

# Offline 5: K-means Clustering

## Dataset preparation:

Use dataset [g\\_data](#). Code for loading dataset into 2D python list: [here](#)

## Train:

1.  $K = 4$
2. Load dataset into 2D list "Data"
3. Randomly select  $K$  different data points from "Data" and store them into 2D list "Centers"
4. Initialize a 2D list named "Clusters" which contains  $K$  1D lists for the  $K$  centers
5. **for** each sample/ data point "S" **in** "Data":
6.       identify the center " $C_i$ " that is the closest to "S"
7.       Append "S" in " $i$ "th list of "Clusters"
8.  $itr = 1$ , " $Shift$ " = 0
9. **while** True:
10.       **for** each 1D list "L" in "Clusters":
11.               Determine the average of the data points. This is the new center of this list.
12.               Update the center of this list in "Centers"
13.       **if**  $itr > 1$  **and** " $Shift$ " < 50: **break** (convergence)
14.       " $Shift$ " = 0
15.       Initialize a 2D list named "Temp\_Clusters" which contains  $K$  1D lists for the  $K$  centers
16.       **for** each sample/ data point "S" **in** "Data":
17.               identify the center " $C_i$ " that is the closest to "S"
18.               Append "S" in " $i$ "th list of "Temp\_Clusters"
19.               **if** S belongs to different clusters in "Clusters" and "Temp\_Clusters" **then**
20.                       " $Shift$ " = " $Shift$ " + 1
21.       Now "Temp\_Clusters" 2D list contains  $K$  1D lists
22.       Assign "Temp\_Clusters" to "Clusters"
23.        $itr = itr + 1$
24. "Clusters" will contain your desired clusters and "Centers" will contain your desired centers at the end of loop
25. Plot them with appropriate color
26. " $inertia$ " = 0
27. **for** each 1D list "L" **in** "Clusters":
28.       " $inertia$ " = " $inertia$ " + sum of distances 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