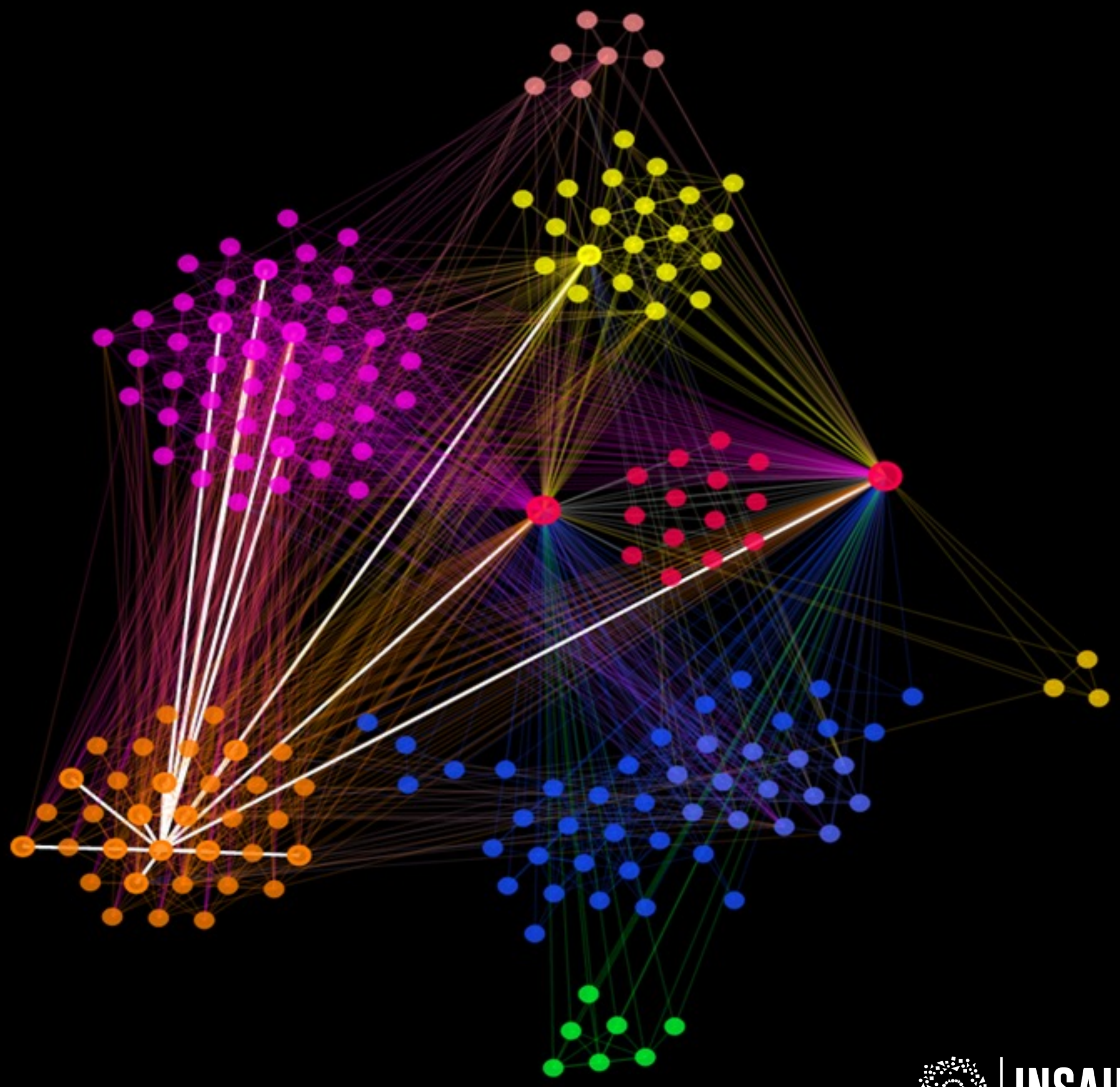
K – Means Clustering



©INSAID All rights reserved.

Course Overview

You are here...

Term	CDF	GCD	GCDAI	PGPDSAI
Term 1	Data Analytics with Python	Data Analytics with Python	Data Analytics with Python	Data Analytics with Python
Term 2	Data Visualization Techniques	Data Visualization Techniques	Data Visualization Techniques	Data Visualization Techniques
Term 3	EDA & Data Storytelling	EDA & Data Storytelling	EDA & Data Storytelling	EDA & Data Storytelling
		Minor Project	Minor Project	Minor Project
Term 4		Machine Learning Foundation	Machine Learning Foundation	Machine Learning Foundation
Term 5		Machine Learning Intermediate	Machine Learning Intermediate	Machine Learning Intermediate
Term 6		Machine Learning Advanced (Mandatory)	Machine Learning Advanced (Mandatory)	Machine Learning Advanced (Mandatory)
		Data Visualization with Tableau (Elective - I)	Data Visualization with Tableau (Elective - I)	Data Visualization with Tableau (Elective - I)
		Data Analytics with R (Elective - II)	Data Analytics with R (Elective - II)	Data Analytics with R (Elective - II)
		Capstone Project	Capstone Project	Capstone Project
Term 7		Bonus: Industrial ML (ML – 4 & 5)	Basics of AI, TensorFlow, and Keras	Basics of AI, TensorFlow, and Keras
Term 8			Deep Learning Foundation	Deep Learning Foundation
Term 9			NPL – I/CV – I	CV – I
Term 10			NLP – II/CV – II	NLP – I
			Capstone Project	Capstone Project
Term 11				CV – II
Term 12				NLP – II
				NLP – III + CV – III
				AutoVision & AutoNLP
				Building AI product

©INSAID All rights reserved.

Term Context

- K – Nearest Neighbor
- **K-means Clustering** You are here…
- Ensemble Learning
- Optimization

©INSAID All rights reserved.

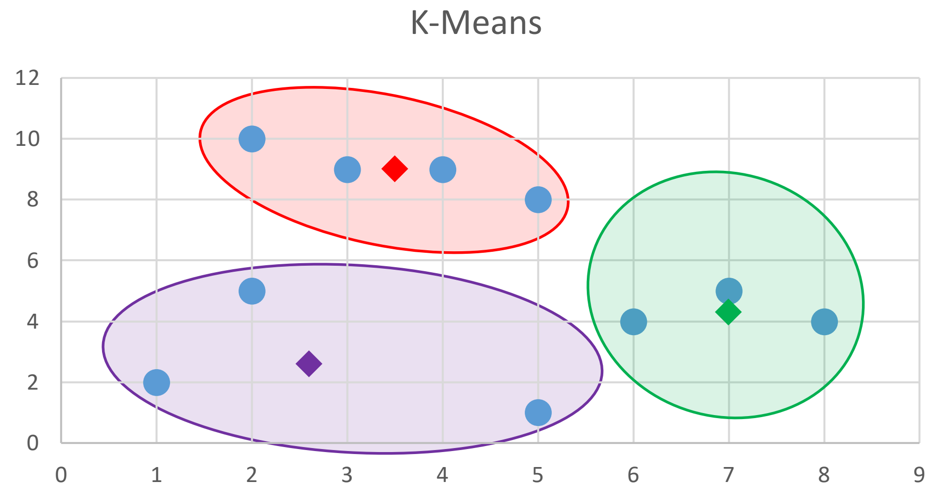
Agenda

©INSAID All rights reserved.

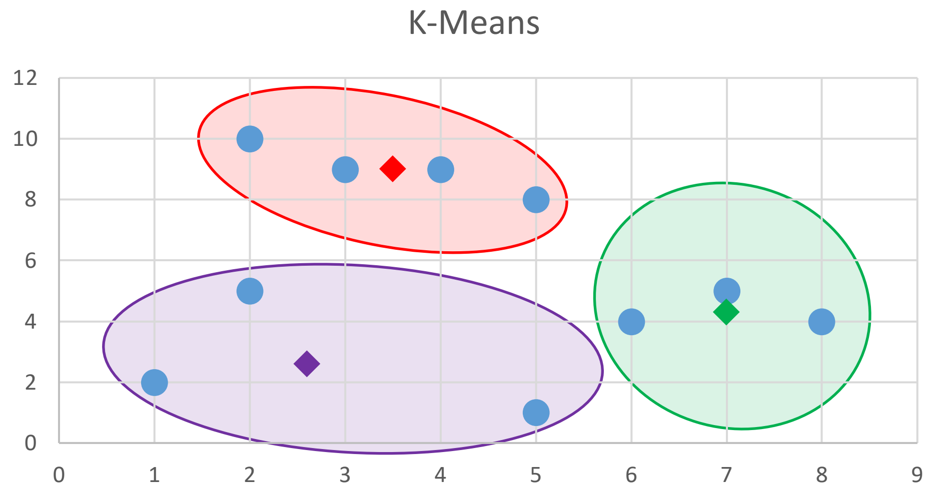
Stopping Criteria

- There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

  1. **Centroids** of newly formed clusters **do not change**.
  2. **Points** remain in the **same cluster** after many iterations.
  3. **Maximum** number of **iterations** are reached.



Iteration 5 for K = 3



Iteration 6 for K = 3

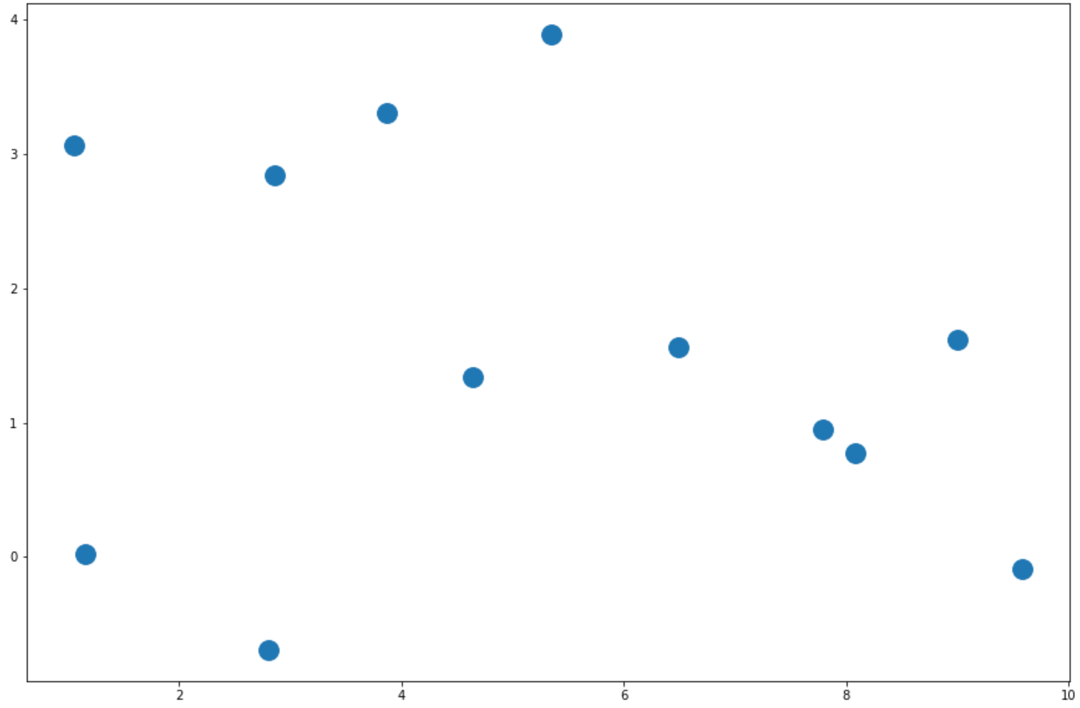
©INSAID All rights reserved.

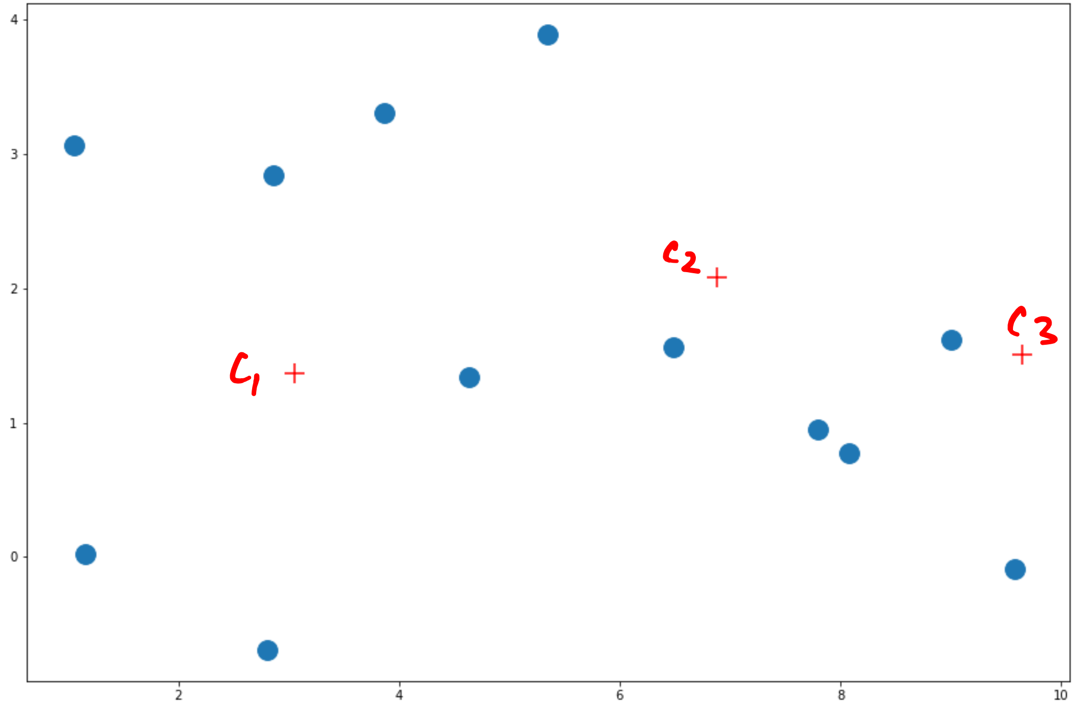
Agenda

©INSAID All rights reserved.

K-Means Clustering (Example) – Iteration 1

- Let's apply K-Means clustering with **K=3** to the following set of points. We will limit the number of iterations to 5.
- Step 1: Assume K=3 random points anywhere in the graph as cluster representatives.

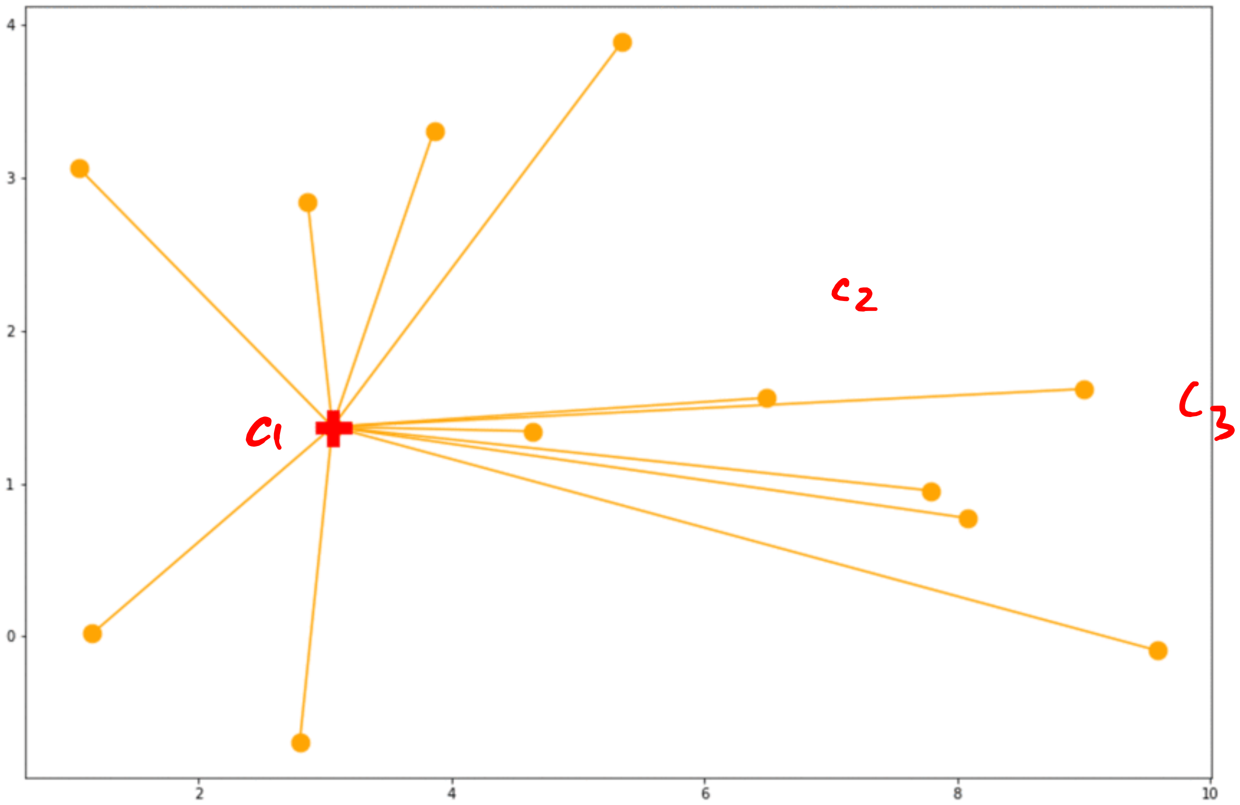




©INSAID All rights reserved.

K-Means Clustering (Example) – Iteration 1

- Step 2: Calculate the **distance** of each **cluster representatives** with **every point** in the data

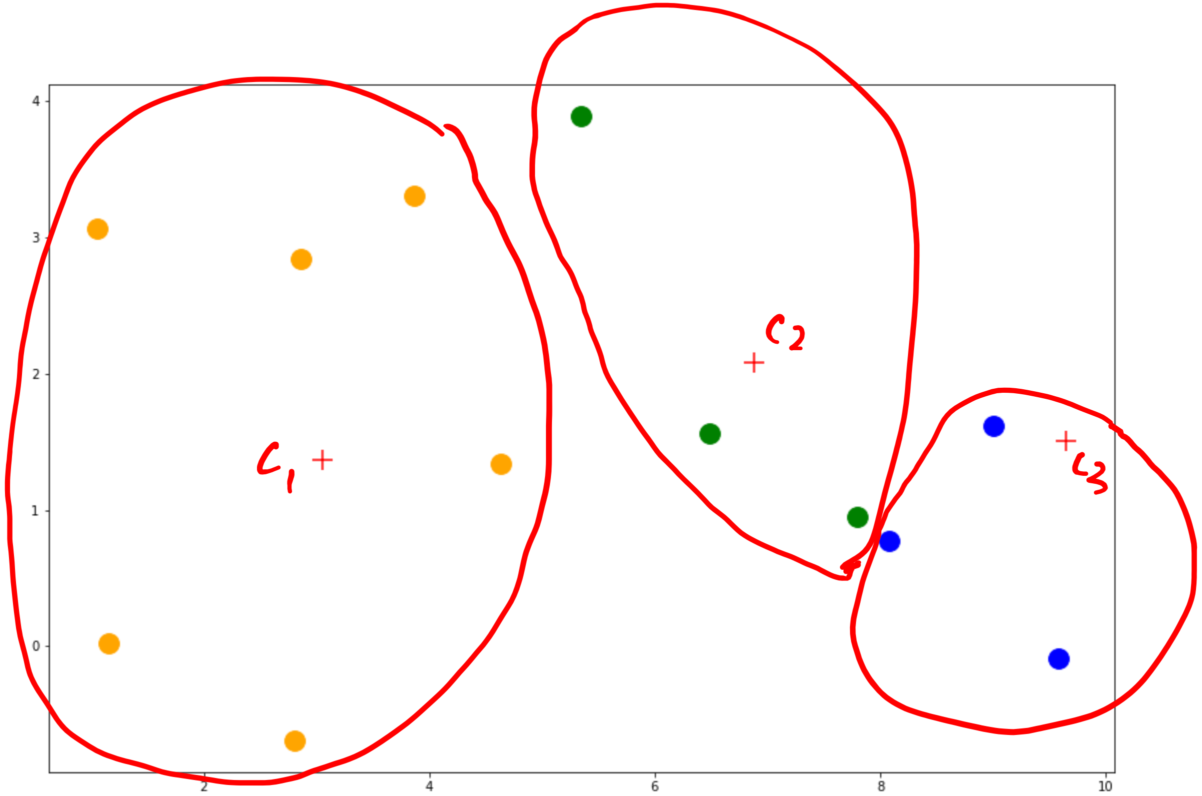


Distance of cluster representative 1 with all the data points

©INSAID All rights reserved.

K-Means Clustering (Example) – Iteration 1

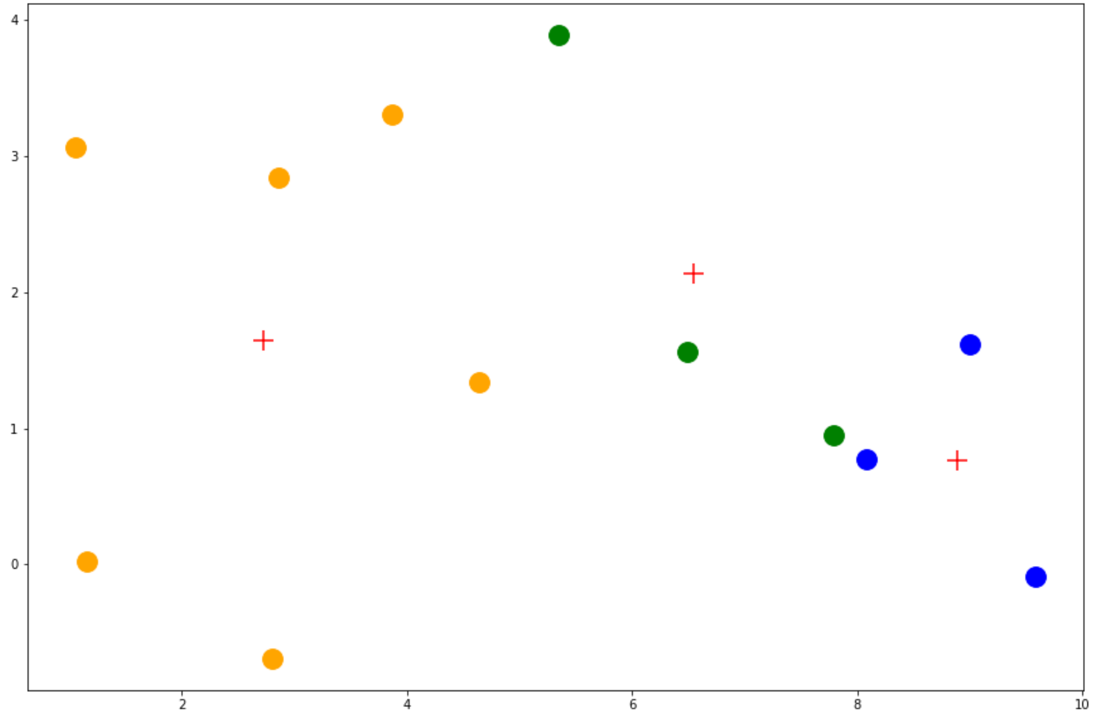
- Step 3: **Assign** the **nearest** data points to the chosen **cluster representatives**.



©INSAID All rights reserved.

K-Means Clustering (Example) – Iteration 1

- Step 4: Calculate the **new** center **representatives** by taking the **average** of the **points** in **each cluster**.
- The total **inertia** of these clusters is **20.65.**

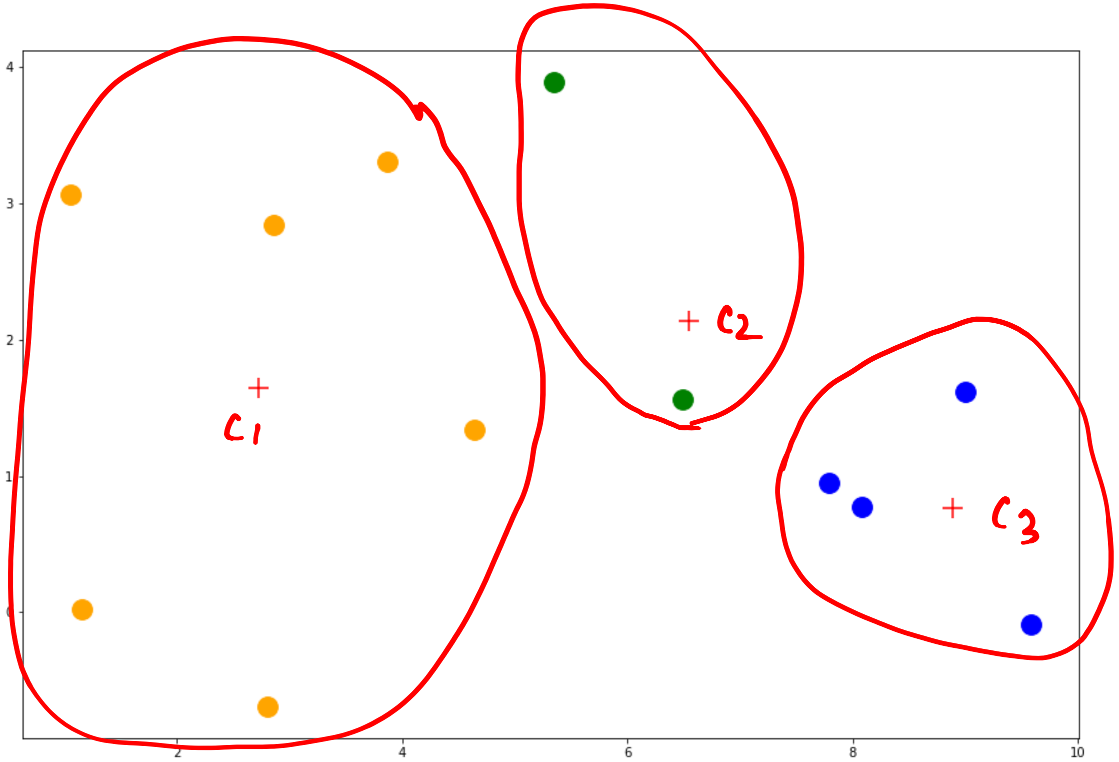


©INSAID All rights reserved.

INSAID

K-Means Clustering (Example) – Iteration 2

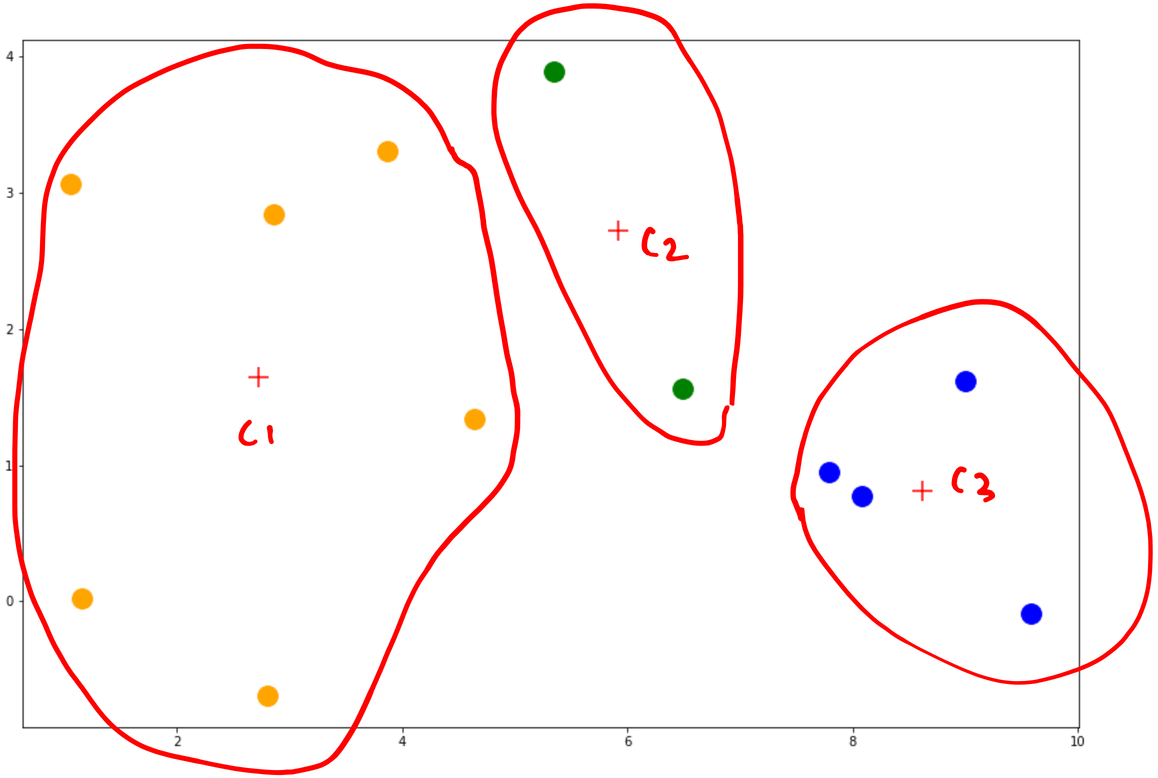
- We will repeat steps 2 and 3 with the new cluster representatives. The new clusters are:



©INSAID All rights reserved.

K-Means Clustering (Example) – Iteration 2

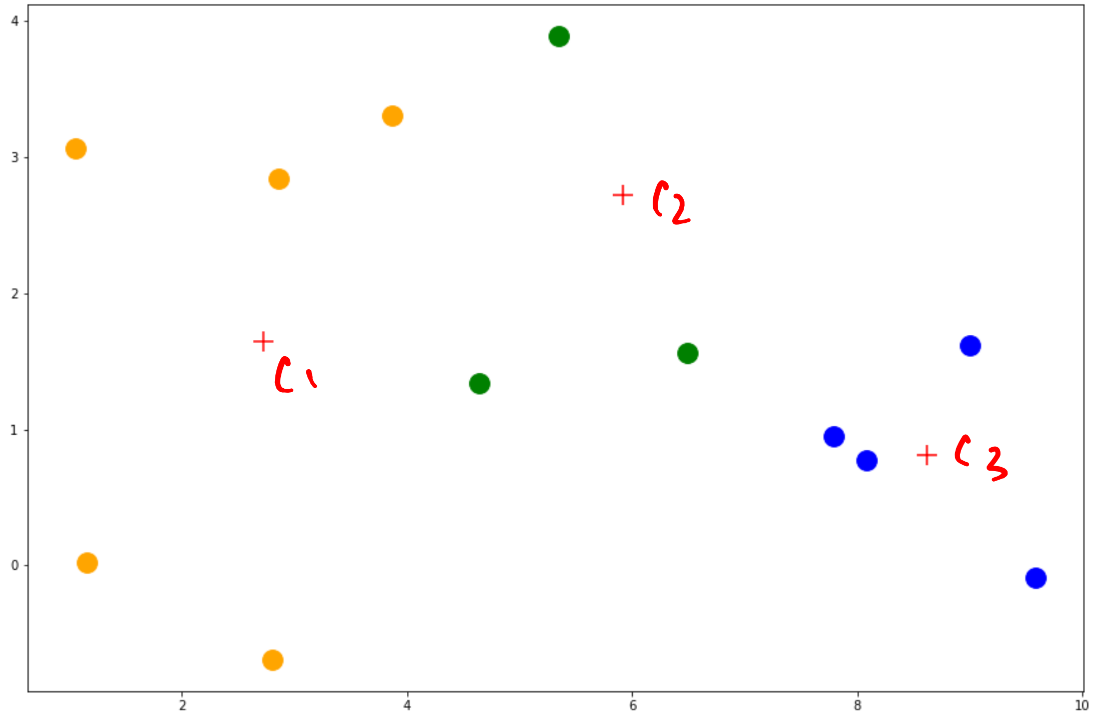
- We will re-calculate the new cluster representatives. The total **inertia** for these clusters are **18.5**.



©INSAID All rights reserved.

K-Means Clustering (Example) – Iteration 3

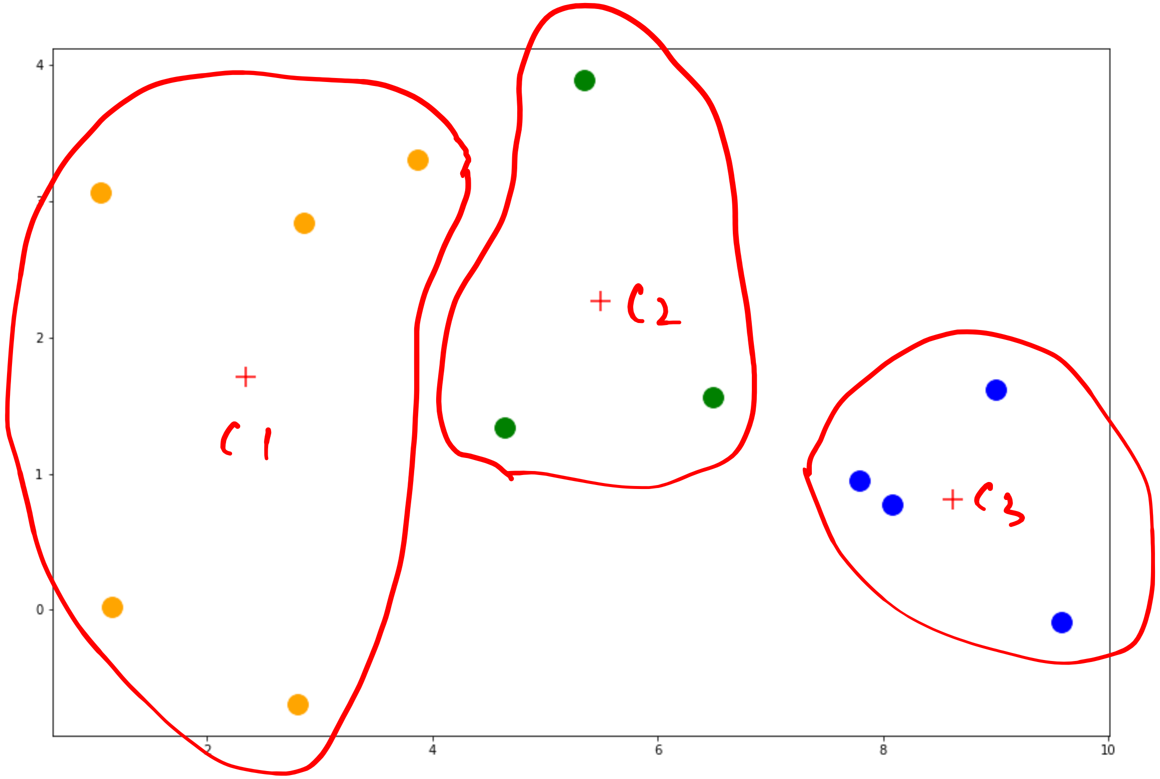
- We will repeat steps 2 and 3 with the new cluster representatives. The new clusters are:



©INSAID All rights reserved.

K-Means Clustering (Example) – Iteration 3

- We will re-calculate the new cluster representatives. The total **inertia** for these clusters are **18.07**.

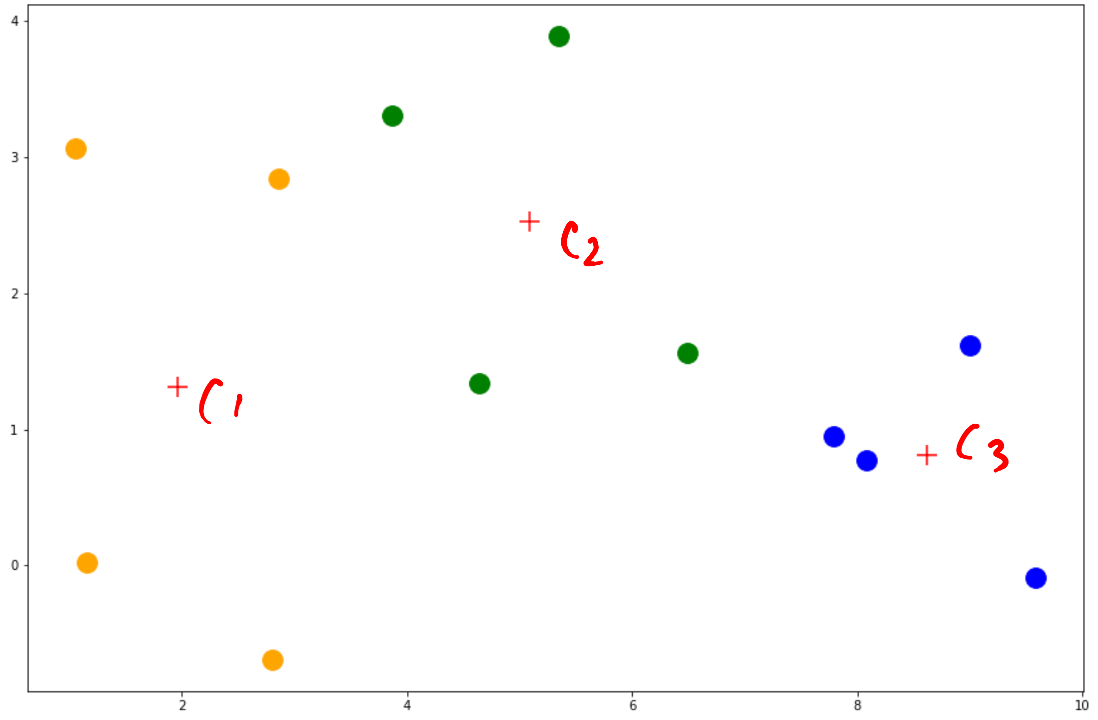


©INSAID All rights reserved.

INSAID

K-Means Clustering (Example) – Iteration 4

- We will repeat steps 2 and 3 with the new cluster representatives. The new clusters are:

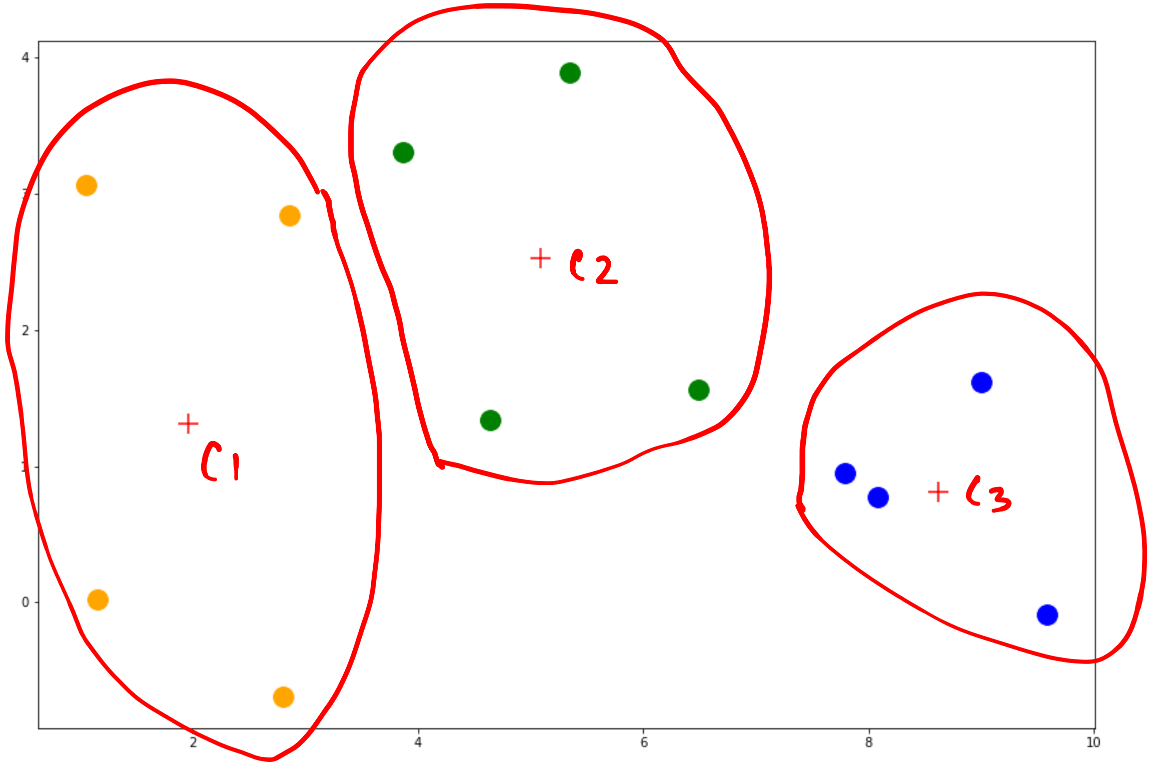


©INSAID All rights reserved.

INSAID

K-Means Clustering (Example) – Iteration 4

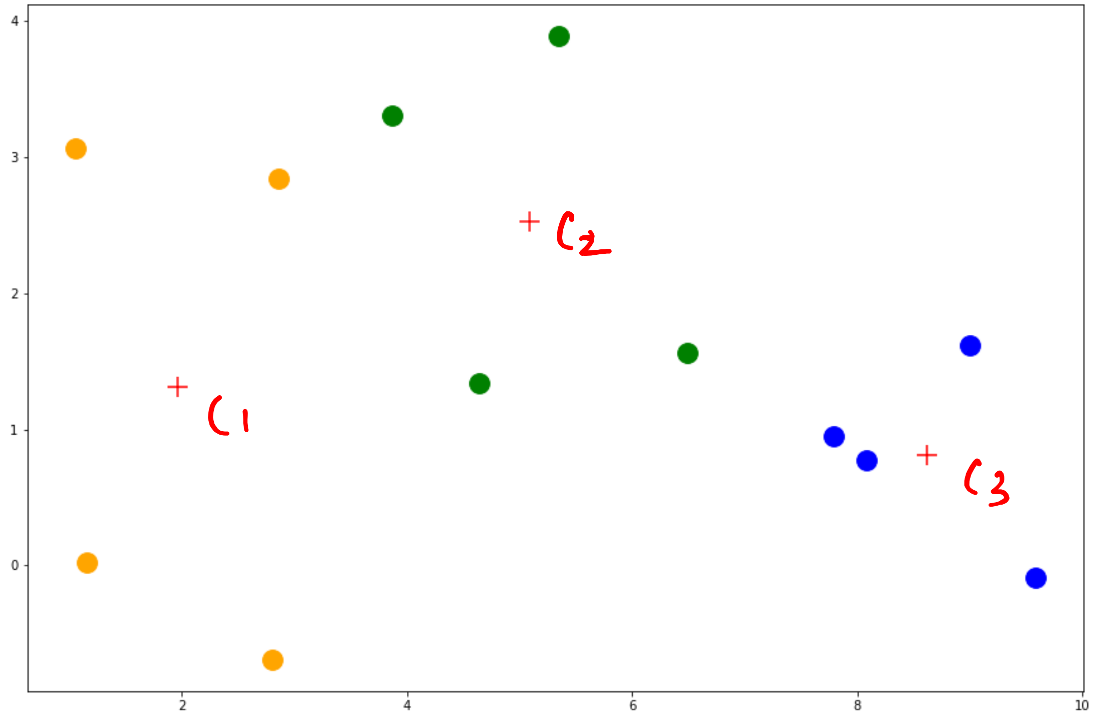
- We will re-calculate the new cluster representatives. The total **inertia** for these clusters are **17.25.**



©INSAID All rights reserved.

K-Means Clustering (Example) – Iteration 5

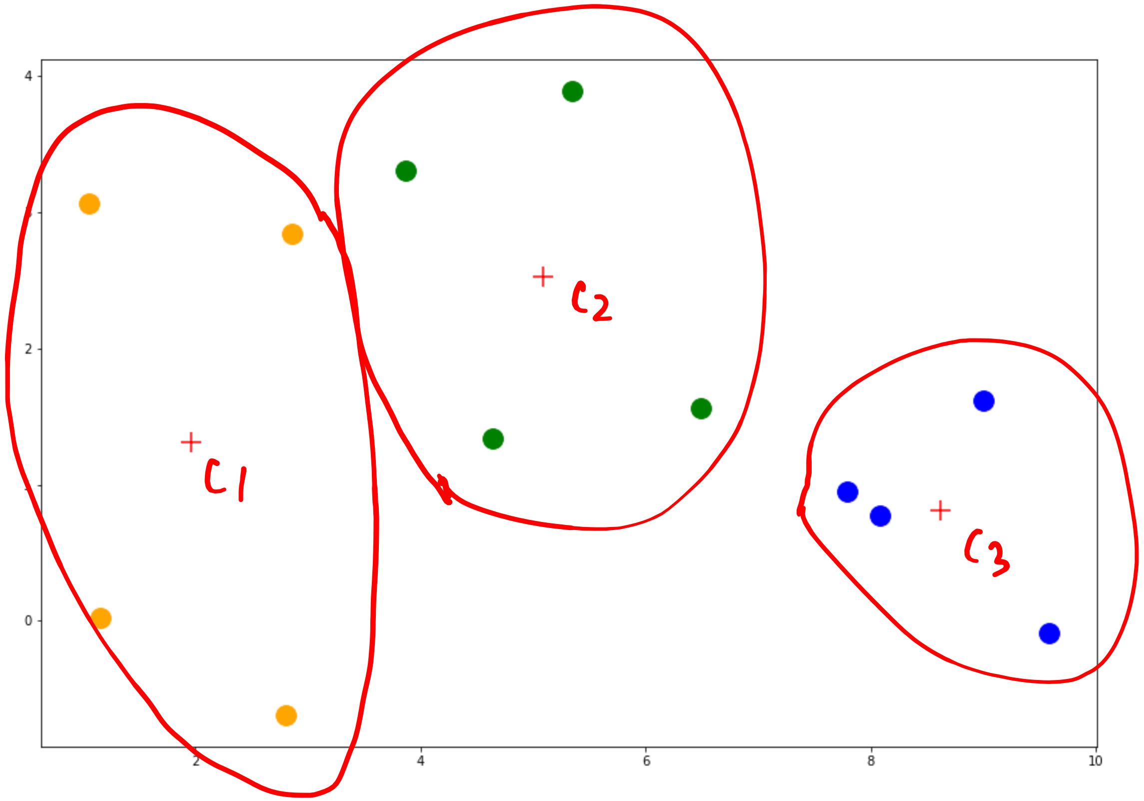
- We will repeat steps 2 and 3 with the new cluster representatives. The new clusters are:



©INSAID All rights reserved.

K-Means Clustering (Example) – Iteration 5

- We will re-calculate the new cluster representatives. The total **inertia** for these clusters are **16.83**.
- We have reached our estimated number of **iterations** and the cluster **representatives** remain the **same**.
- We will **stop** the iterations here. These are our **final clusters**:



©INSAID All rights reserved.

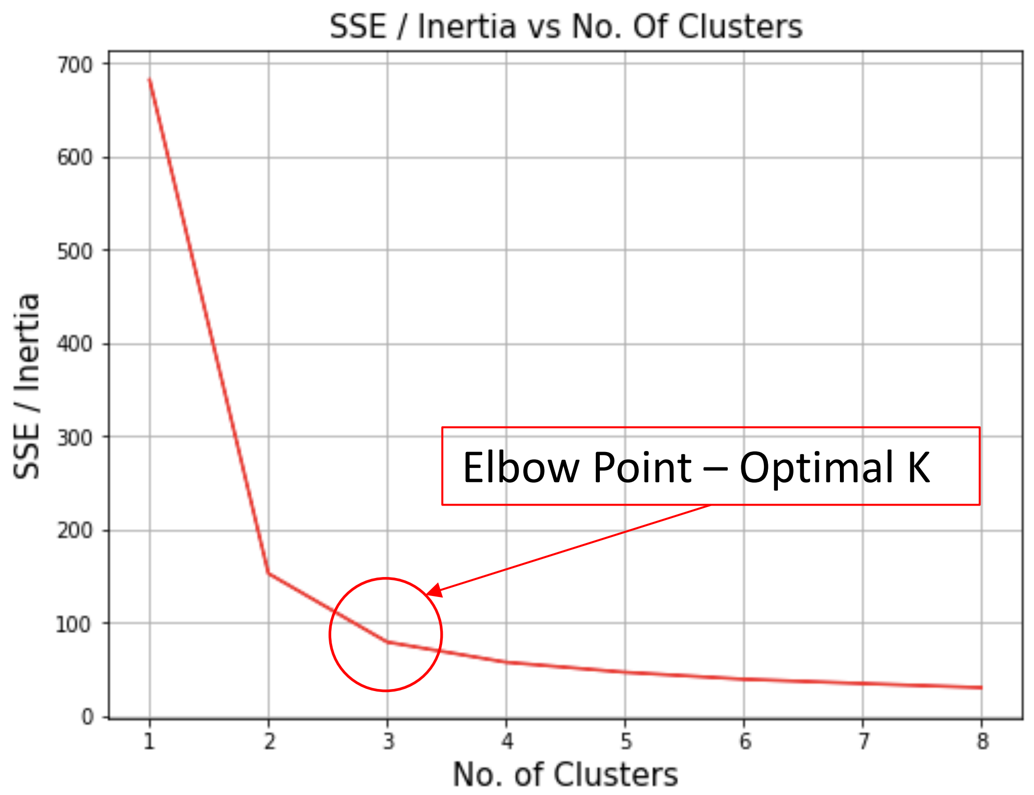
Agenda

©INSAID All rights reserved.

INSAID

Elbow Method – Optimal Value

- It is one of the most popular methods to determine the **optimal value of K**.
- We use it to choose a K when we observe **negligible change** in the **inertial values** between different values of K.



©INSAID All rights reserved.

Agenda

©INSAID All rights reserved.

Advantages

- Relatively **simple** to implement.
- **Scales** to large data sets.
- Guarantees **convergence**.
- Can **warm-start** the positions of centroids.
- Easily **adapts** to new examples.
- **Generalizes** to clusters of different shapes and sizes, such as **elliptical** clusters.

©INSAID All rights reserved.

Agenda

©INSAID All rights reserved.

Limitations

- Choosing k **manually**. (Unsupervised ML)
- **Dependent** on **initial** values.
- **Prone** to varying **sizes** and **density** of clusters.
- Prone to **outliers**.
- Prone to **Curse of Dimensionality**. (huge no of cols.)
- **Convergence** to a **local minimum** may produce **wrong** results.

©INSAID All rights reserved.

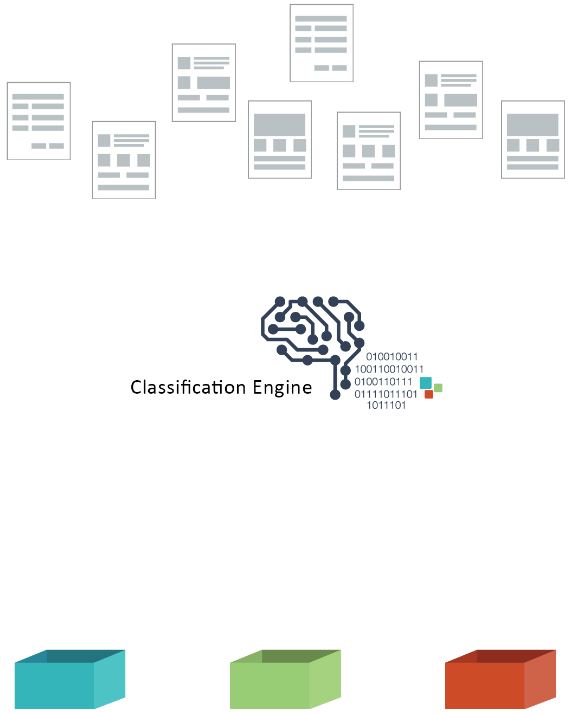
Agenda

©INSAID All rights reserved.

INSAID

Applications – Document Classification

- By classifying text, we are aiming to assign **multiple categories** to a **document**, making it easier to manage and sort.
- This is especially useful for publishers, news sites, blogs or anyone who deals with a lot of content.



©INSAID All rights reserved.

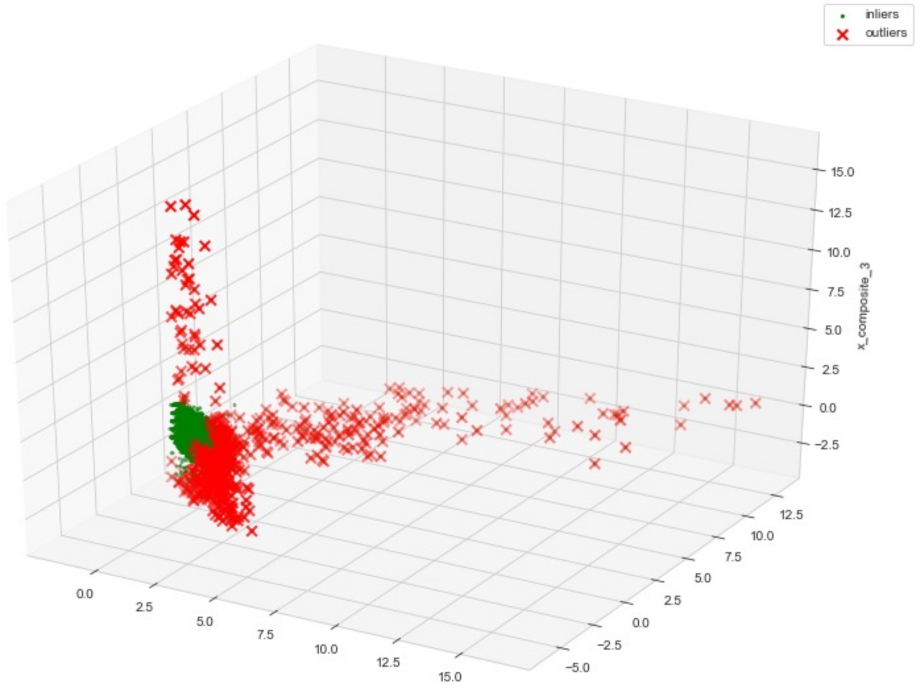
Applications – Image Compression

- Image compression is reducing the size that an image takes while storing or transmitting.
- While compressing, the **colors** are **clustered** towards some **major colors** by grouping them towards the major colors.

©INSAID All rights reserved.

Applications – Outlier Detection

- In the k-means based outlier detection technique the **data points** are partitioned in to k groups by assigning them to the **closest cluster centers**.



©INSAID All rights reserved.

Applications – Image Segmentation

- Image Segmentation is the process of assigning a **label** to **every pixel** in an image such that **pixels** with the **same label** **share** certain **characteristics**.

©INSAID All rights reserved.

Thank you

©INSAID All rights reserved.