Offline 4: K-means Clustering

Dataset preparation:

Use dataset ‘dataset.txt’ in the given folder.

Train:

```
K=4
Load dataset into 2D list "Data"
Randomly select K different data points from “Data” and store them into 2D list "Centers"
Initialize a 2D list named "Clusters" which contains K 1D lists for the K centers
for each sample/ data point "S" in "Data":
    identify the center “C_i” that is the closest to “S”
    Append "S" in "i"th list of "Clusters"
itr = 1, “Shift” = 0
while True:
    for each 1D list "L" in "Clusters":
        Determine the average of the data points. This is the new center of this list.
        Update the center of this list in “Centers”
    if itr > 1 and "Shift" < 50 break (convergence)
    “Shift” = 0
    Initialize a 2D list named "Temp_Clusters" which contains K 1D lists for the K centers
    for each sample/ data point "S" in "Data":
        identify the center “C_i” that is the closest to “S”
        Append "S" in "i"th list of "Temp_Clusters"
        if S belongs to different clusters in “Clusters” and “Temp_Clusters” then
            “Shift” = “Shift” + 1
    Now "Temp_Clusters" 2D list contains K 1D lists
    Assign "Temp_Clusters" to "Clusters"
    itr = itr + 1
"Clusters" will contain your desired clusters and "Centers" will contain your desired centers at
the end of loop
Plot them with appropriate color
“inertia” = 0
for each 1D list "L" in "Clusters":
    “inertia” = “inertia” + sum of distances-square of data points of “L” from the center
```

Report:

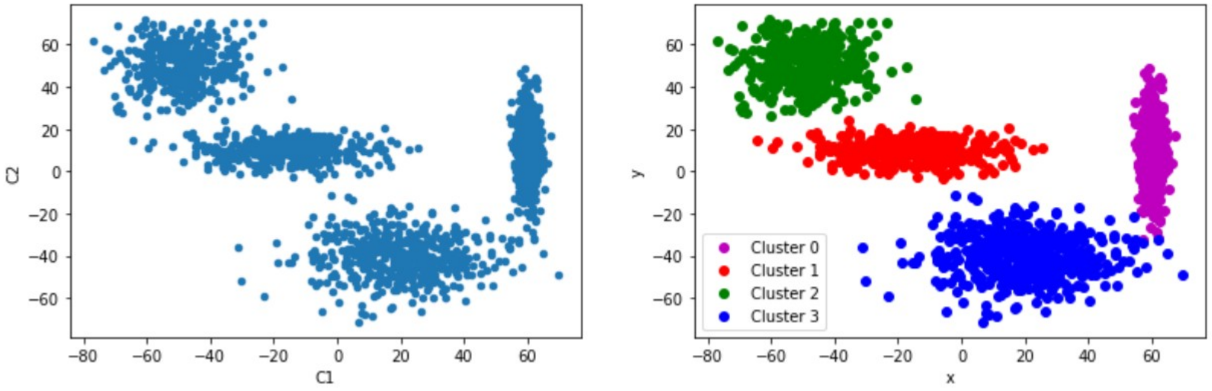
- □ Plot the data for *K* = 2, 4, 6, 7 and note down inertia.

Instruction

- Submit a .ipynb file and a report (report template) .pdf file.
- **You must follow the given algorithm**
- **DO NOT USE LIBRARIES SUCH AS: "Sklearn", "Scikit learning" or "pandas" for this assignment**
- **Use your student id as seed**
- **Copying will result in -100% penalty**
- **Your marks will fully depend on your viva and understanding.**
    - Full Algorithm: 16
    - Plotting: 4

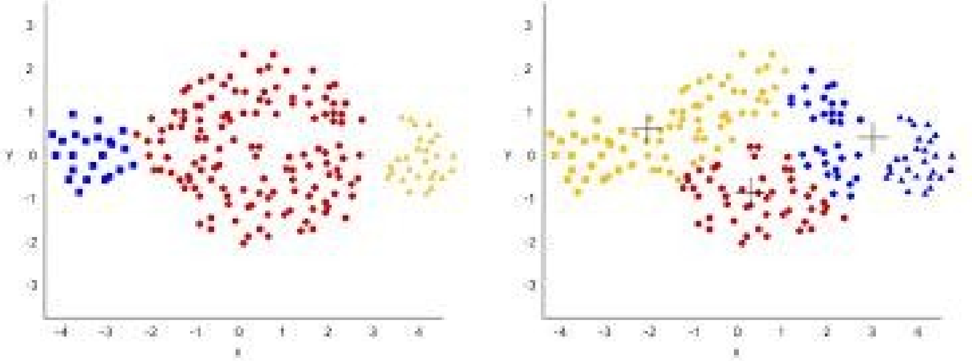
Resources

▶ [k-means clustering]



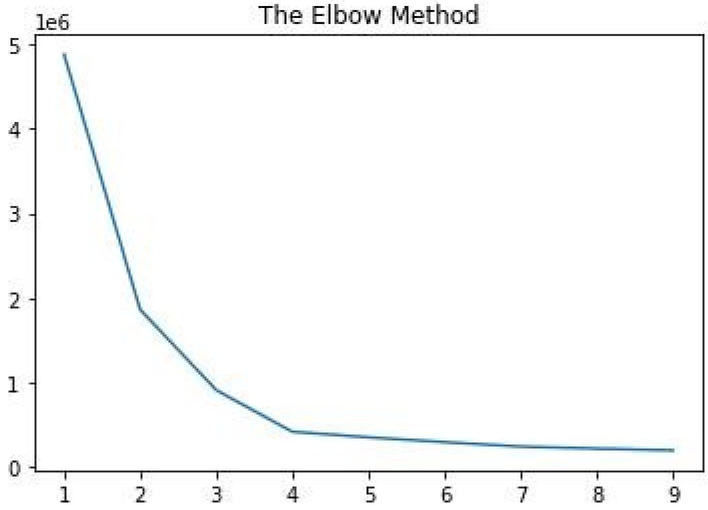
```
Select K random data points as the centers of K clusters
Assign each datapoint to the closest clusters (by calculating the distance from centers).
While True:
    Recalculate the center of the clusters (which is the mean of the data points)
    Reassign each datapoint to the closest cluster
    If no datapoint changes cluster then
        break
```

Limitations:

- Need to know K in advance
- Depended on initial assignment of the centers
- 

How to choose the K?

- Inertia measures how well a dataset was clustered by K-Means. It is calculated by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster. A good model is one with low inertia AND a low number of clusters (K).



-