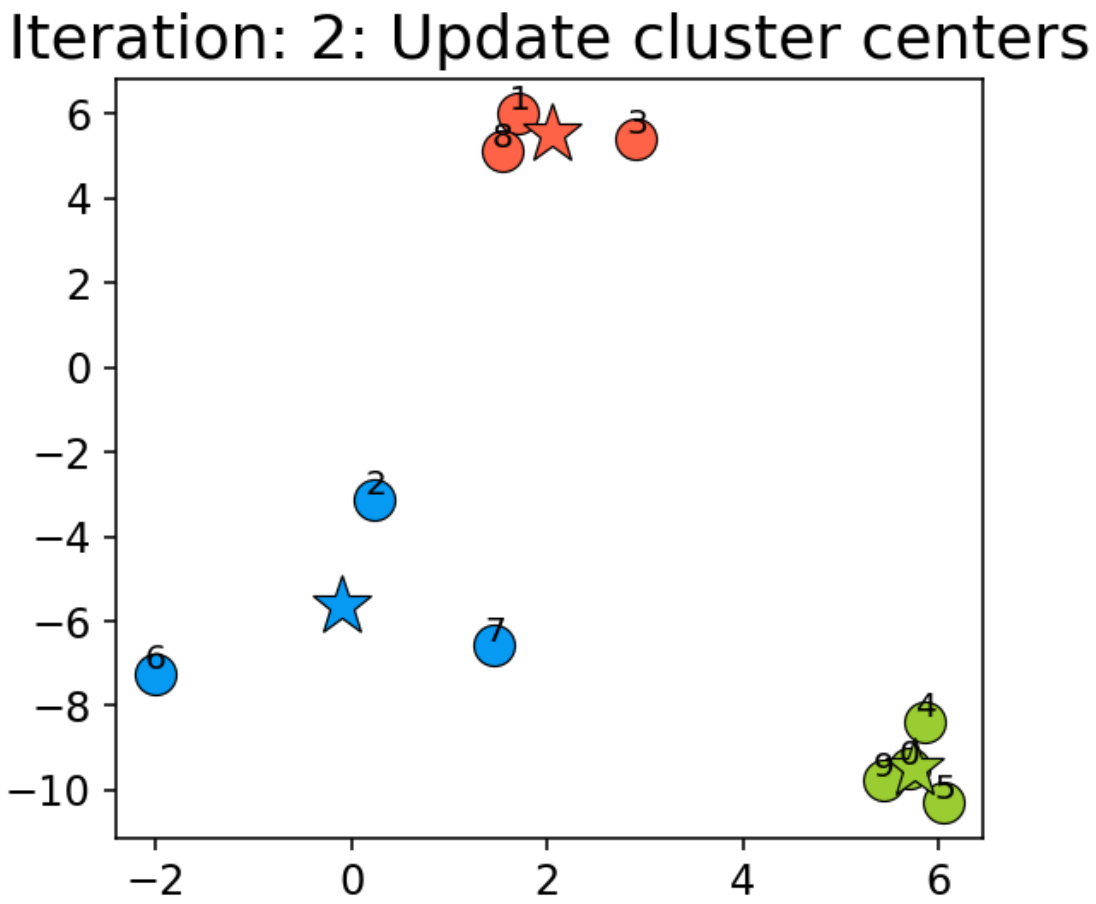
K-Means algorithm: Iterative Process



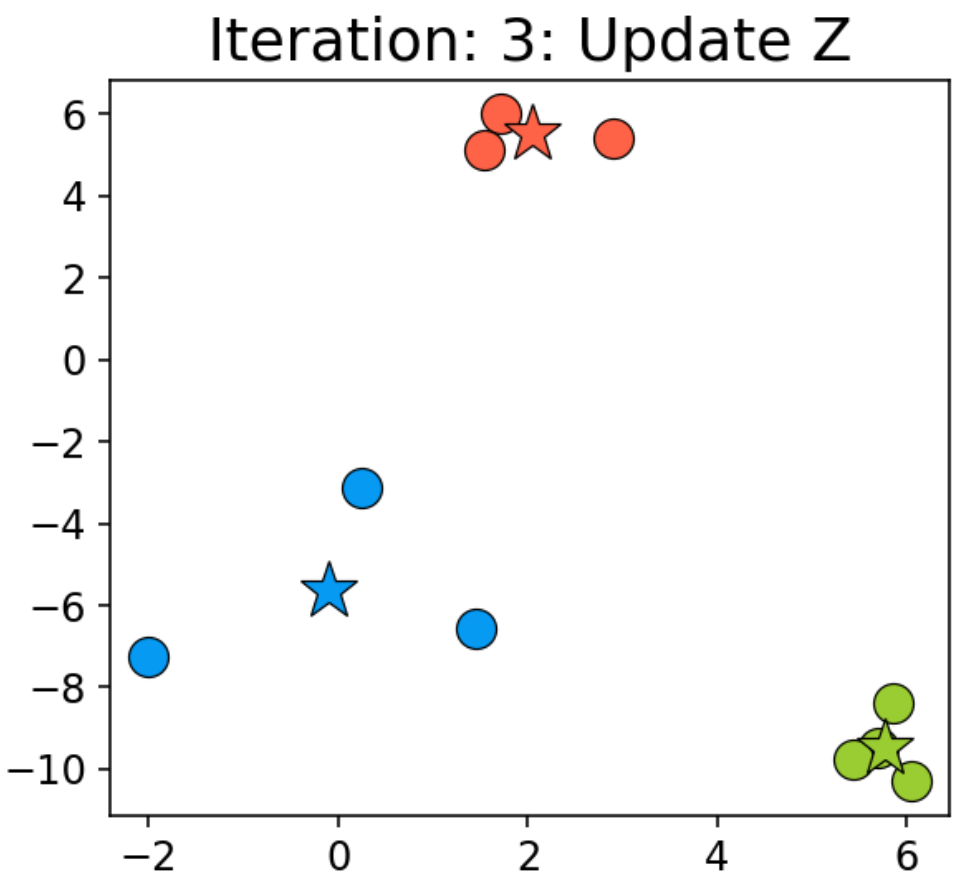


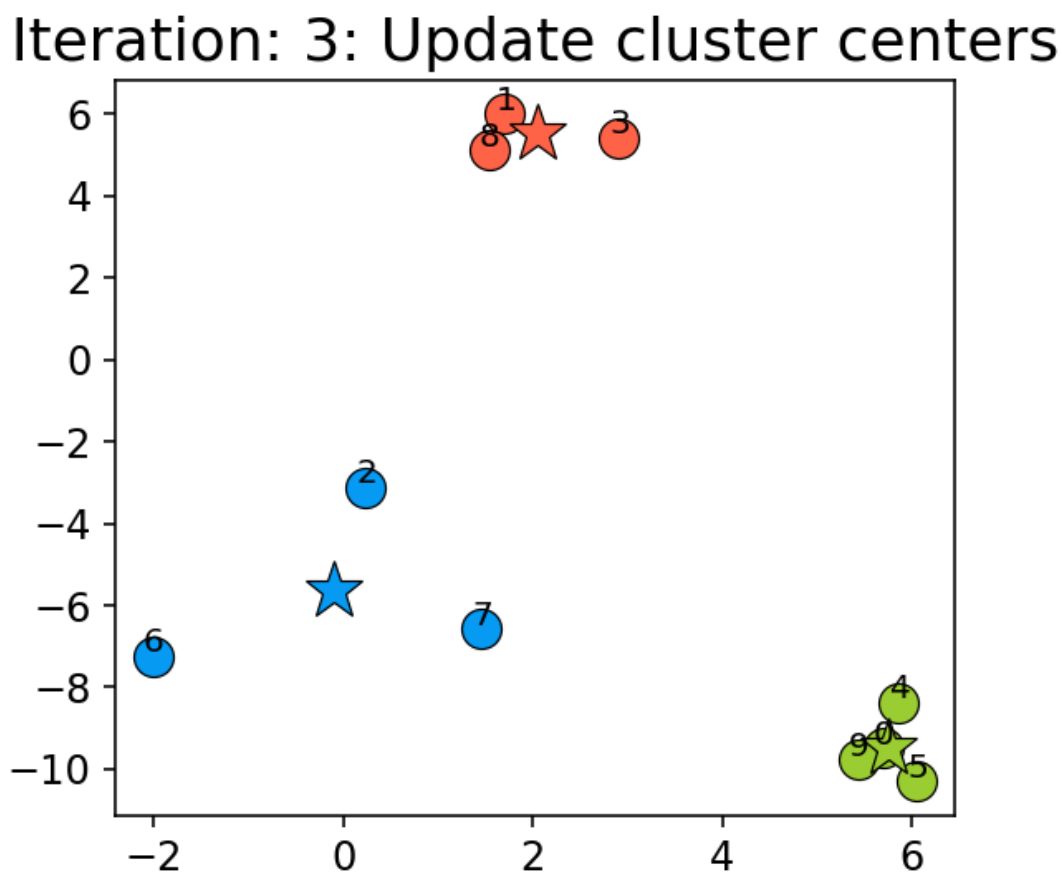
K-Means algorithm: Iterative Process





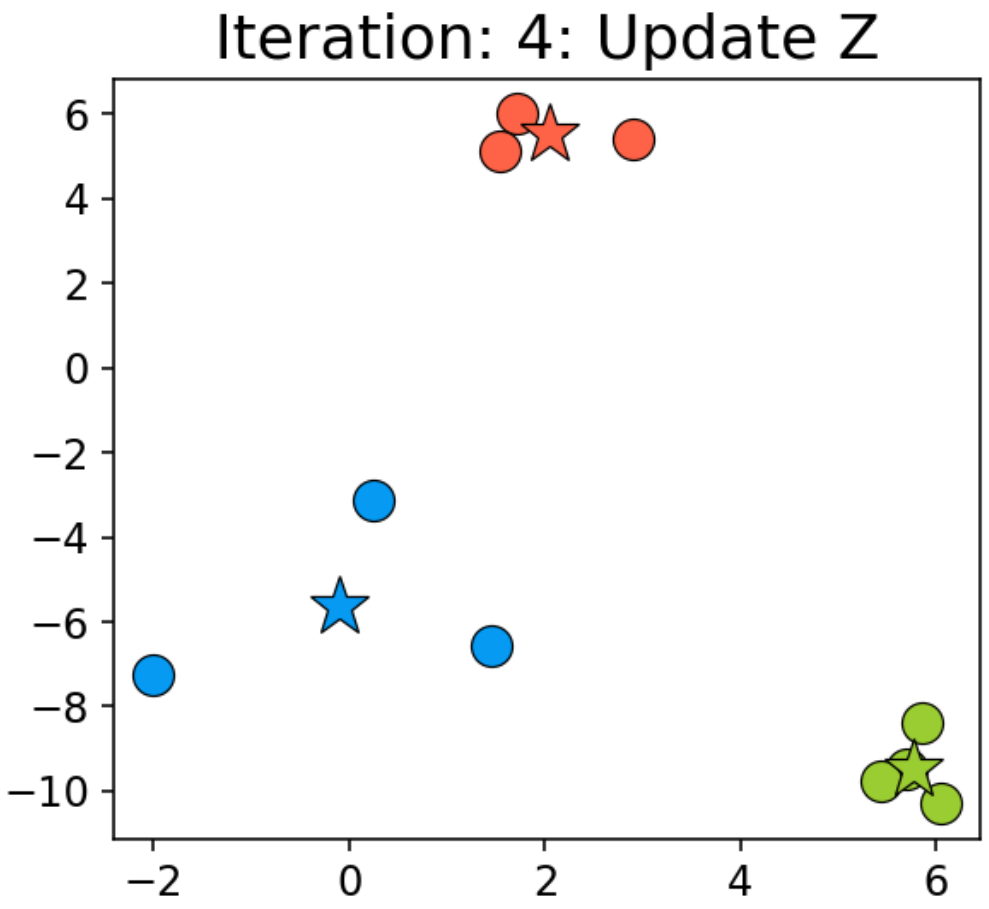
K-Means algorithm: Iterative Process

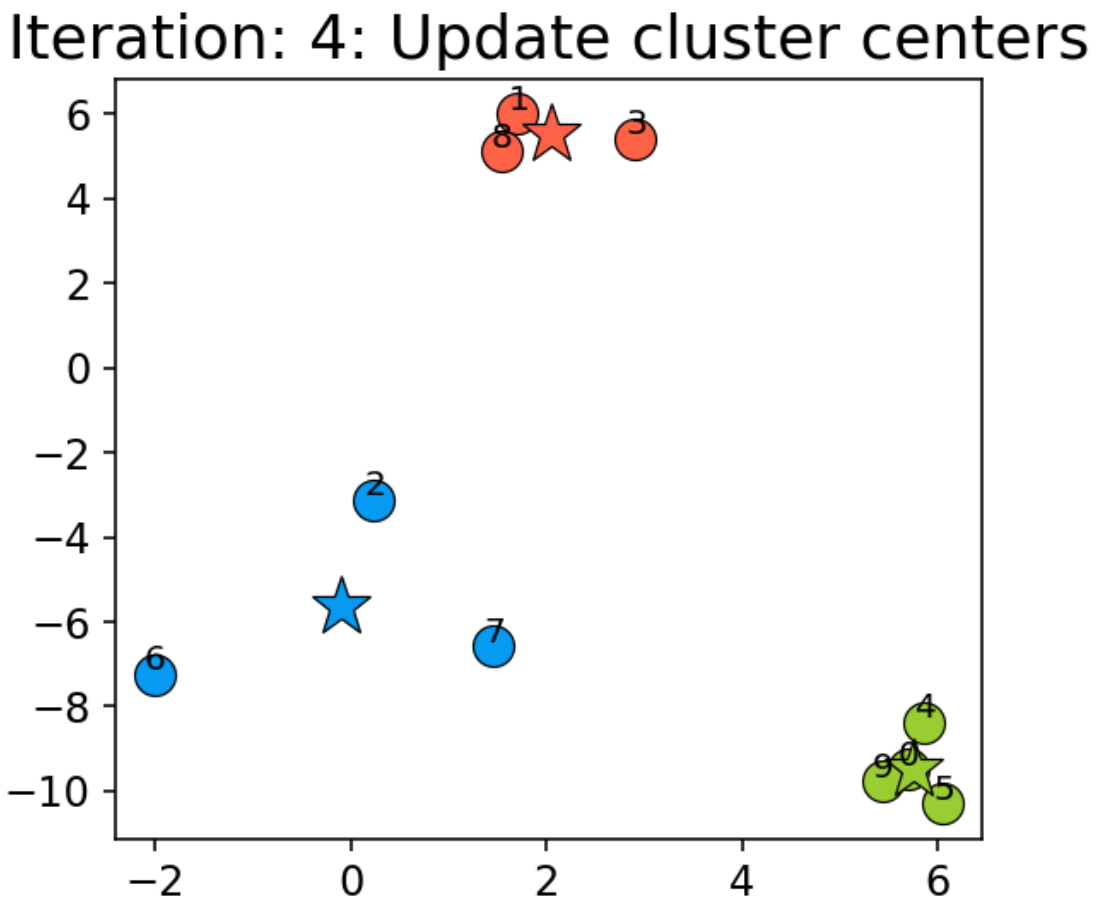




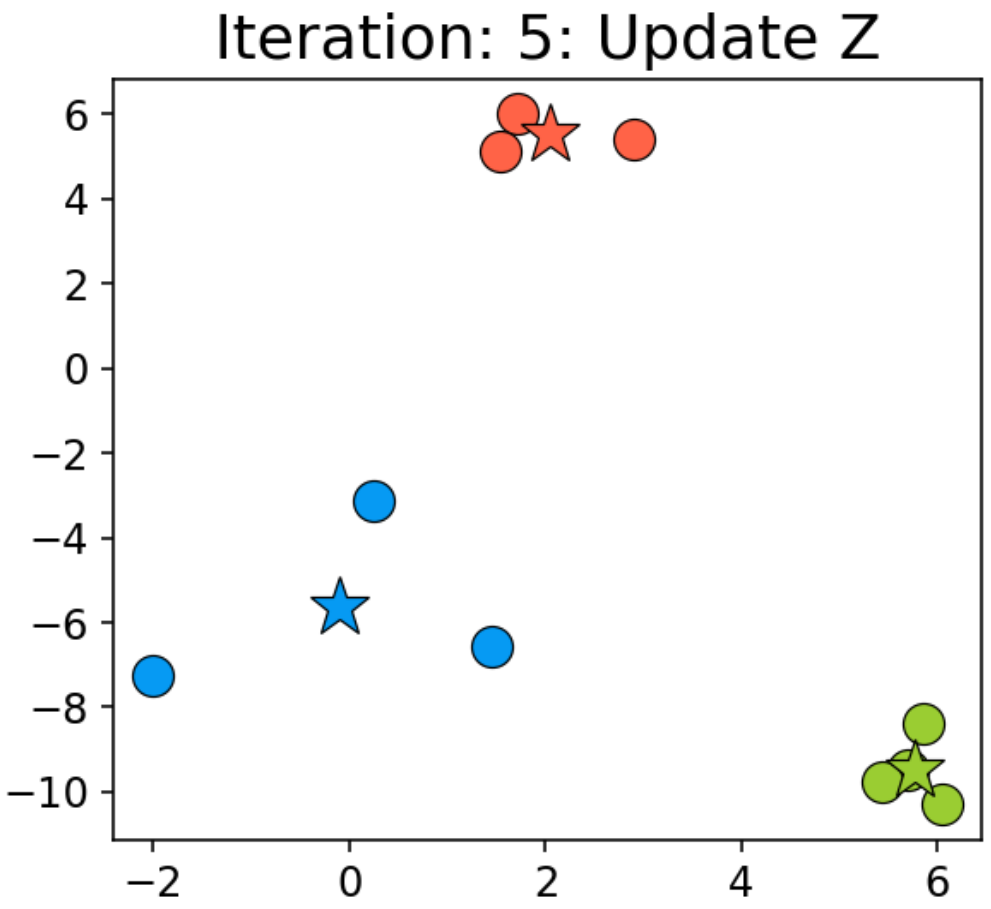
K-Means algorithm: Iterative Process

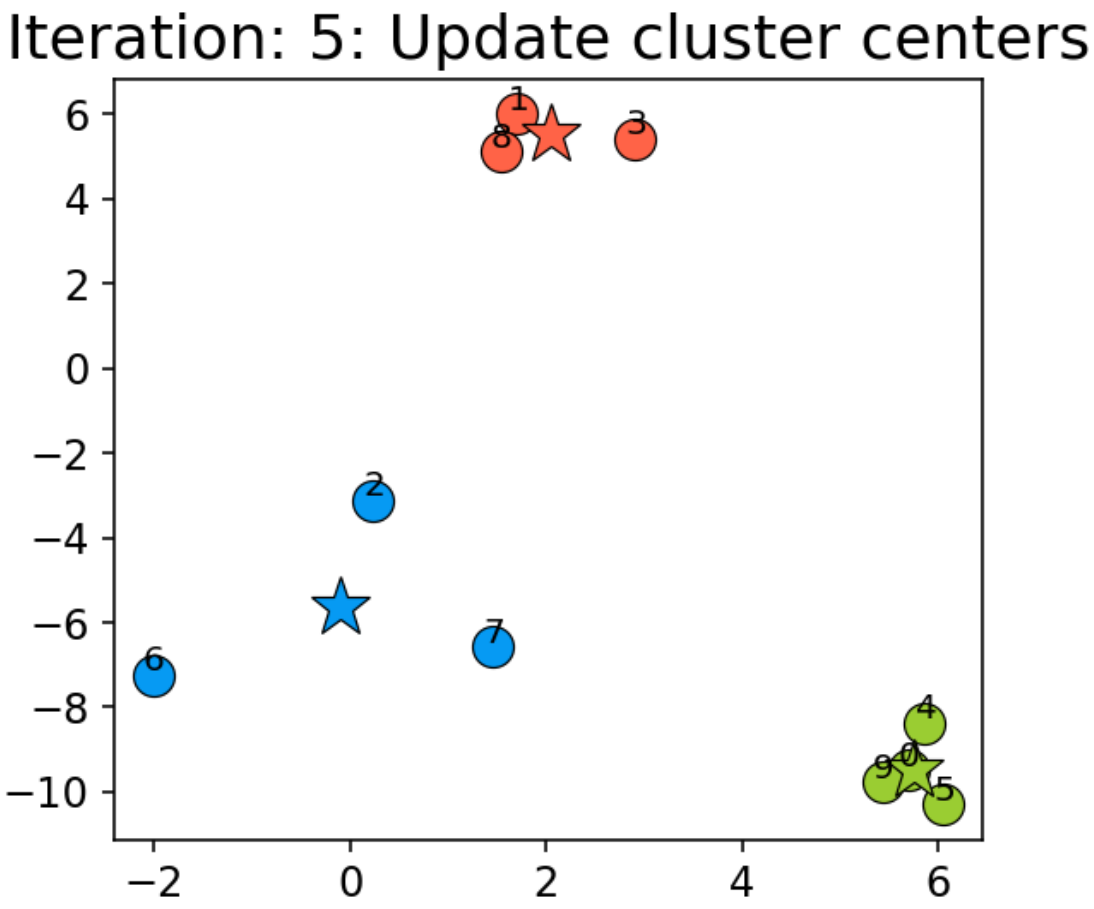
McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology



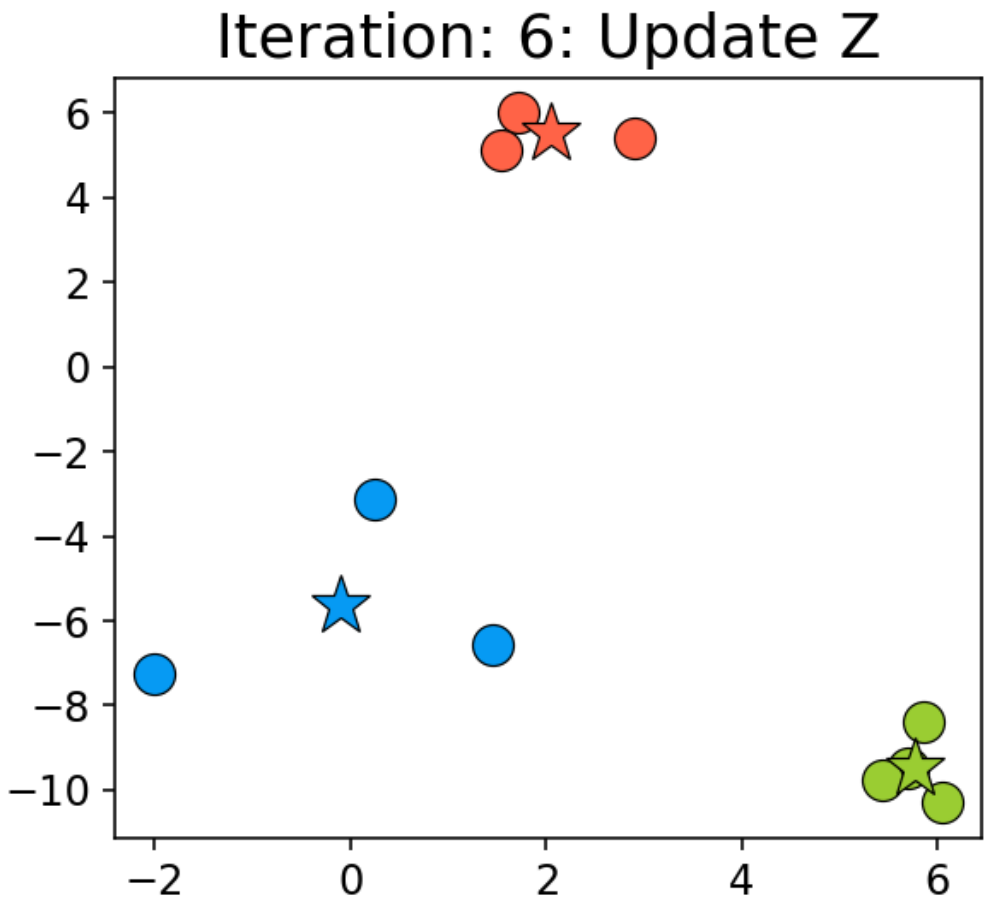


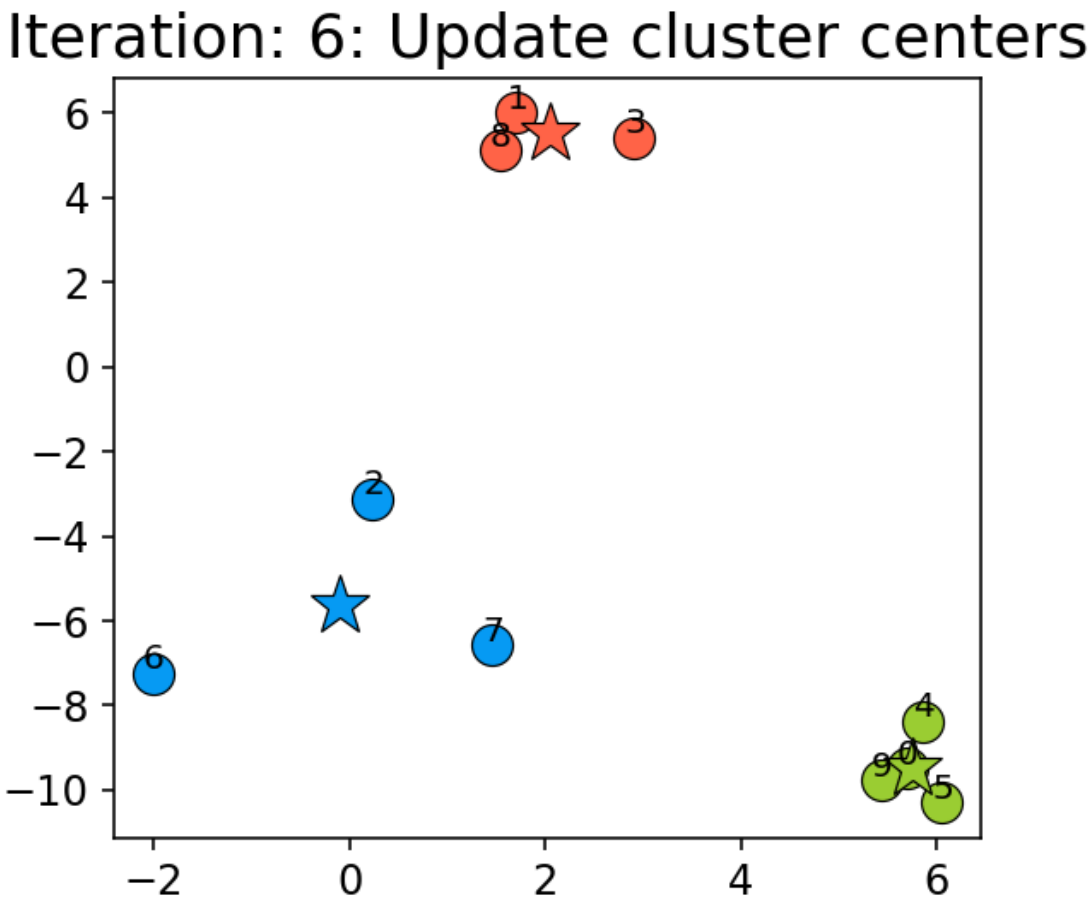
K-Means algorithm: Iterative Process





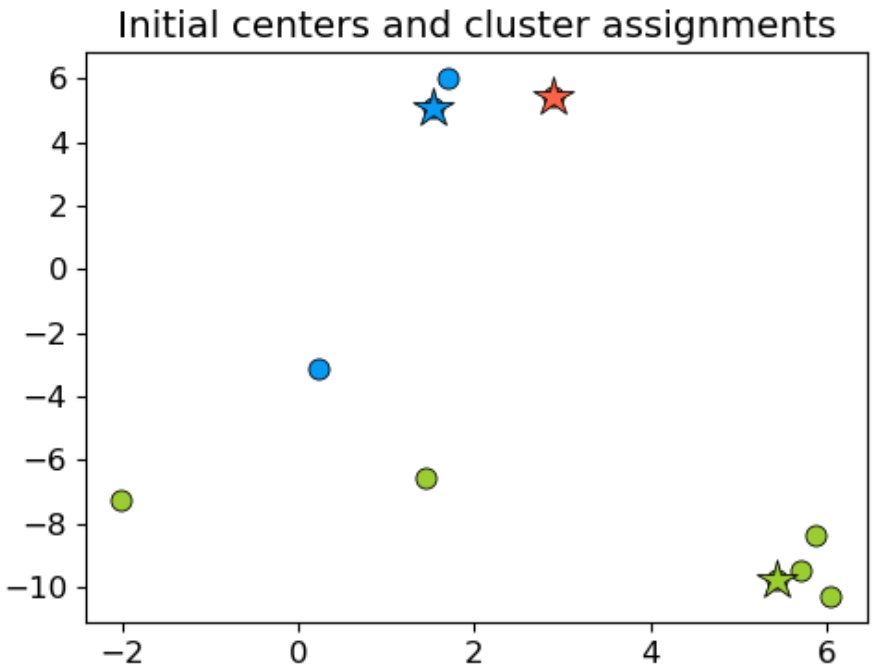
K-Means algorithm: Iterative Process

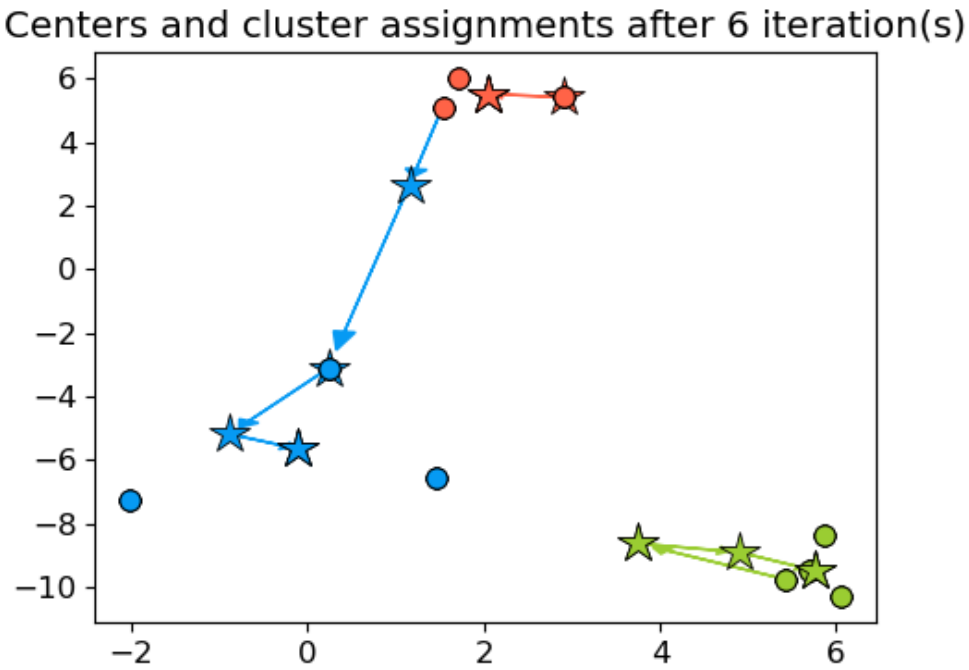




When to stop?

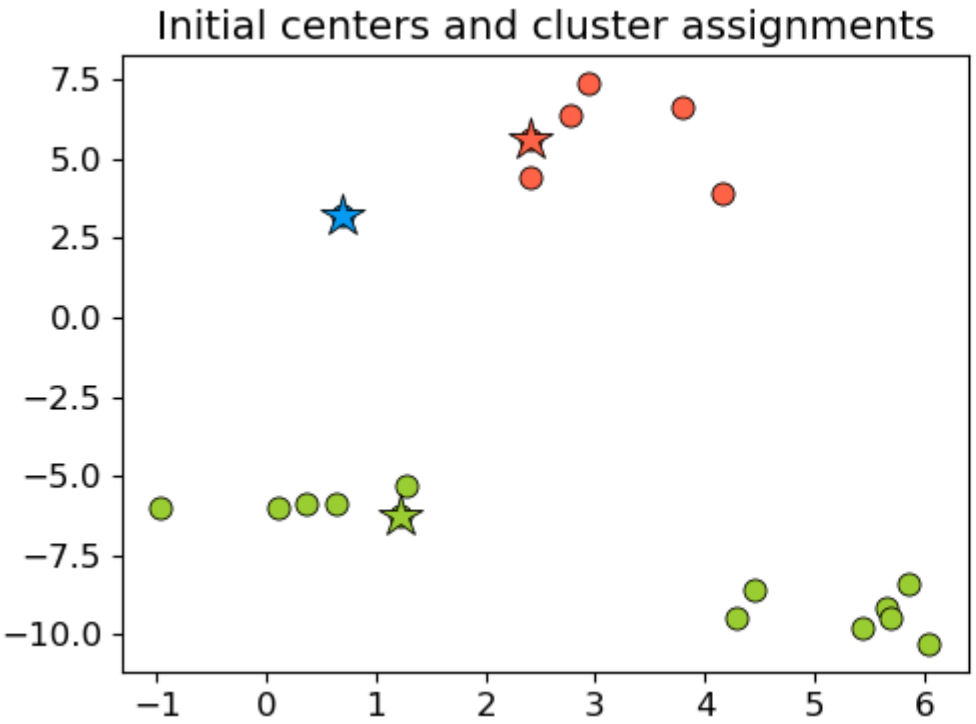
- Seems like after iteration 4 our centroids aren't changing anymore.
- The algorithm has converged. So we stop!
- K-Means always converges. It doesn't mean it finds the "right" clusters. It can converge to a sub-optimal solution.

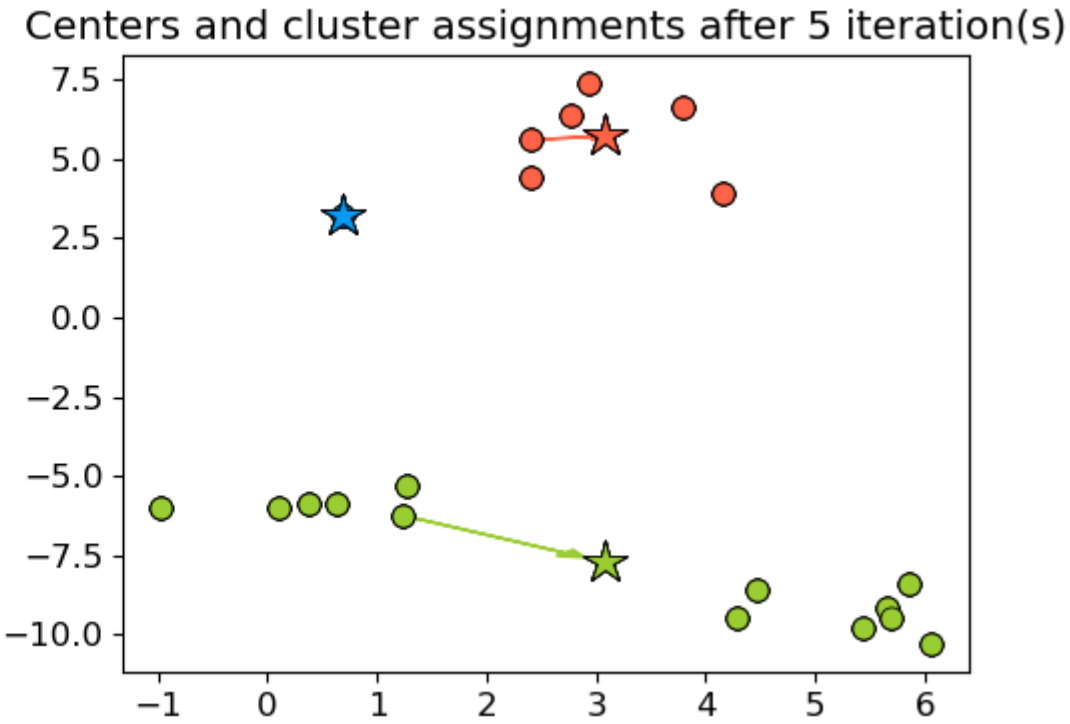




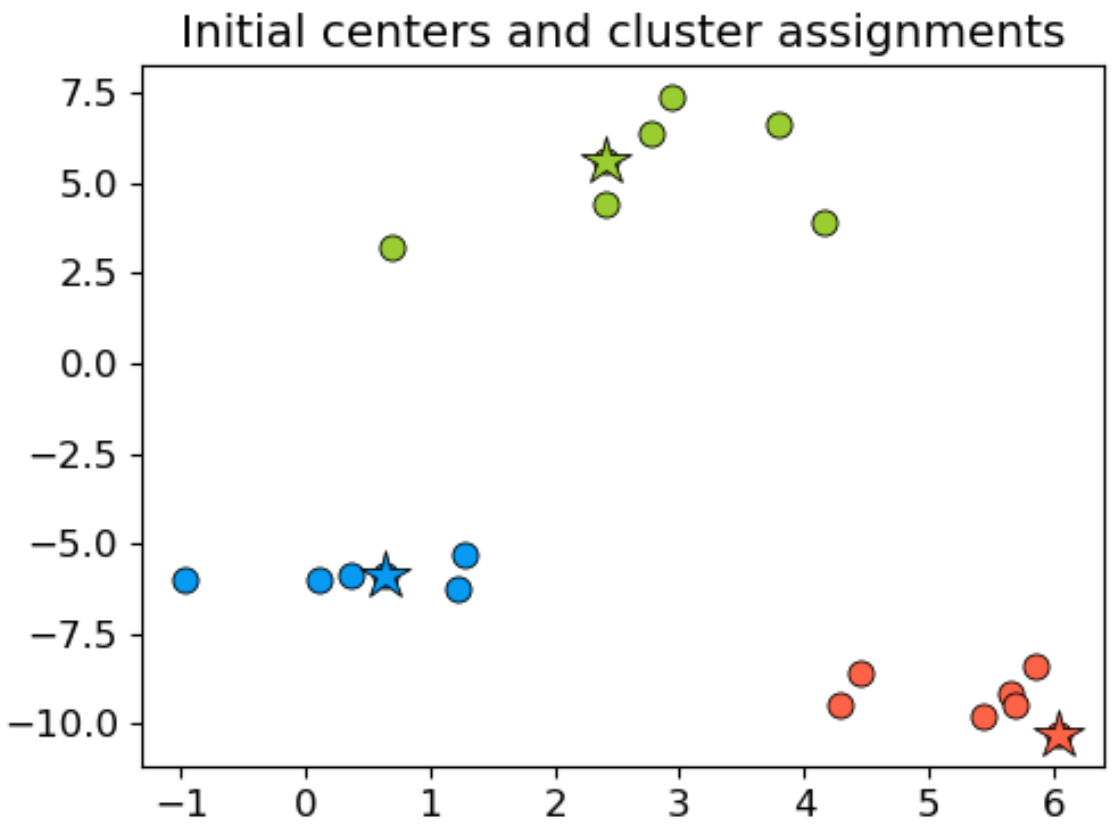
Initialization Step

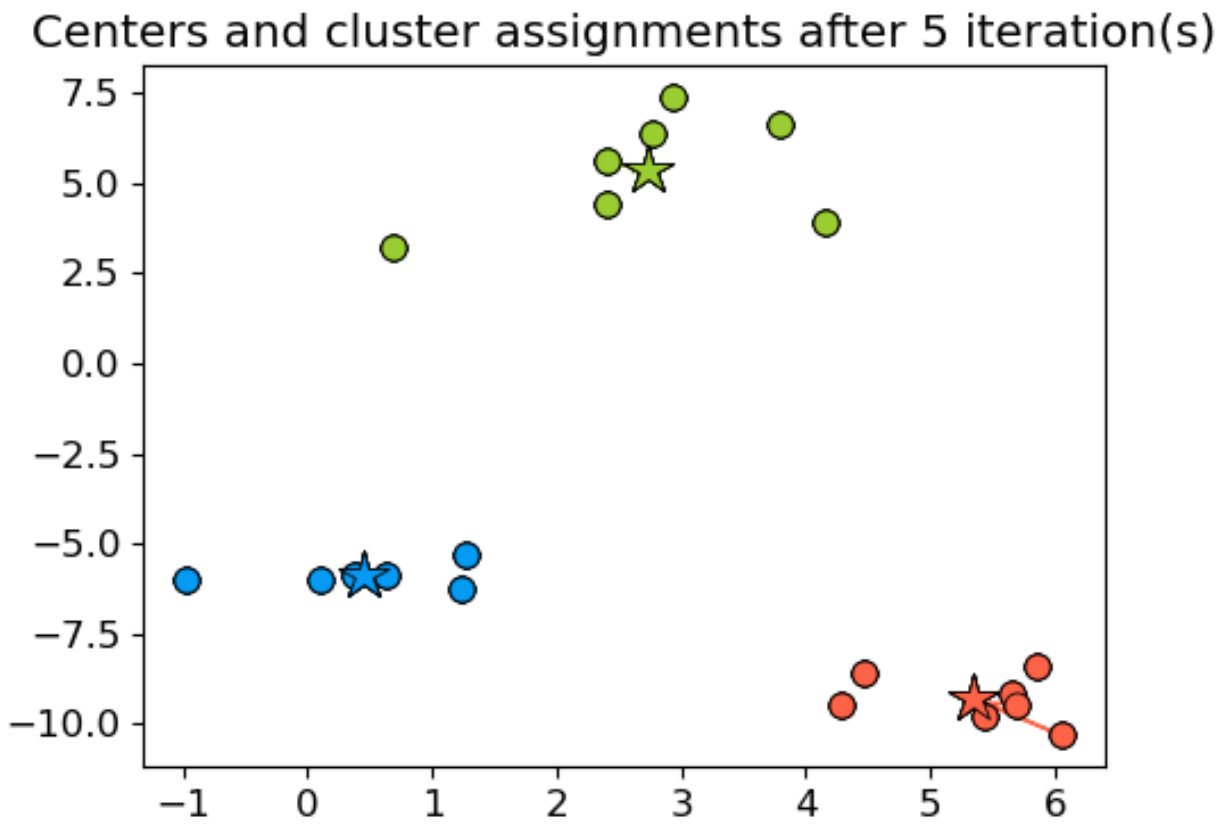
- The initialization step in K-Means is crucial because it determines the starting cluster centers, which can affect the algorithm's convergence and final clustering quality.





Initialization Step





Hyperparameter Tuning for K-Means Clustering

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Since K-Means is an **unsupervised algorithm**, it doesn't have labeled data to guide learning. However, we can optimize its **main hyperparameter: the number of clusters (k)**.
  - Elbow Method
  - Silhouette Score

Elbow Method

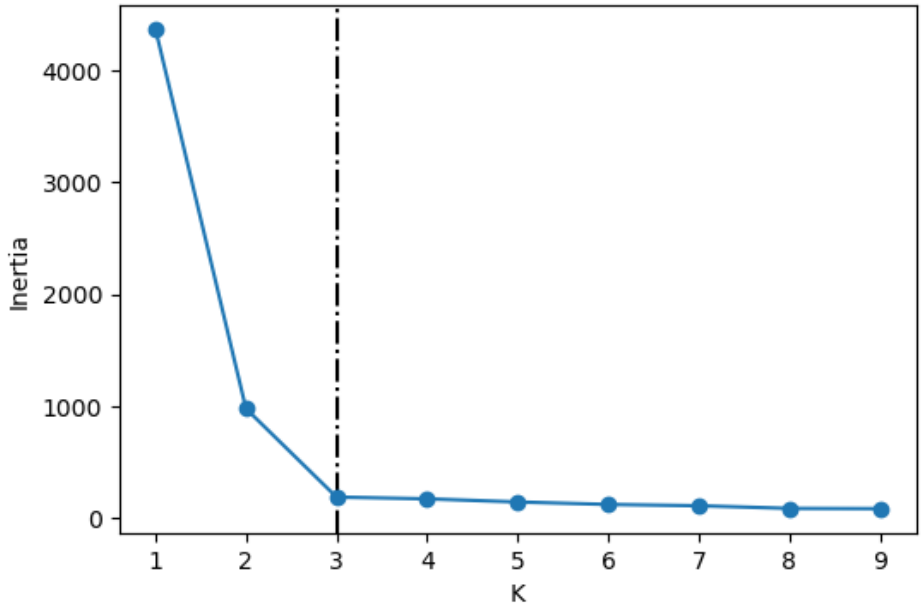
- Concept: Plot the sum of squared errors (SSE) for different values of k and look for an "elbow" where the SSE stops decreasing significantly.
- **Inertia** (also called **Sum of Squared Errors, SSE**) is a metric that measures how **tightly grouped** the data points are around their cluster centers.
  It is the **sum of squared distances** between each data point and its assigned cluster center.

$$\text{Inertia} = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{c}_{\text{cluster}_i}\|^2$$

$\mathbf{x}_i$ is a data point,

$\mathbf{c}_{\text{cluster}_i}$ is the center of its assigned cluster,

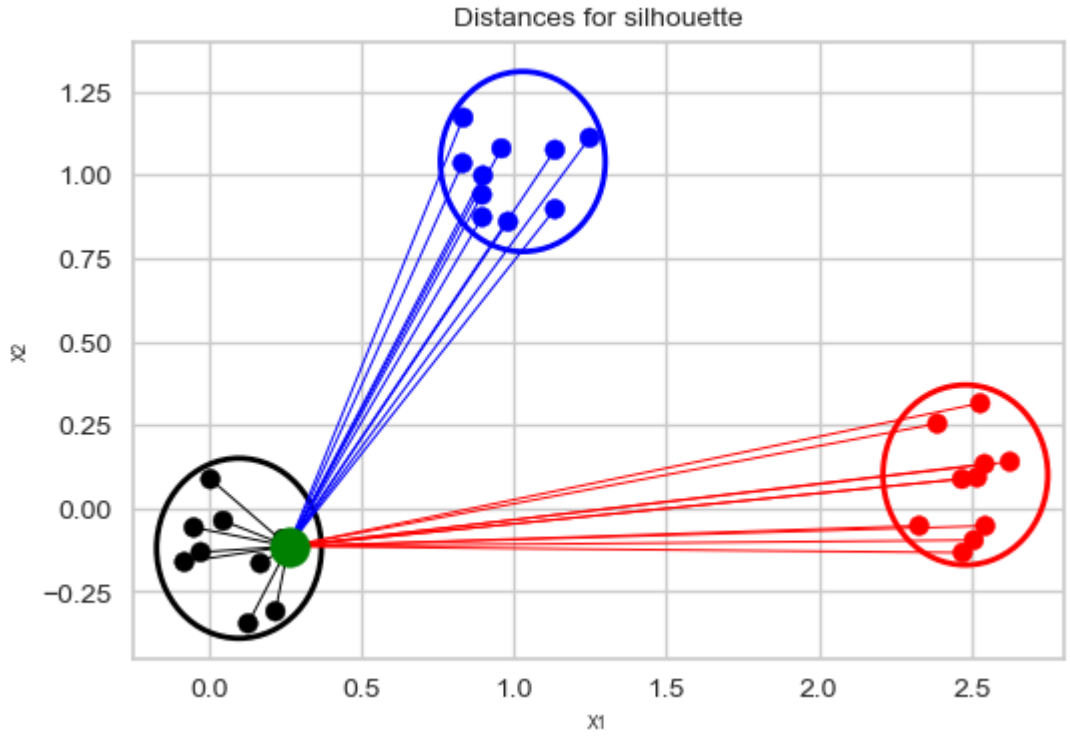
$\|\cdot\|^2$ is the squared Euclidean distance.



The Silhouette method

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- **Concept:** Measures how well a point fits inside its cluster compared to other clusters.
- **Ranges:** Between -1 and 1. Higher = better clustering.

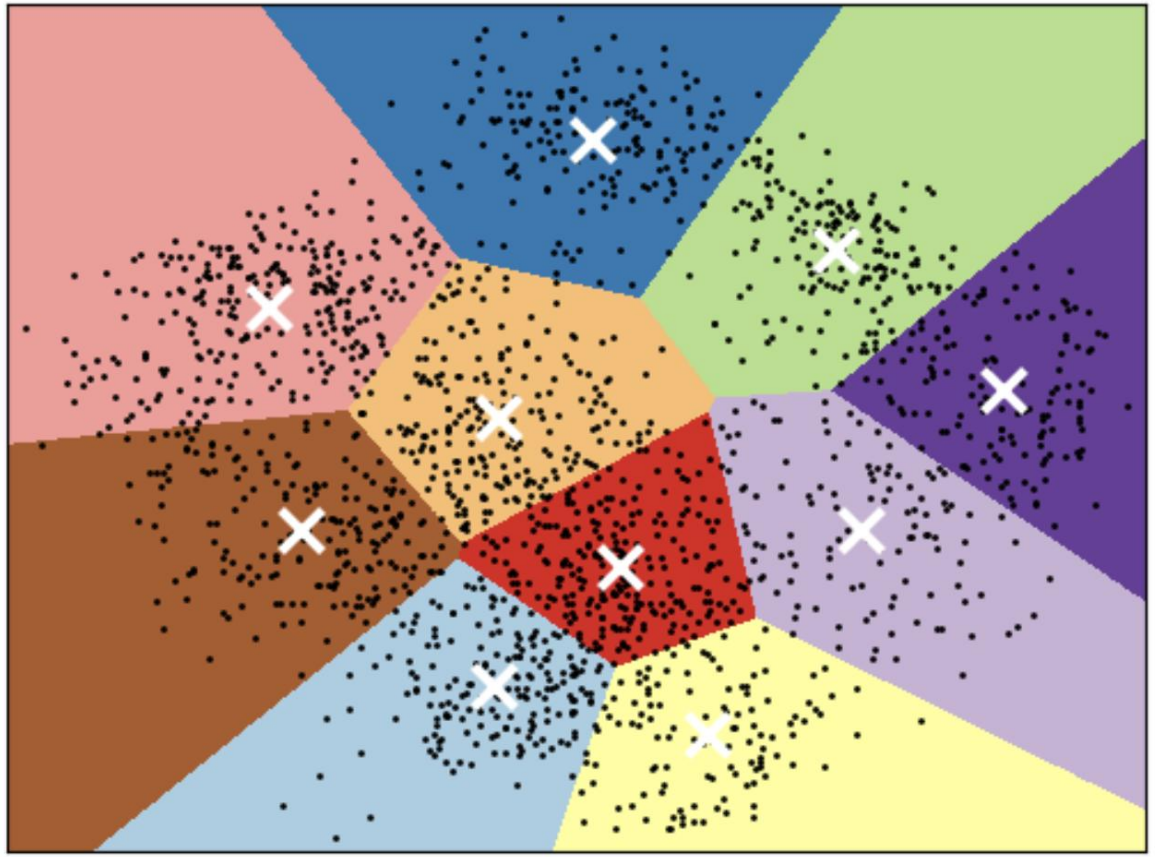


K-Means limitations

- Relies on random initialization and so the outcome may change depending upon this initialization.
- K-Means clustering requires to specify the number of clusters in advance.
- Very often you do not know the centers in advance. The elbow method or the silhouette method to find the optimal number of clusters are not always easy to interpret.
- Each point has to have a cluster assignment.

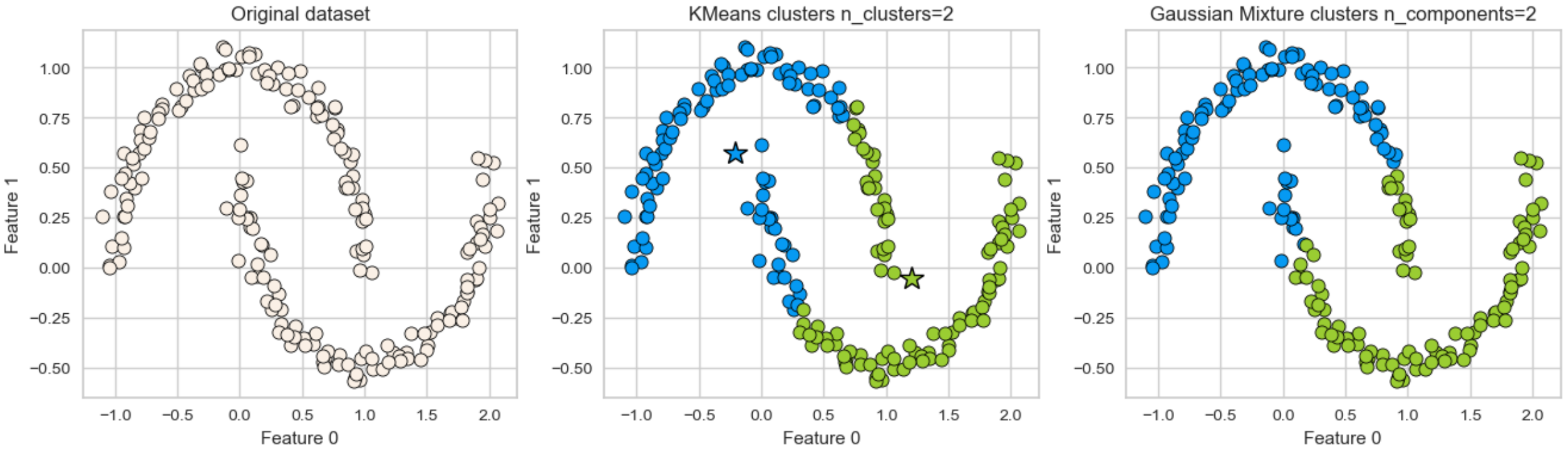
K-Means limitations

- K-Means partitions the space based on the closest mean.
- Each cluster is defined solely by its center and so it can only capture relatively simple shapes.
- So the boundaries between clusters are linear; It fails to identify clusters with complex shapes



K-Means Failure Case 1

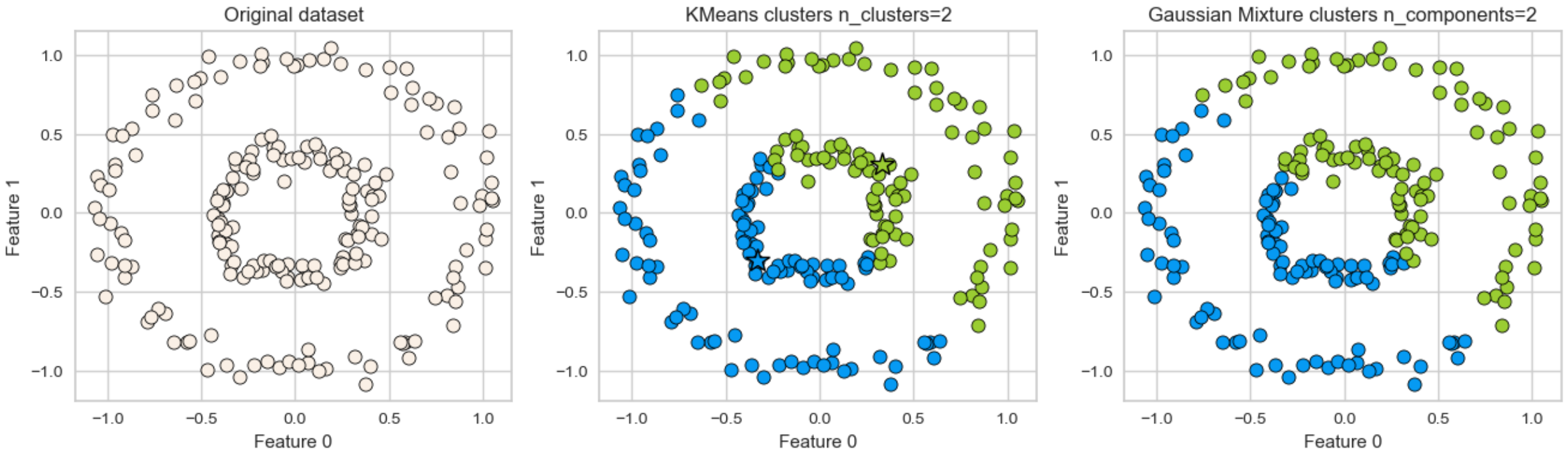
McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology



K-Means performs poorly if the clusters have more complex shapes (e.g., two moons data below).

K-Means Failure Case 2

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology



K-Means is unable to capture complex cluster shapes.

K-Means Failure Case 3

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

It assumes that all directions are equally important for each cluster and fails to identify non-spherical clusters.

Contents

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

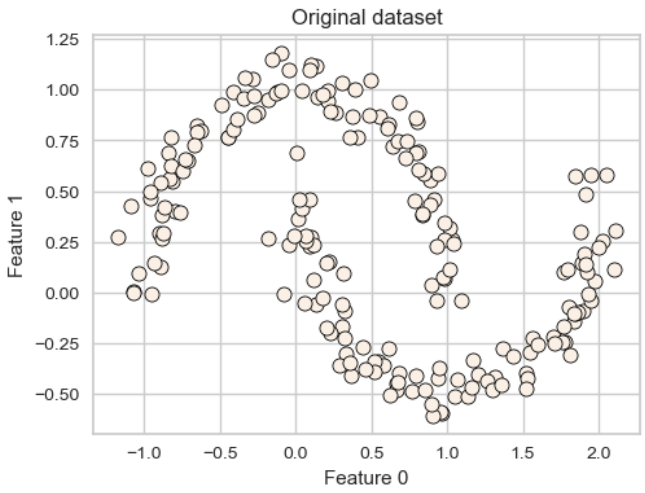
- Introduction
- K-Means Clustering
- DBSCAN
- Hierarchical Clustering
- Summary

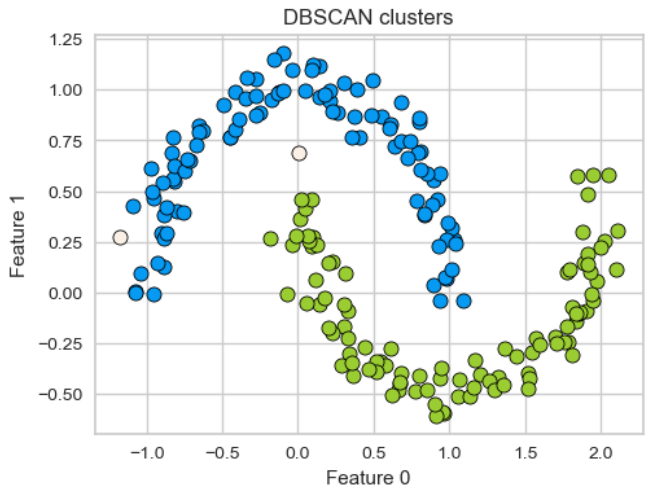
DBSCAN

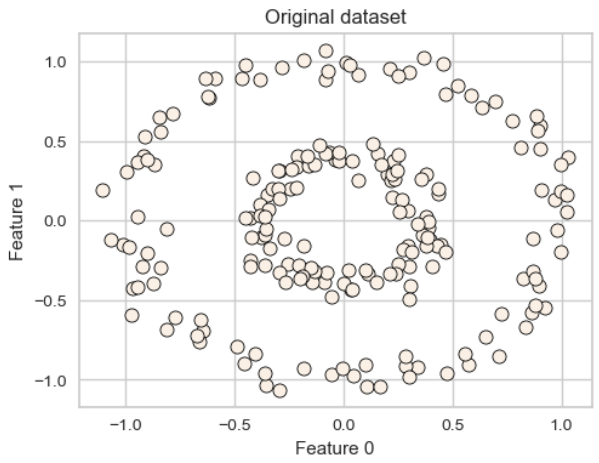
- DBSCAN is a density-based clustering algorithm.
- Intuitively, it's based on the idea that clusters form dense regions in the data and so it works by identifying "crowded" regions in the feature space.
- It can address some of the limitations of K-Means we saw above.
  - It does not require the user to specify the number of clusters in advance.
  - It can identify points that are not part of any clusters.
  - It can capture clusters of complex shapes.

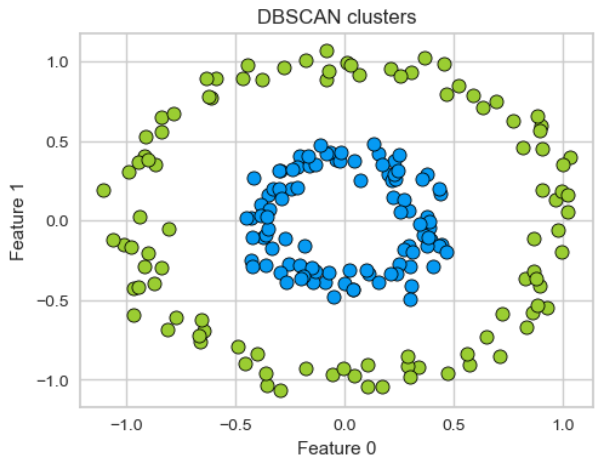
DBSCAN captures complex clusters

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology









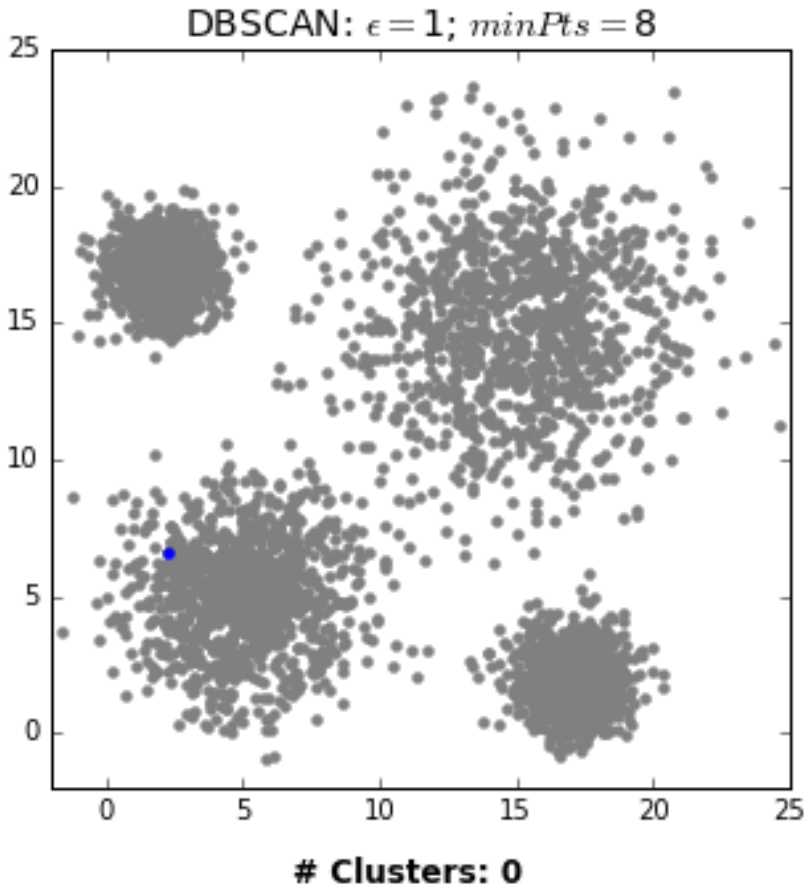
DBSCAN is able to capture half moons shape

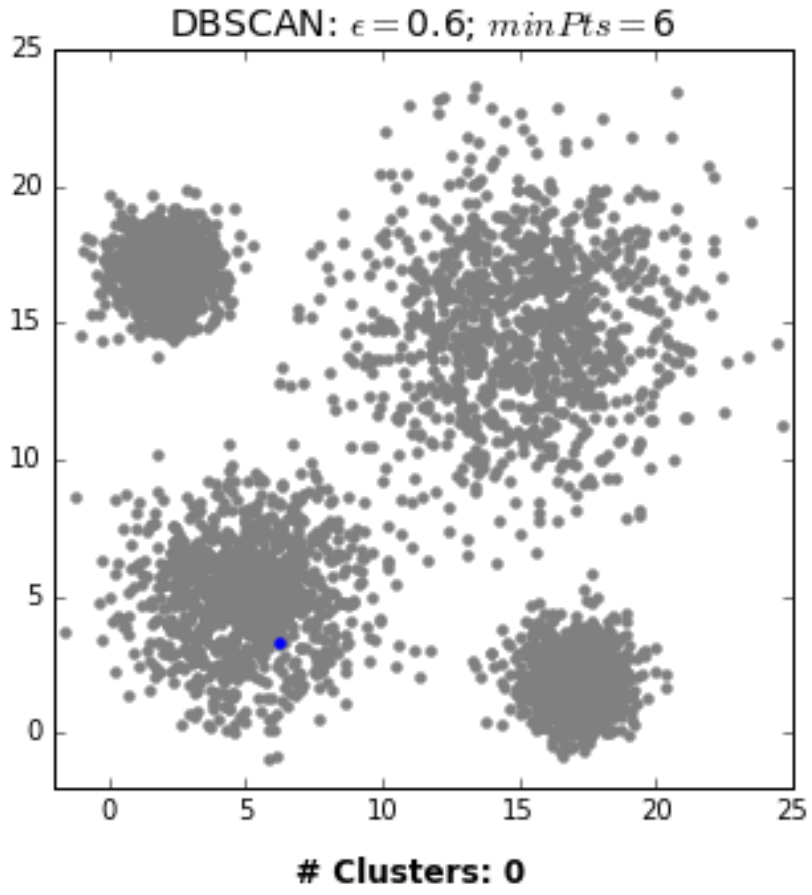
We don’t not have to specify the number of clusters.

DBSCAN: Iterative algorithm

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Based on the idea that clusters form dense regions in the data.





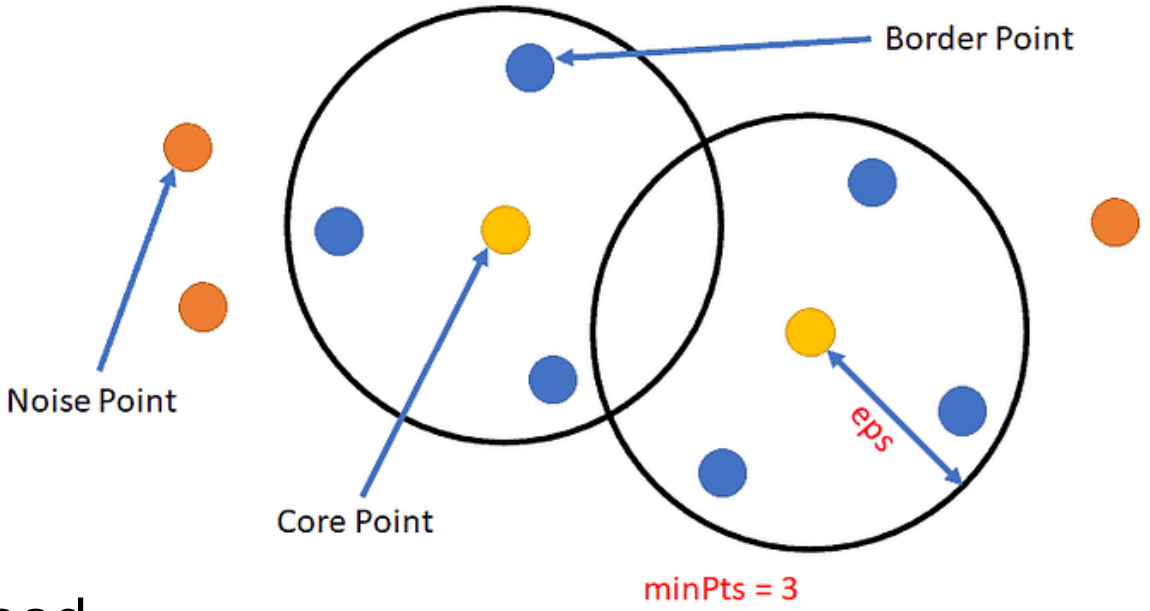
Source

DBSCAN Hyperparameters

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Min Samples (min_samples) – Minimum Points in a Cluster

- Defines the **minimum number of points** (including itself) that a core point must have in its **eps-neighborhood**.
- Controls **cluster density**:
  - **Higher min_samples** → More **strict**, fewer clusters.
  - **Lower min_samples** → More **lenient**, may lead to more small clusters.

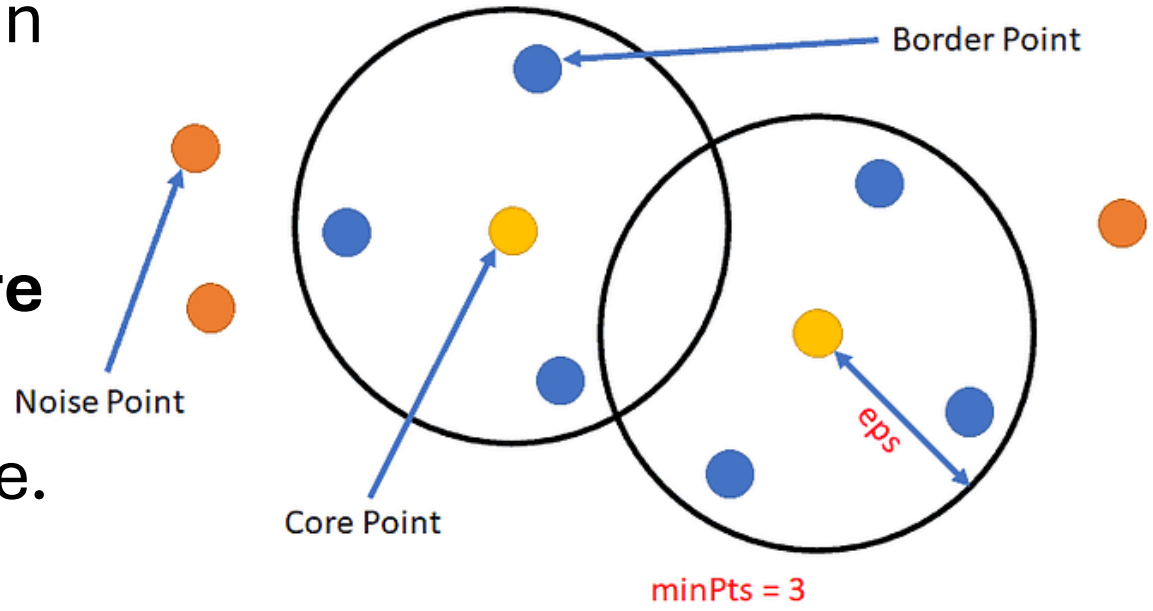


DBSCAN Hyperparameters

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Epsilon (eps) – Neighborhood Radius

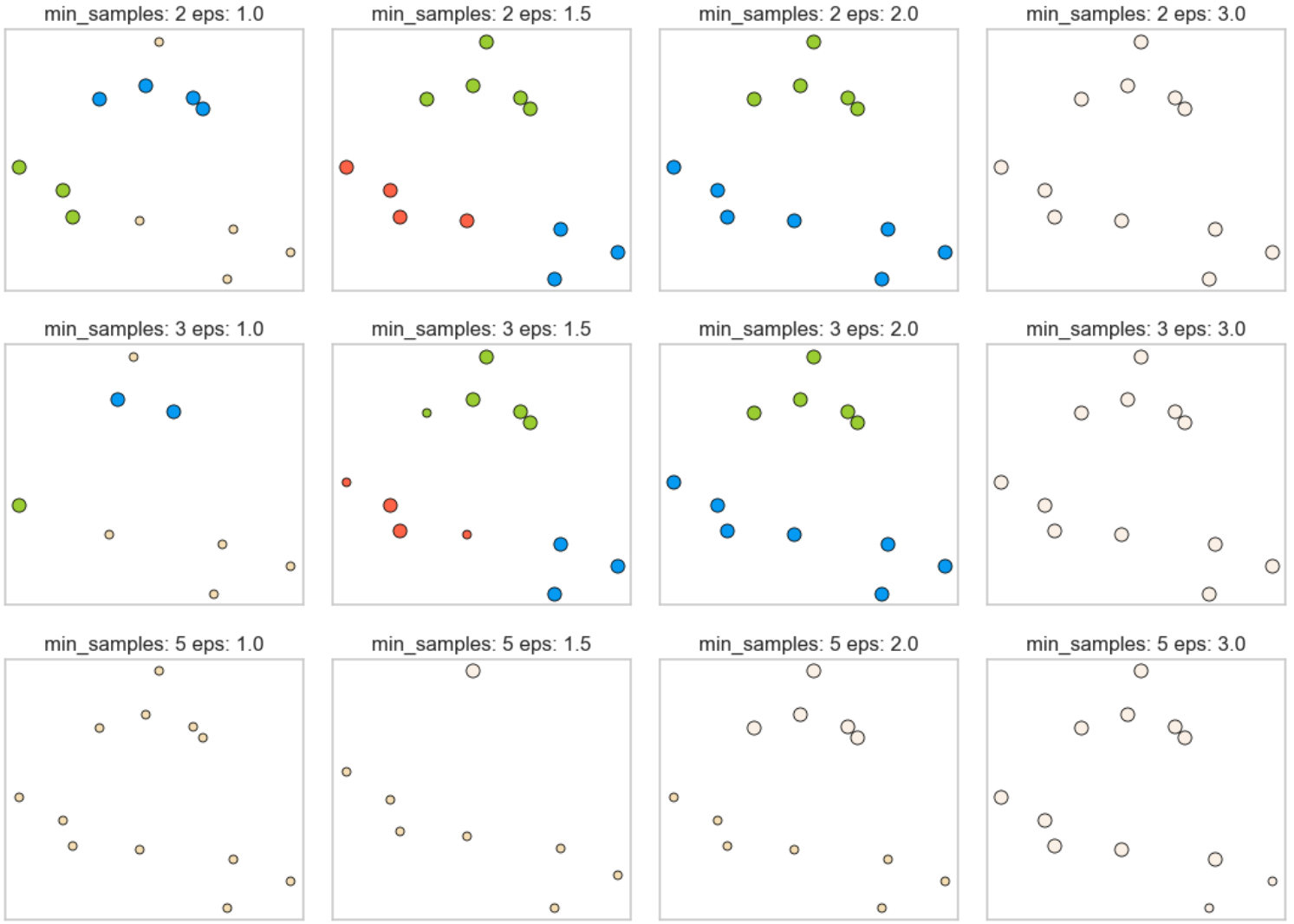
- Defines the **maximum distance** between two points to be considered **neighbors**.
- If a point has **at least min_samples points within eps**, it is considered a **core point**.
- **Smaller eps** → More clusters, more noise.
- **Larger eps** → Fewer clusters, risk of merging distinct clusters.



DBSCAN Hyperparameters

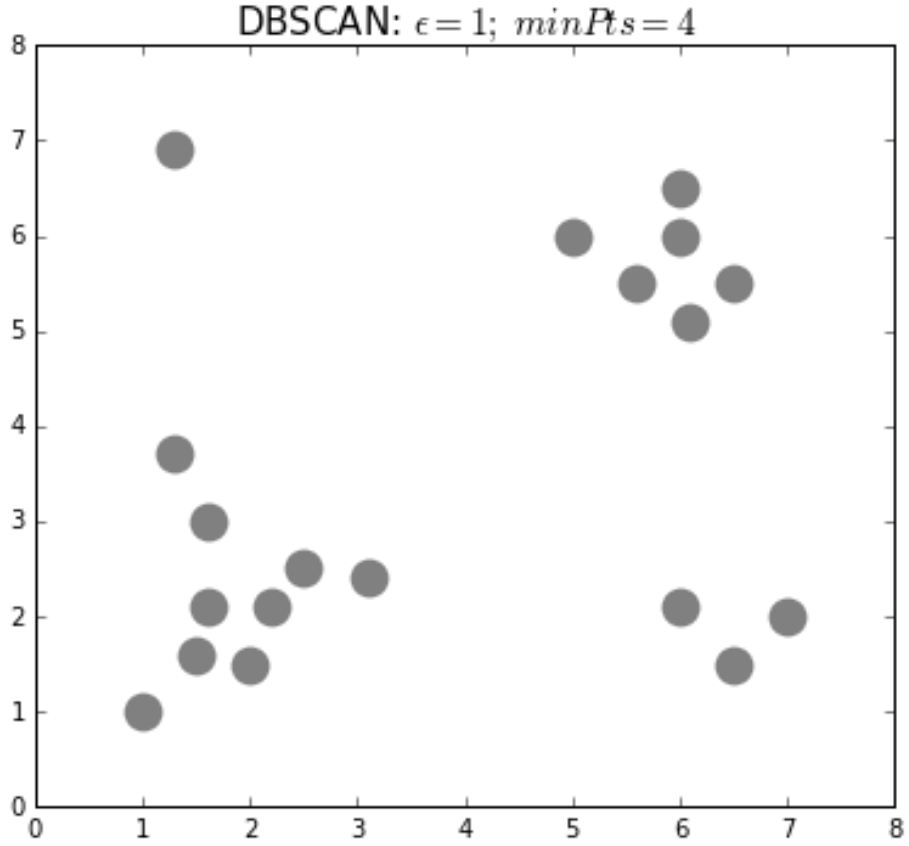
McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Increasing eps (↑) (left to right in the plot above) means more points will be included in a cluster.
  - eps = 1.0 either creates more clusters or more noise points, whereas eps=3.0 puts all points in one cluster with no noise points.
- Increasing min_samples (↑) (top to bottom in the plot above) means points in less dense regions will either be labeled as their own cluster or noise.
  - min_samples=2, for instance, has none or only a fewer noise points whereas min_samples=5 has several noise points.
- Here min_samples = 2.0 or 3.0 and eps = 1.5 is giving us the best results.



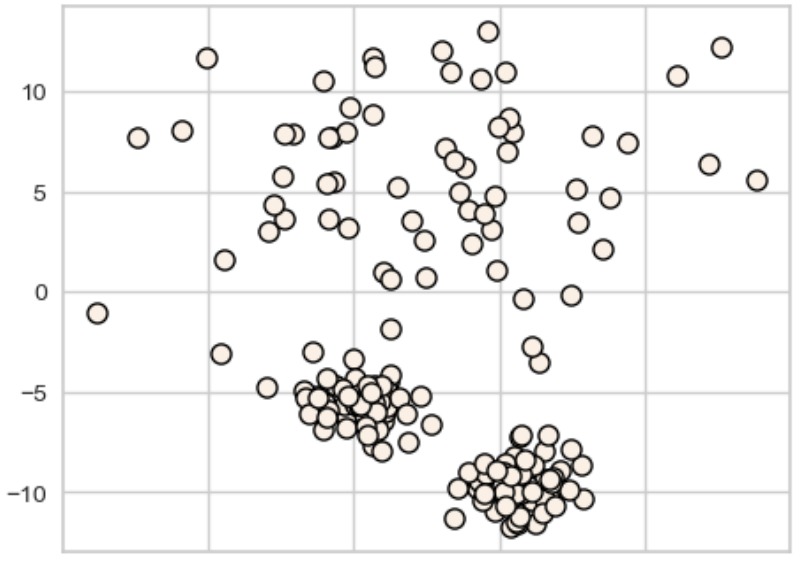
DBSCAN Steps

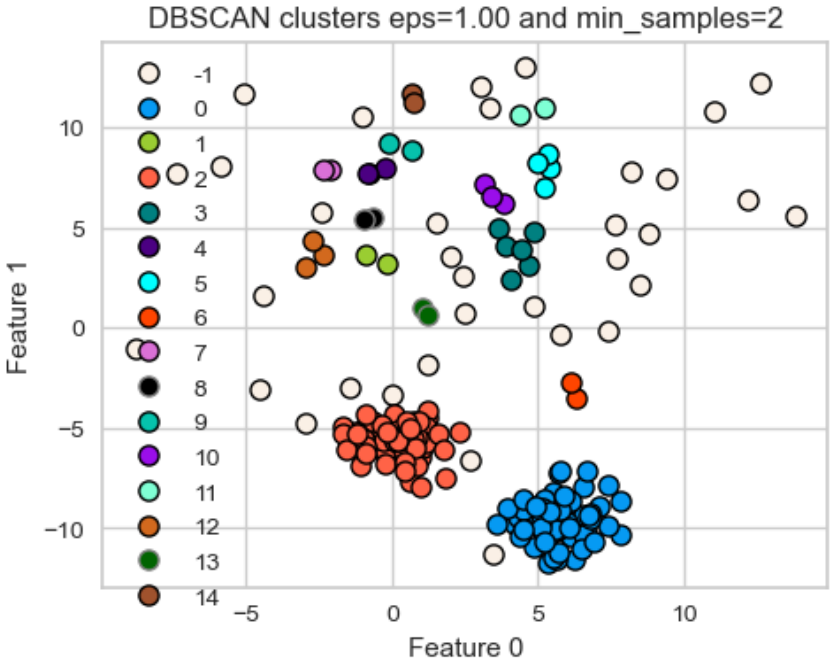
- **Define Hyperparameters.**
- **Classify Points**:
  - **Core Point**: Has ≥ min_samples in eps radius
  - **Border Point**: Near a core point but not dense enough
  - **Noise**: Does not belong to any cluster
- **Expand Clusters**:
  - Start from a **core point** → Expand by adding neighbors
  - Border points join clusters, **noise remains unassigned**
- Repeat Until All Points Are Processed



DBSCAN: failure cases

- K-Means performs better compared to DBSCAN. But it has the benefit of knowing the value of K in advance.





Contents

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

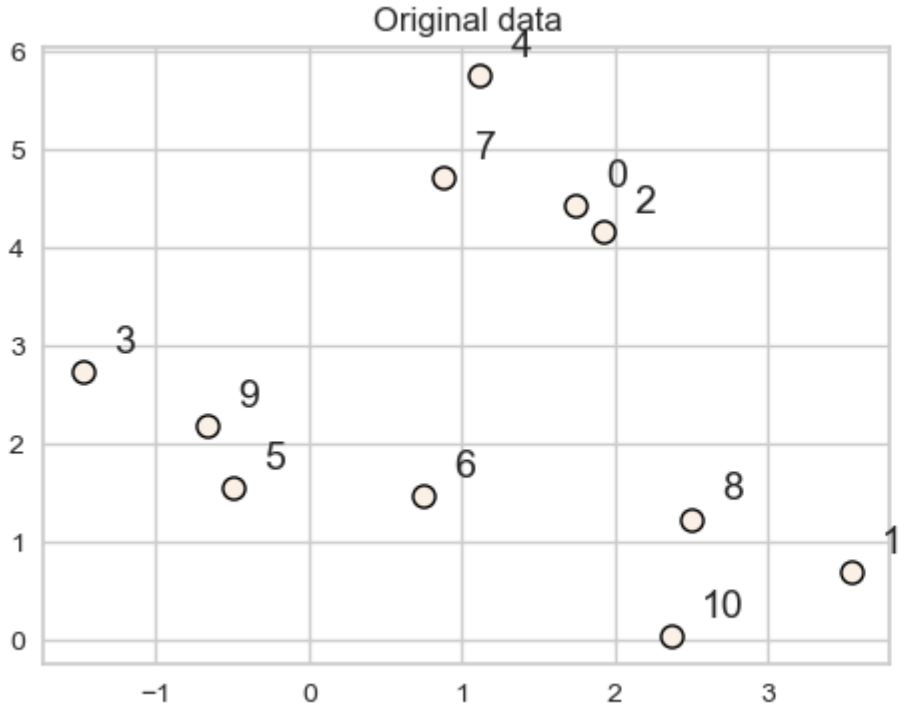
- Introduction
- K-Means Clustering
- DBSCAN
- Hierarchical Clustering
- Summary

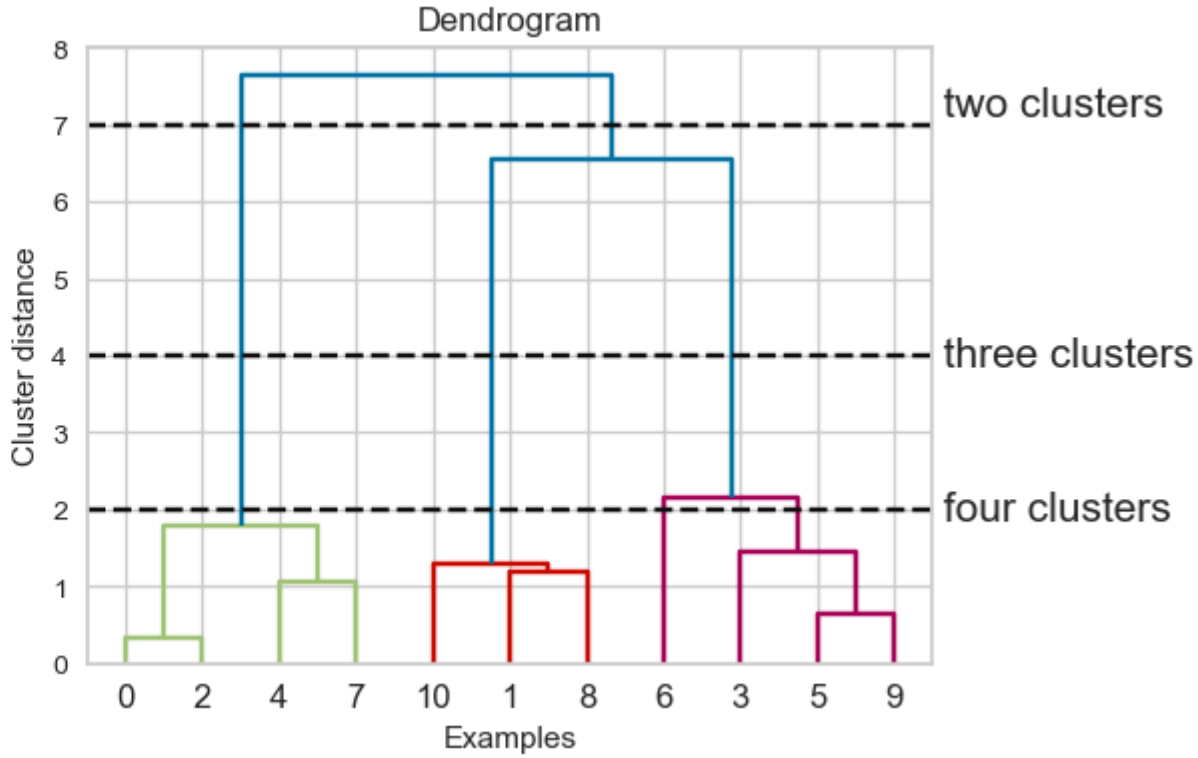
Hierarchical Clustering

- Deciding how many clusters we want is a hard problem.
- Often, it's useful to get a complete picture of similarity between points in our data before picking the number of clusters.
- Hierarchical clustering is helpful in these scenarios.
- Hierarchical clustering is a technique used to group similar data points together based on their similarity creating a hierarchy or tree-like structure.
- key idea is to begin with each data point as its own separate cluster and then progressively merge or split them based on their similarity.

Hierarchical clustering

- Hierarchical clustering can be visualized using a tool called **a dendrogram**.



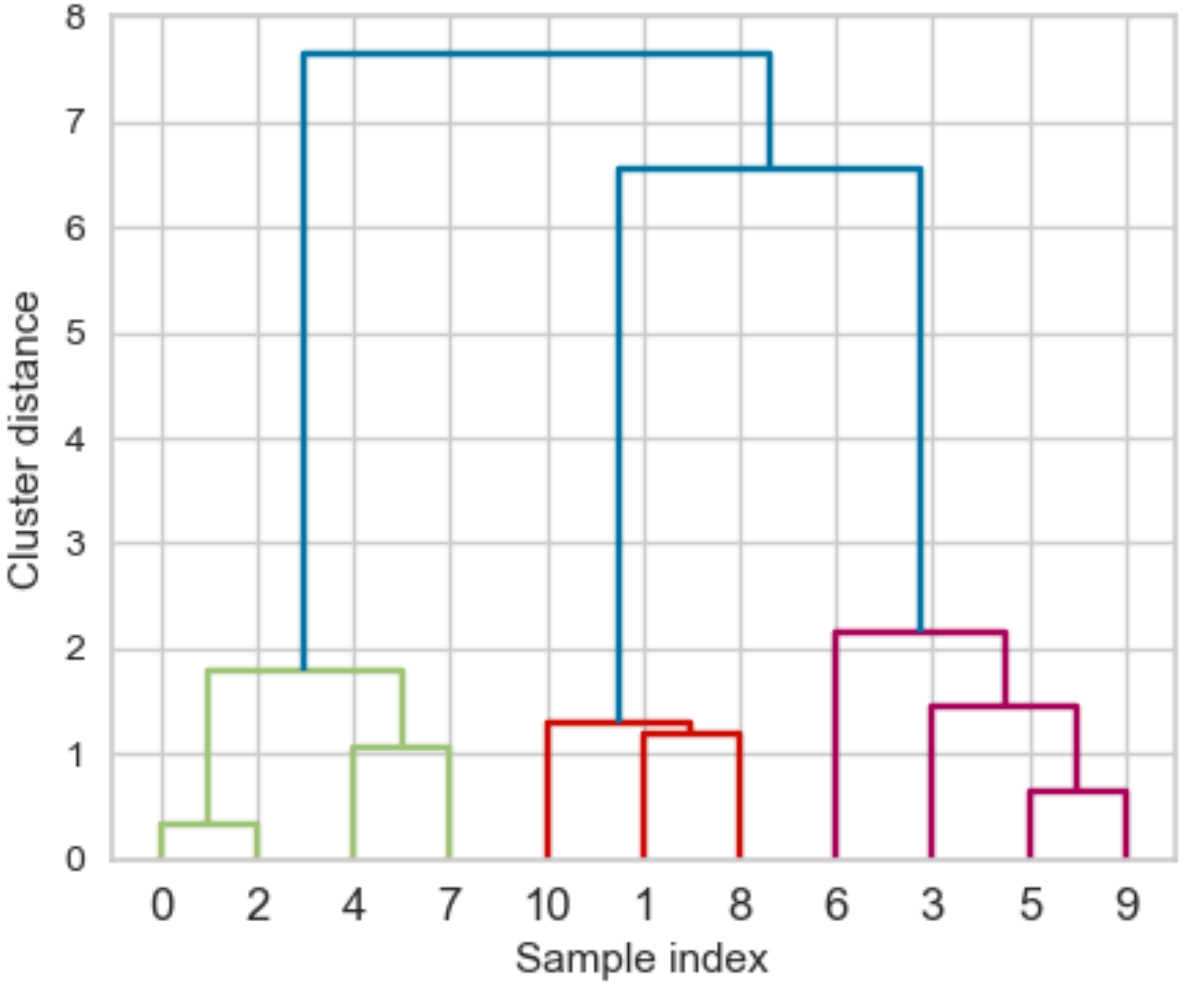


•Every point goes through the journey of being on its own (its own cluster) and getting merged with some other bigger clusters.

•The intermediate steps in the process provide us clustering with different number of clusters.

Dendrograms

- Dendrogram is a tree-like plot.
- On the x-axis we have data points.
- On the y-axis we have distances between clusters.
- We start with data points as leaves of the tree.
- New parent node is created for every two clusters that are joined.
- The length of each branch shows how far the merged clusters go.
  - In the dendrogram shown going from three clusters to two clusters means merging far apart points because the branches between three cluster to two clusters are long.



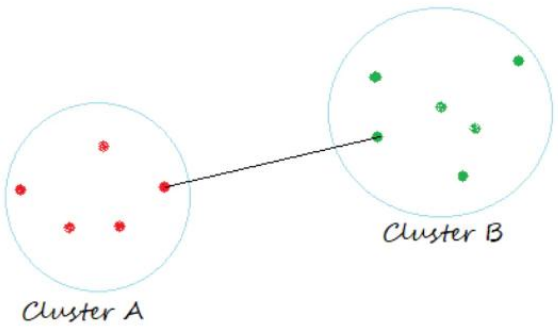
Linkage in dendrograms

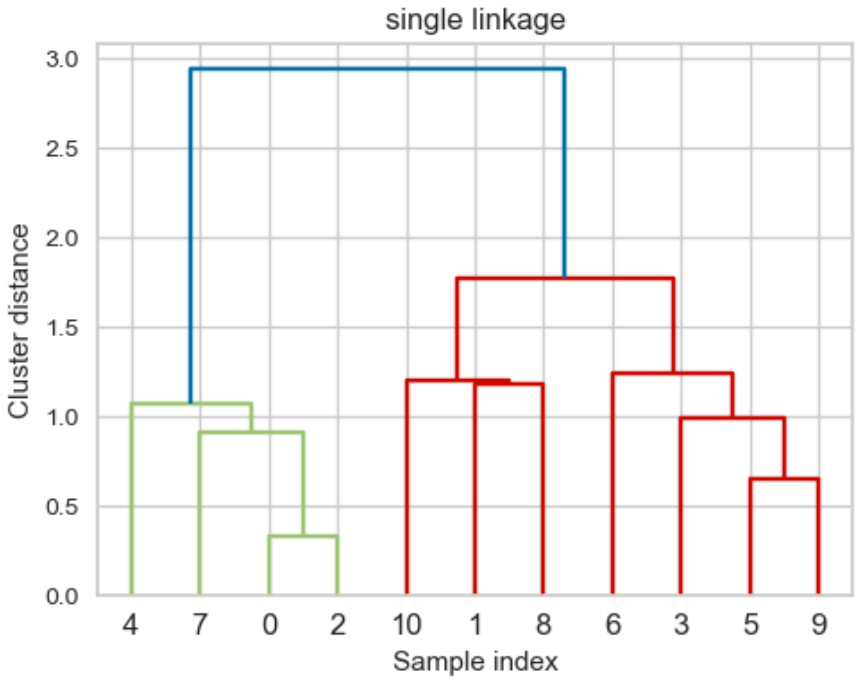
McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

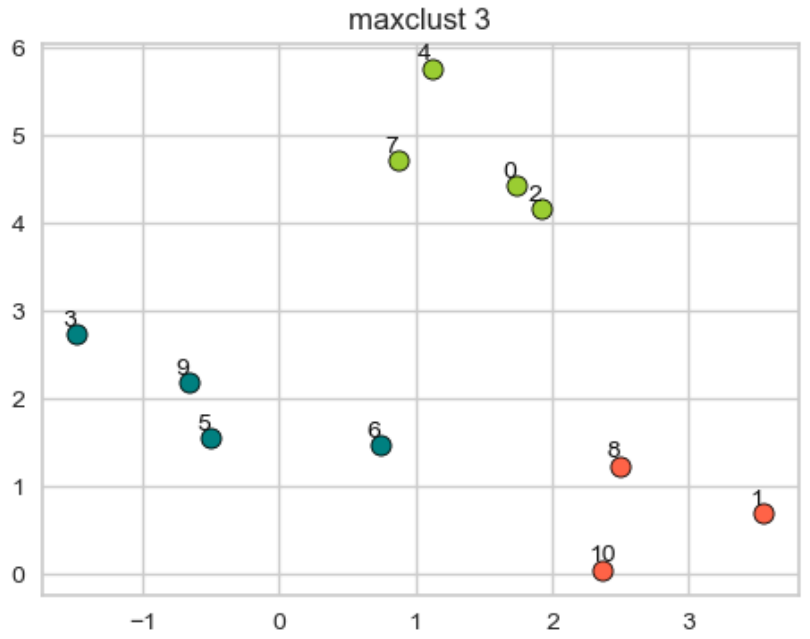
- Linkage determines **how the distance between two clusters is calculated** when merging them in hierarchical clustering. It affects the shape of the **dendrogram** and the resulting clusters.
- Single linkage → smallest minimal distance, leads to loose clusters
- Complete linkage → smallest maximum distance, leads to tight clusters
- Average linkage → smallest average distance between all pairs of points in the clusters
- Ward linkage → smallest increase in within-cluster variance, leads to equally sized clusters

Single Linkage

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

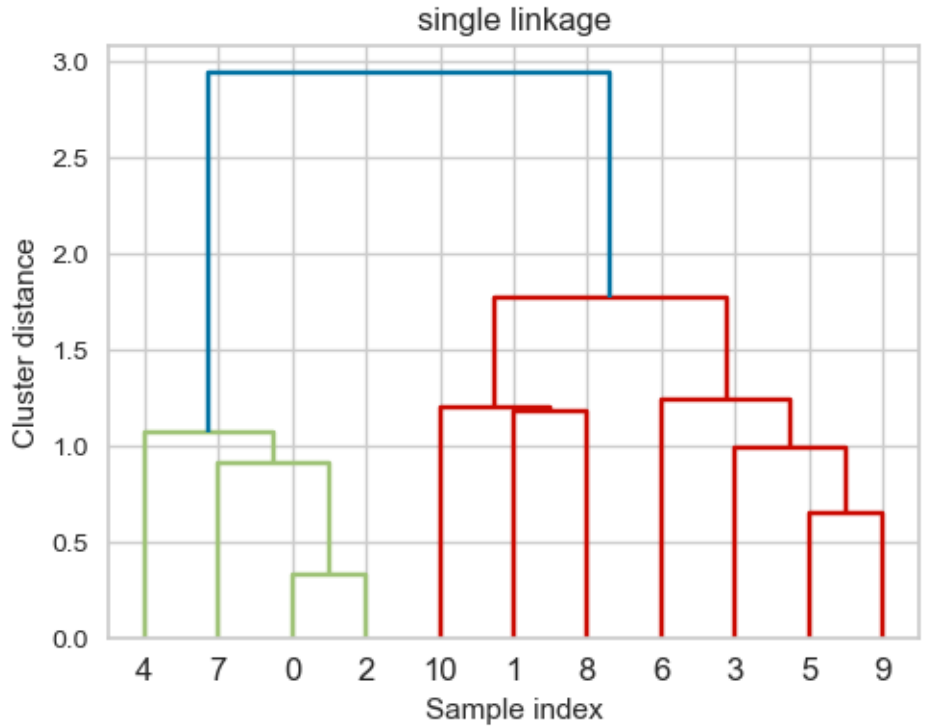


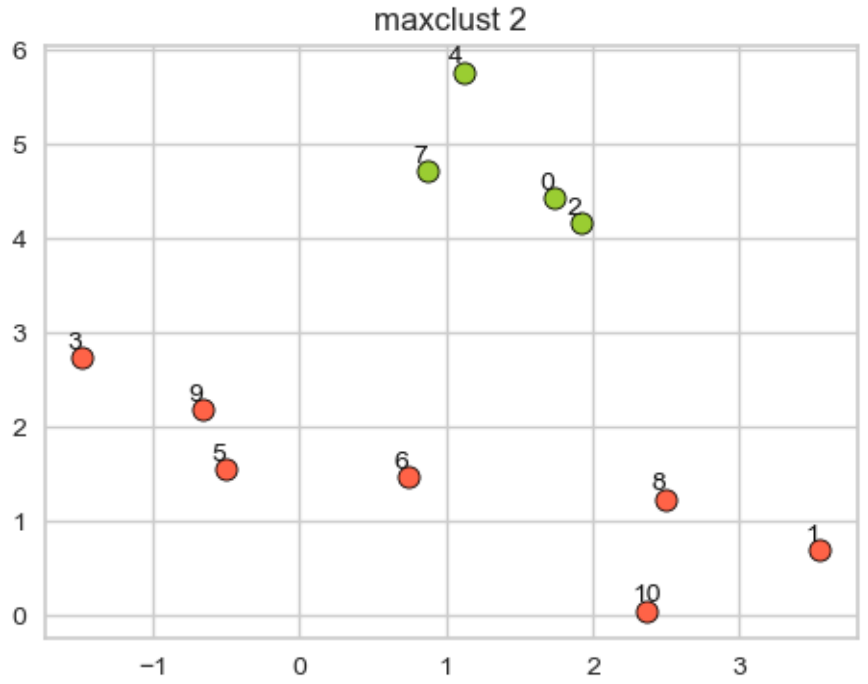




Single linkage measures the **distance between the closest (nearest) points** in two clusters. When merging clusters, it picks the **smallest** distance between any two points—one from each cluster.

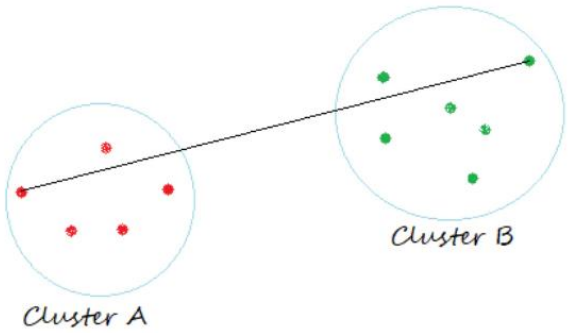
Single Linkage

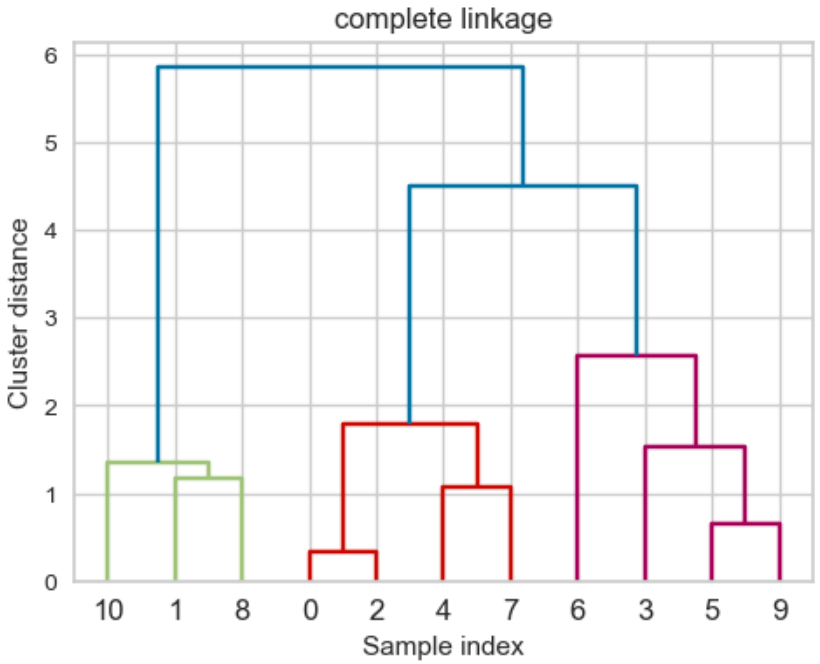


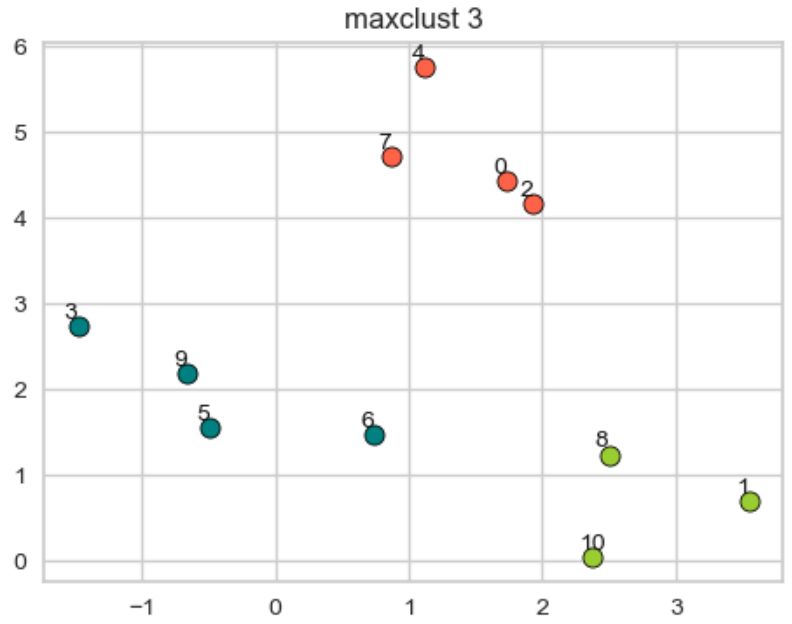


Suppose we decide to go from 3 clusters to 2 clusters. Which clusters would be merged with single linkage criterion?
It will merge the clusters with the smallest minimal distance.

Complete Linkage



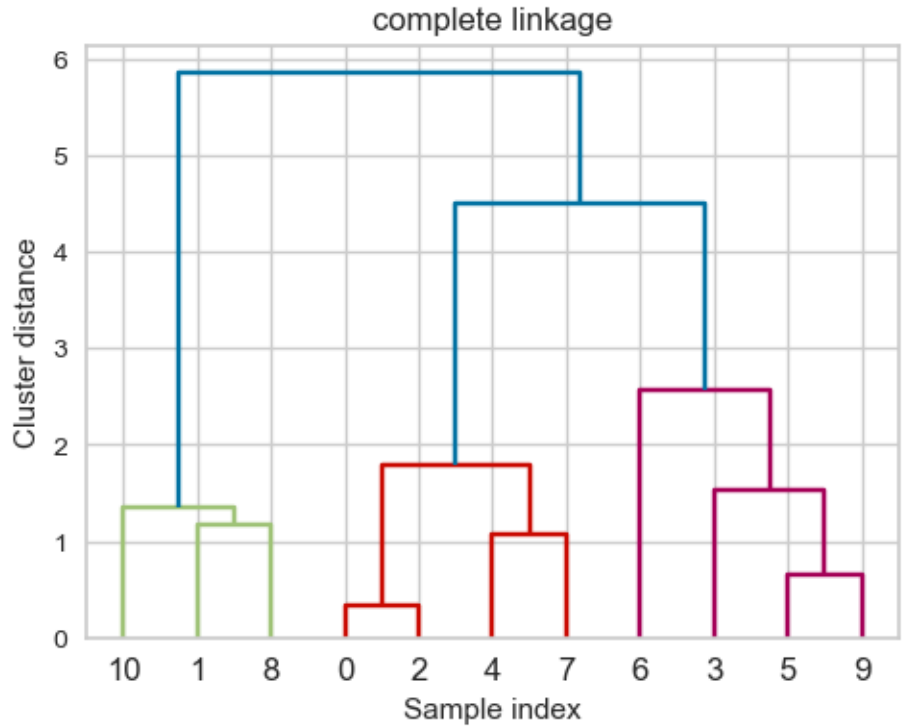


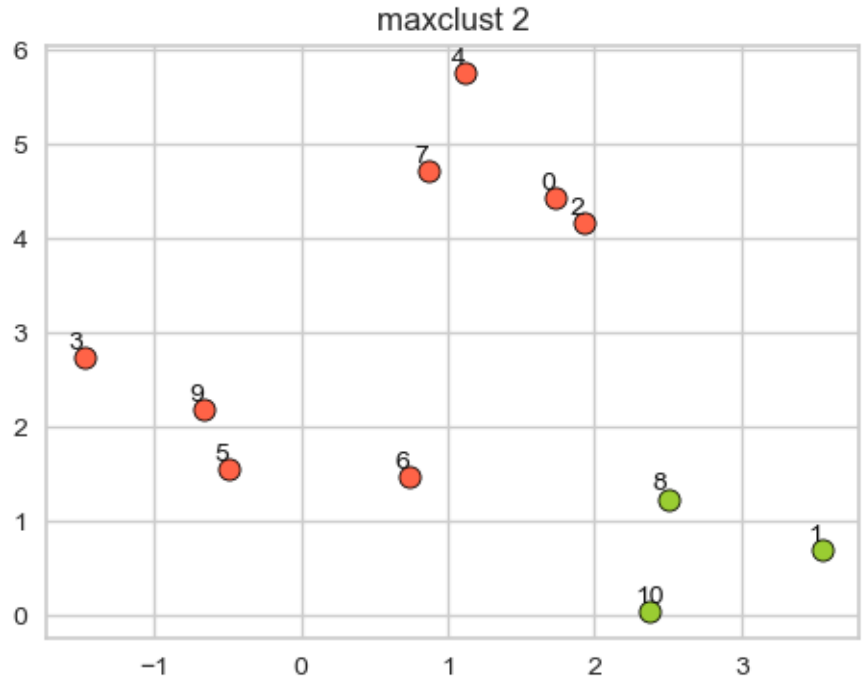


Merges two clusters that have the smallest maximum distance between their points

Complete Linkage

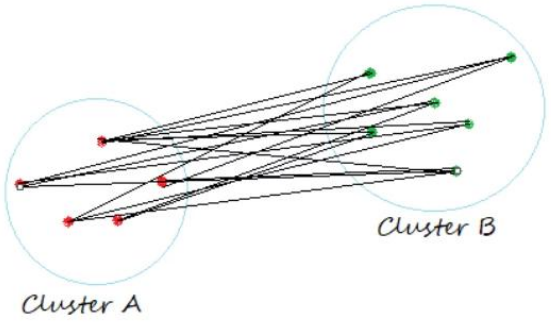
McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

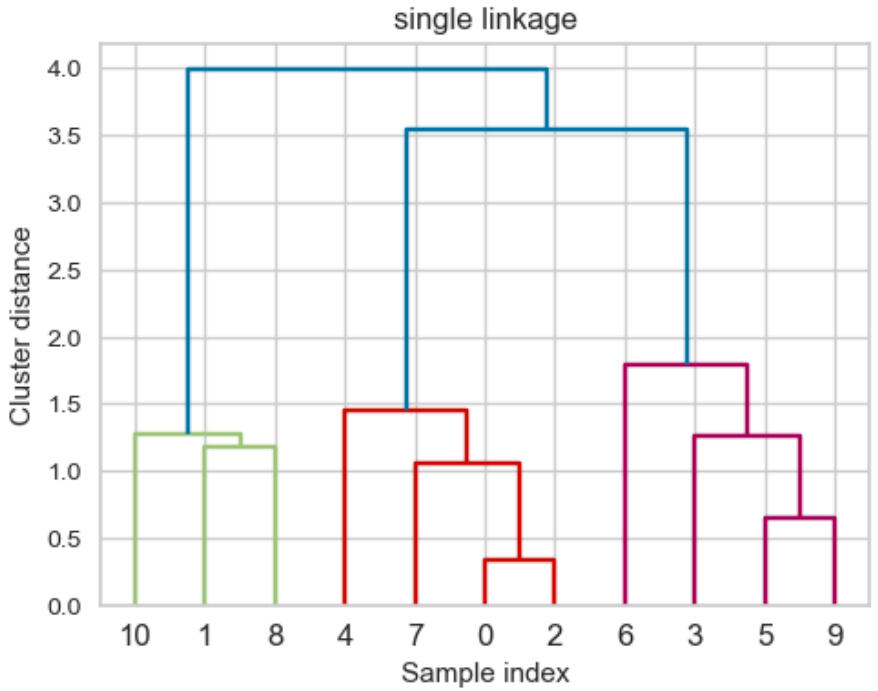


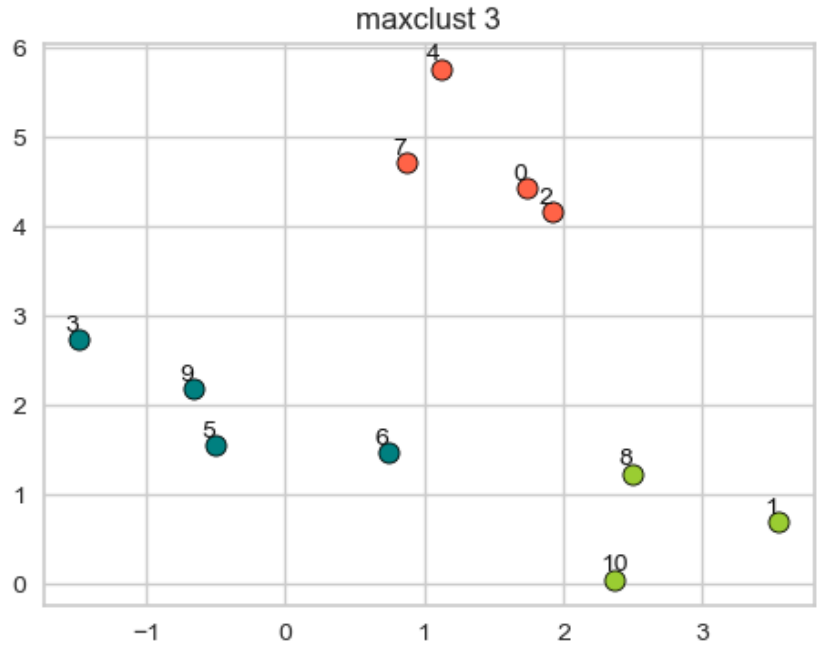


Suppose we decide to go from 3 clusters to 2 clusters. Which clusters would be merged with complete linkage criterion? It will merge the clusters with the smallest maximum distance between their points

Average Linkage

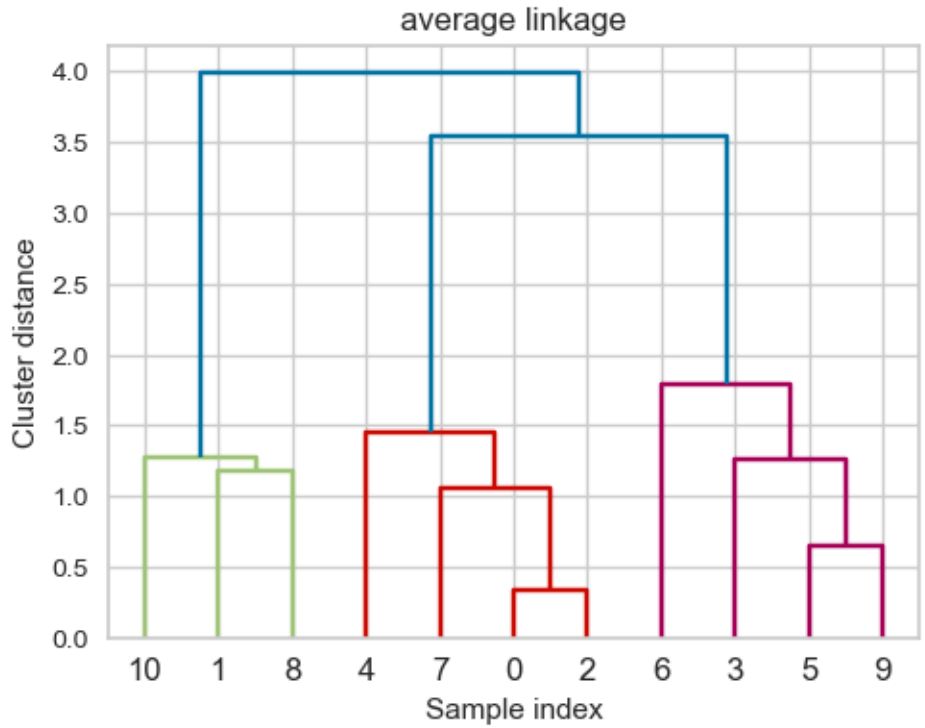


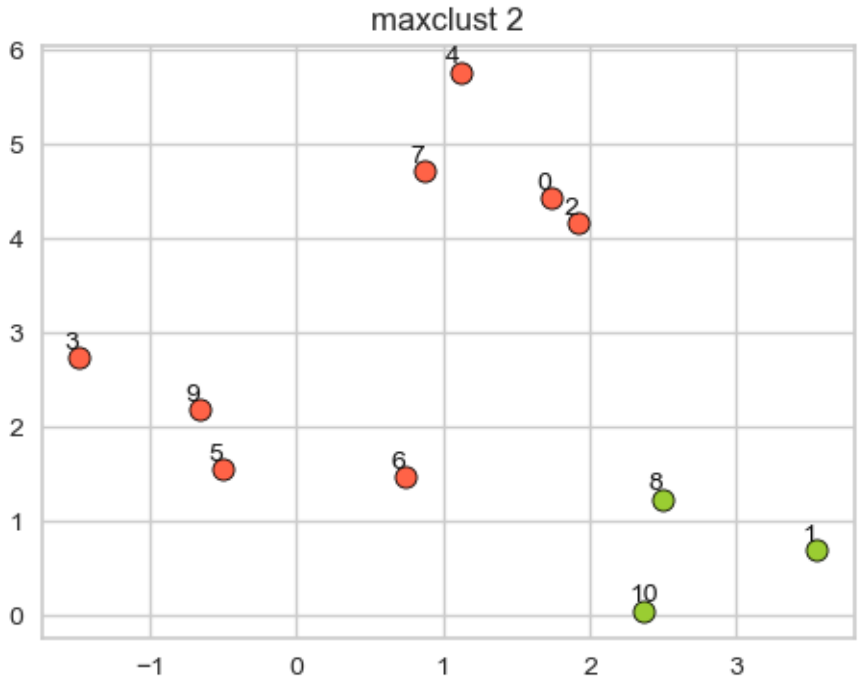




Merges two clusters that have the smallest average distance between all their points.

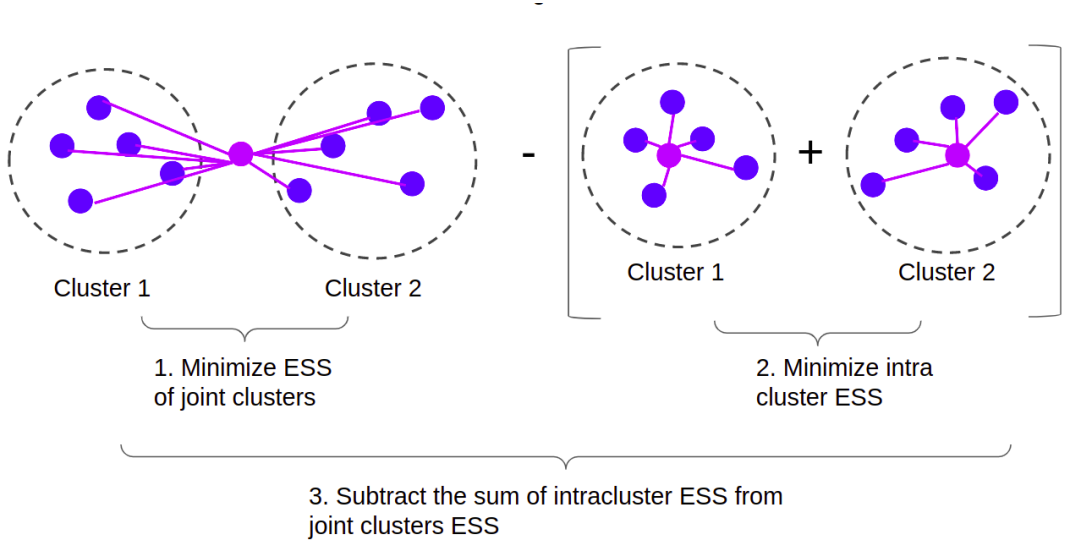
Average Linkage

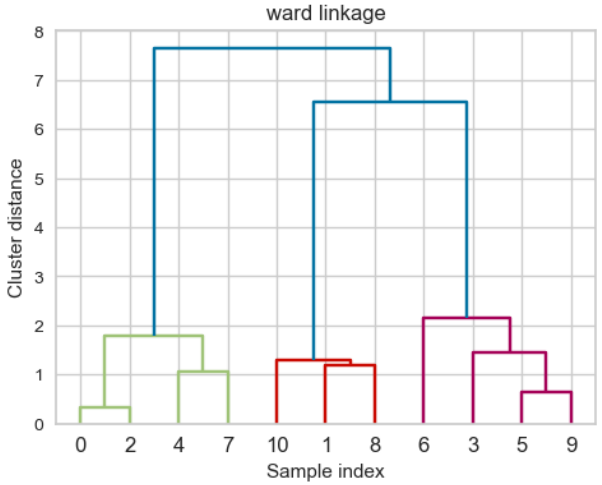


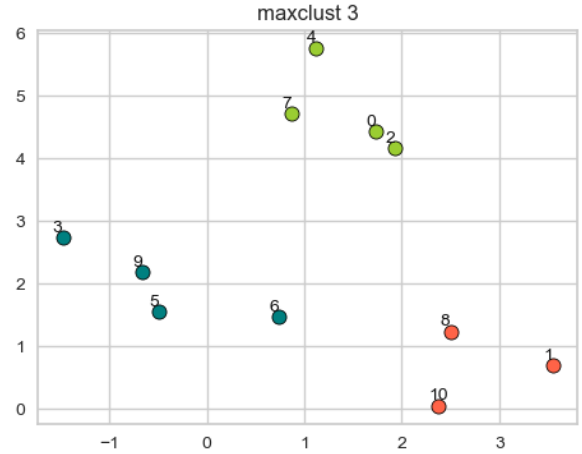


Suppose we decide to go from 3 clusters to 2 clusters. Which clusters would be merged with average linkage criterion? It will merge the clusters with the smallest average distance.

Ward Linkage

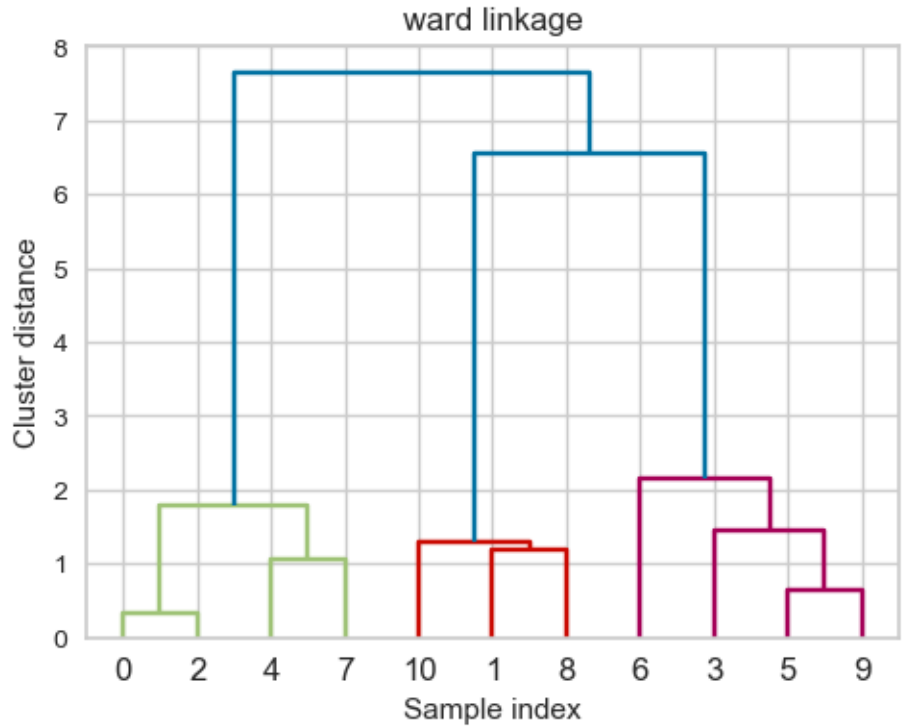


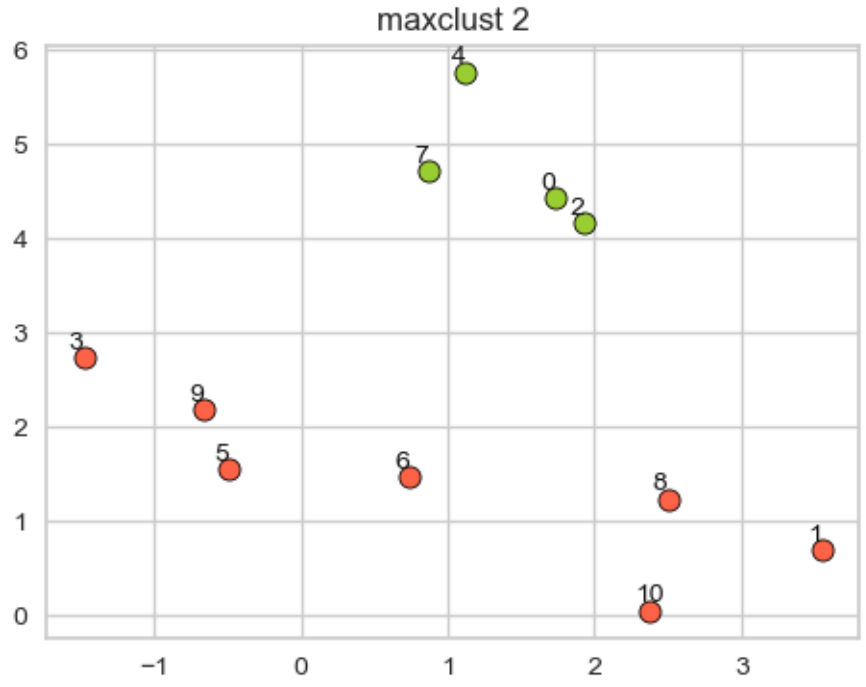




- It specifies the distance between two clusters, computes the sum of squares error (ESS), and successively chooses the next clusters based on the smaller ESS.
- Ward's Method seeks to minimize the increase of ESS at each step.
- Therefore, minimizing error.
- Picks two clusters to merge such that the variance within all clusters increases the least.
- Often leads to equally sized clusters.

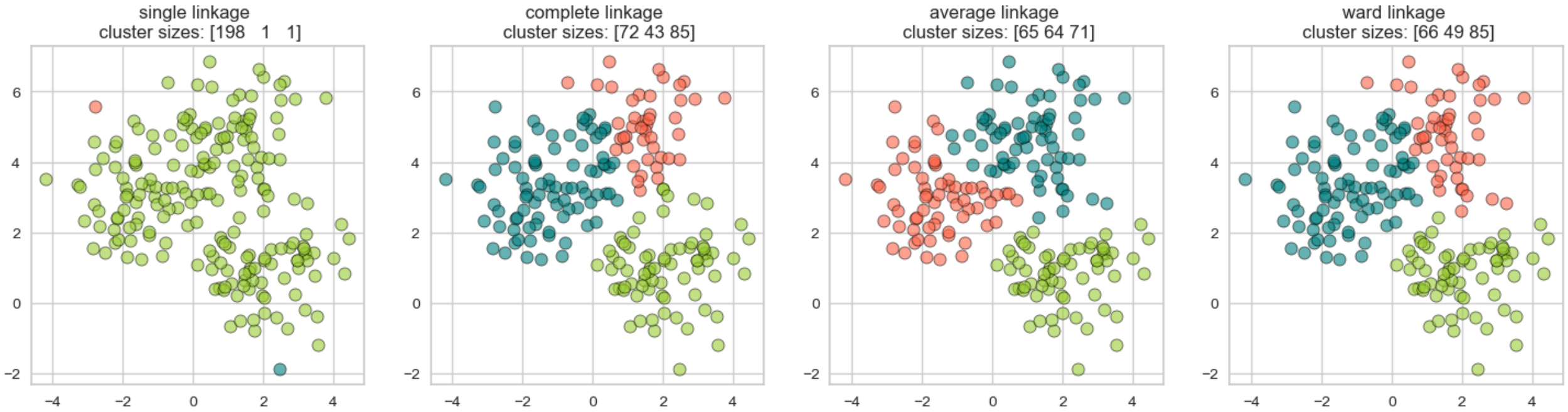
Ward Linkage





Suppose we decide to go from 3 clusters to 2 clusters. Which clusters would be merged with ward linkage criterion? It will merge the clusters with the smallest within-cluster variance

Linkage



Single linkage → smallest minimal distance, leads to loose clusters
Complete linkage → smallest maximum distance, leads to tight clusters
Average linkage → smallest average distance between all pairs of points in the clusters
Ward linkage → smallest increase in within-cluster variance, leads to equally sized clusters

Truncation in Hierarchical Clustering

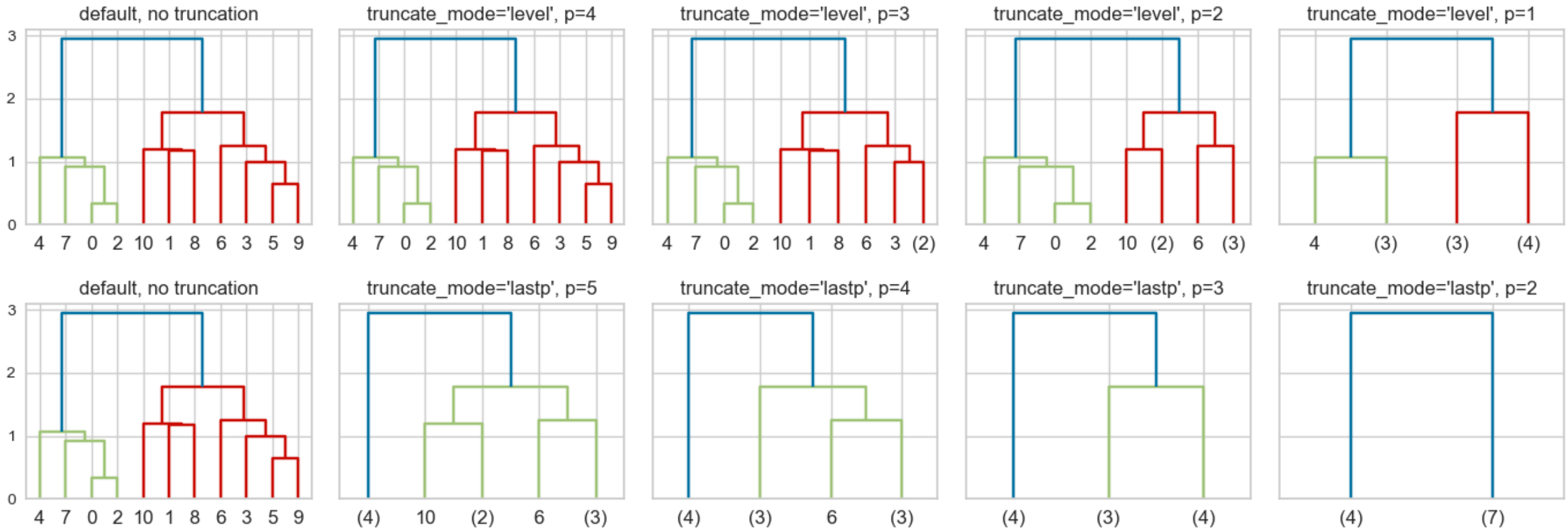
- Truncation in hierarchical clustering means **cutting the dendrogram at a specific height** to form a desired number of clusters.

Why Use Truncation?

- **Controls the number of clusters** by setting a threshold.
- **Removes unnecessary details** from the dendrogram.
- **Focuses on high-level structures** instead of individual points.

Truncation in Hierarchical Clustering

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology



`level` → Maximum depth of the tree is p

`lastp` → Only p leaves are shown

Use **level** if you want to control **how deep the tree is shown**.
Use **lastp** if you want to **focus on a specific number of final clusters**.

Contents

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Introduction
- K-Means Clustering
- DBSCAN
- Hierarchical Clustering
- Summary

- K-Means: Fast, requires k, assumes spherical clusters, sensitive to outliers.
- DBSCAN: Density-based, detects arbitrary shapes, handles noise, needs ε & MinPts tuning.
- Hierarchical: Builds a tree (dendrogram), no need for k, computationally expensive.

Comparison:

- K-Means → Best for large, well-separated clusters.
- DBSCAN → Ideal for noisy data & irregular clusters.
- Hierarchical → Useful for visualizing relationships in small datasets.

Summary

Feature	K-Means	DBSCAN	Hierarchical Clustering
Cluster Shape	Spherical (circular) clusters	Arbitrary shape	Can capture various shapes
Number of Clusters	Must be specified in advance	Determined automatically	Can be determined via dendrogram
Handling Outliers	Sensitive to outliers	Can ignore noise points	Can be sensitive to outliers
Scalability	Fast for large datasets	Slower on large datasets	Computationally expensive
Works Well For	Large, well-separated clusters	Data with noise and irregular shapes	Hierarchical relationships
Hyperparameters	k (number of clusters)	Epsilon (ε), MinPts	Linkage method, distance metric
Assignment Method	Each point belongs to a cluster	Some points can be noise	Builds a tree of clusters
Main Weakness	Requires k beforehand, struggles with non-spherical clusters	Sensitive to parameter selection	Slow for large datasets