Assignment 1
k-Means Clustering

CMPE 481 Data Analysis and Visualization **Due: November 19th, 2021, 6am**

In this assignment, you will implement the k-means clustering algorithm from scratch.

1. Generate a 2D toy/simple dataset suitable for clustering. See Figure 1 for an example.
2. Apply k-Means algorithm to the dataset and plot moving cluster centers 1) for the first three iterations and 2) for the last iteration. See Figure 2 for a sample plot for the first two steps. Use two different k values, e.g., k=3 and k=7.
3. Plot objective function vs iteration count for all iterations. See Figure 3.
4. Compare your final clustering with the output obtained by the scikit-learn library. Use scikit's k-Means algorithm and plot final cluster centers. Explain the differences between your results and scikit's outputs. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
5. Implement a method to find the best k automatically. You need to perform research for this. Explain the method in one page in detail and provide references. Show the final cluster centers found by this method on your dataset.

Perform all these steps with another -difficult- dataset, possibly with a non-convex or elongated structure. Be creative while designing this dataset ☺

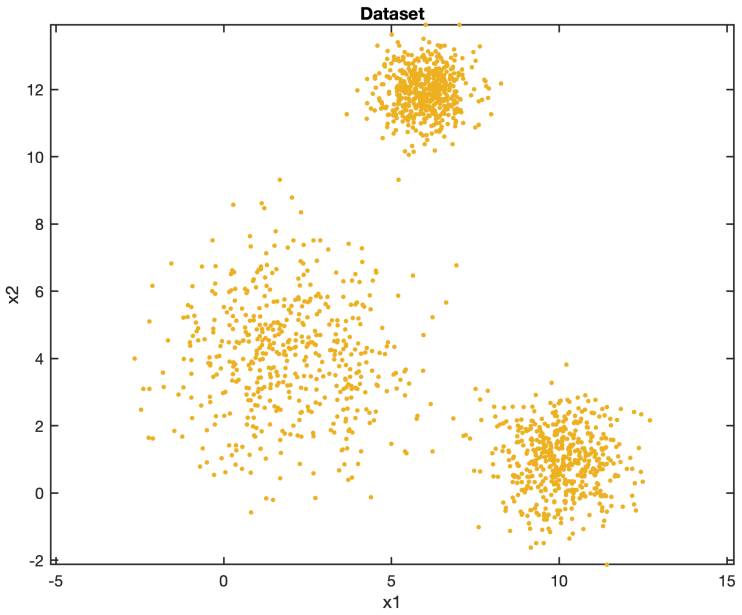
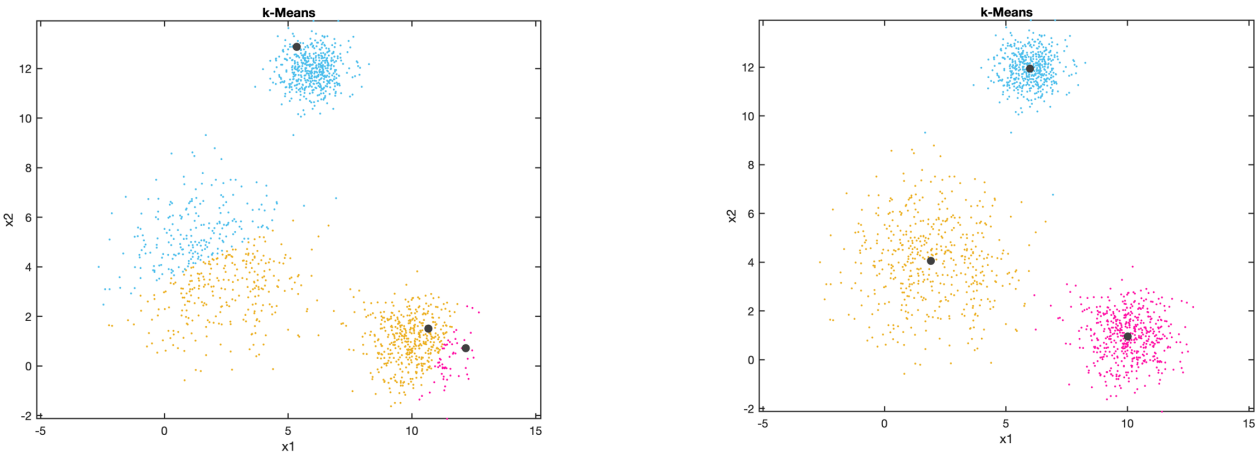


Figure 1. A sample dataset with apparent clusters.



Initial Cluster Centers Iteration 1

Figure 2. Initial cluster centers (left) and clusters centers after the first iteration (right).

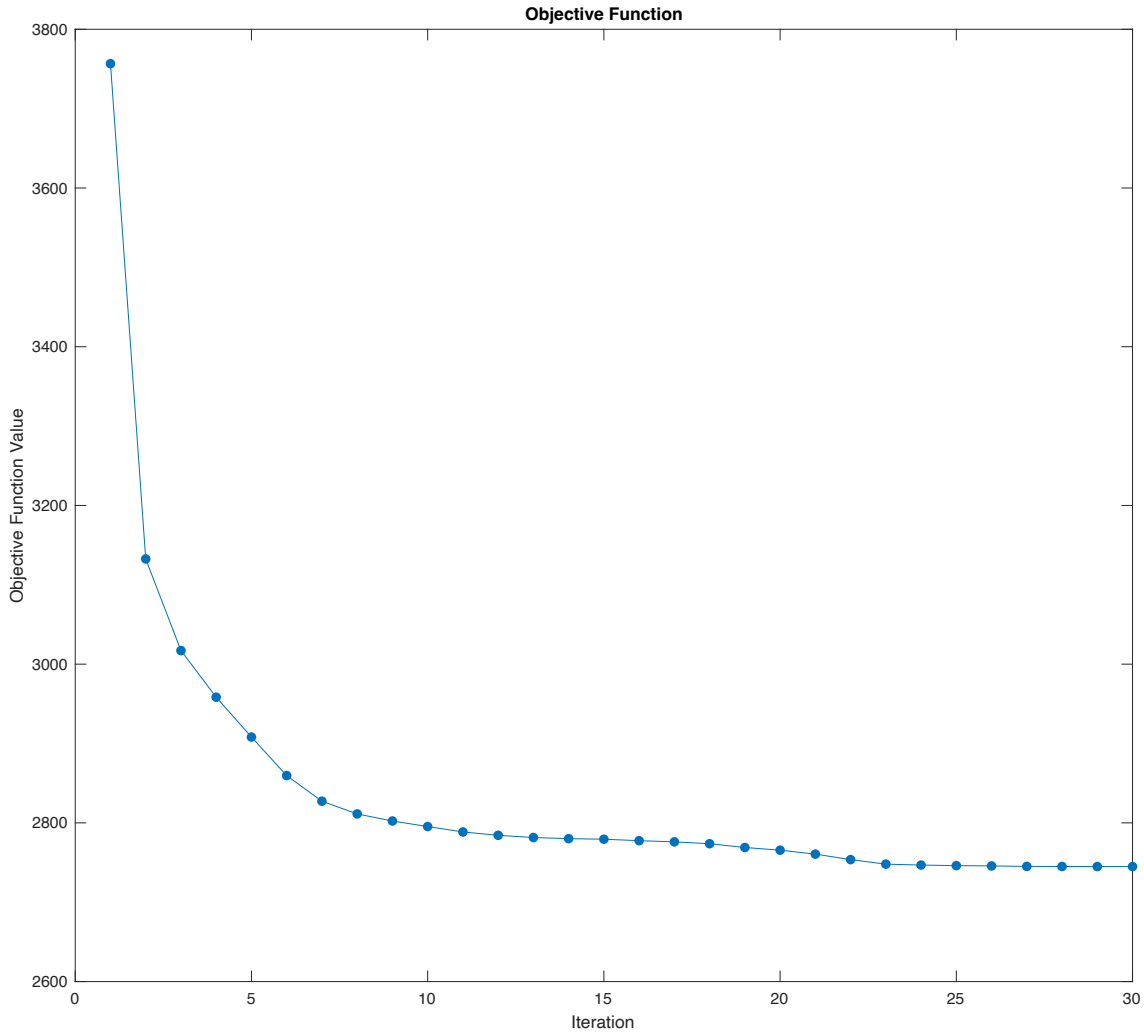


Figure 3. Sample plot for the objective function value (y-axis) vs iteration count (x-axis).

Notes

- If the Euclidean distance between two consecutive cluster centers (summed over all clusters) is less than a very small threshold, k-Means algorithm should stop. Choose the threshold value yourself by experimenting with different values.
- Do not use any machine learning libraries to implement the k-Means algorithm. You should implement the algorithm from scratch.

Evaluation Criteria

	Points
k-Means algorithm	30
Finding the best k	20
Report (Contents, completeness, format, etc.)	40
Compliance to Submission Rules (Directory structure, file formats/naming, organization, etc.)	10
TOTAL	**100**

Submission Guide

Submission Files
Submit a single compressed (.zip) file, named as name_surname.zip, to the Moodle. It should contain all source code files (under the \code directory), report (in PDF format, under the \report directory) and all other files if needed (under \misc directory)

File Naming
Name your report as name_surname.pdf. Name the main code which is used to run your assignment as assignmentX.py, where X is the assignment number and .py is the extension for Python, given as an example.

Late Submission Policy
Maximum delay is two days. Late submission will be graded on a scale of 50% of the original grade.

Mandatory Submission
Submission of assignments is mandatory. If you do not submit an assignment, you will fail the course.

Plagiarism
Leads to grade F and YÖK regulations will be applied