Project 2: Unsupervised Learning (K-means) Report

CSE 575: Statistical Machine Learning (Summer 2024) (Samira Ghayekhloo)

Group 2

Project Team: Arjun Dadhwal, Ching-Chun Yuan, Jeffrey Li

Table of Contents

Introduction

This report presents our implementation of the unsupervised learning K-means algorithm, and its application to a given 2-D dataset, clustering the given samples into predicted classes.

We implemented two different strategies for choosing the initial cluster centers for the algorithm: The first involved randomly picking the initial centers of the clusters from the given samples and the second involved picking the first cluster center randomly but for subsequent cluster centers, we ensured that its similarity distance from all the previously chosen centers is maximal.

For each of these strategies, we tested the implementation on the data for values of K, the number of desired clusters, ranging from K = 2 to 10, and calculated the objective function for each value of K.

After this we plotted the relations between the different K values and objective function for both the strategies.

Strategy 1

For the first strategy we randomly picked the initial centroids for all the clusters, turn by turn for all values K = 2 to K = 10, where K is the number of clusters that we want to form.

For each value of K, we applied the K-Means clustering algorithm to them.

For the algorithm, we looped through each of the samples present in the dataset, and calculated the euclidean distance between them and the centroids of each cluster as the similarity measure.

Based on this, the sample is assigned to the cluster for which its distance from the cluster's centroid is the least.

After we have iterated through all the samples, we will now recompute the new centroid values for each cluster based on the newly assigned sample values.

If none of the samples were reassigned to a different cluster after this iteration, then we will break the loop. Otherwise, we will keep iterating again until the samples are no longer assigned to new clusters, indicating that

[For reference]

The sum-of-squared-error criterion/cost

D_i be the subset of samples from class i,
Let n_i be the number of samples in D_i, and m_i the mean of those samples.

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

The sum of squared error is:

$$J_e = \sum_{i=1}^{C} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

Well-separated and compact data tend to give small errors.

Find a good clustering i.e. to find the partition of the data that minimizes J_e.

Find an optimal set of centroids.

K-Means Clustering

Given n data samples.

Partition them into k clusters/sets D_i, with respective center/mean vectors U_1, U_2,.. so to minimize:

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in D_i} ||\mathbf{x} - \boldsymbol{\mu}_i||^2$$

Strategy 2

The second strategy used is selecting the first center randomly. Then for each subsequent center, choose a point that maximizes the minimum Euclidean distance to any previous chosen center.

$$distance(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

For the clustering the k-means algorithm is applied, each point will be assigned to the nearest center.

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in D_i} ||\mathbf{x} - \boldsymbol{\mu}_i||^2$$

Based on the points, the mean of the points in each cluster will be calculated to update the new center. This will be looped until convergence is reached, where when recomputing the centers will result in the new centers not changing significantly. Otherwise, if the maximum iterations are reached.
n represents the number of points in each cluster D, while x represents the sum of all points that belong to cluster D.

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

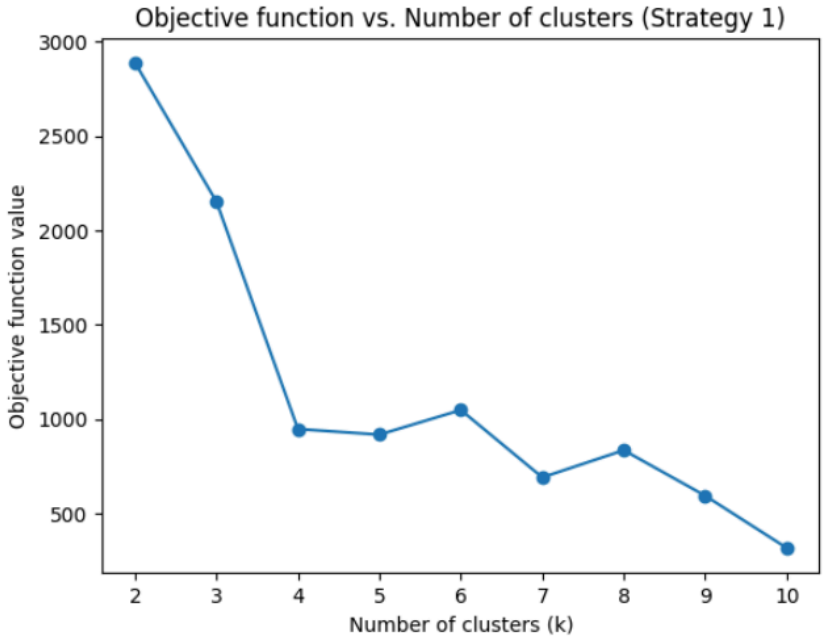
To calculate the Objective Function, similar to strategy 1, the sum of the squared error would be used.

$$J_e = \sum_{i=1}^{C} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

Results

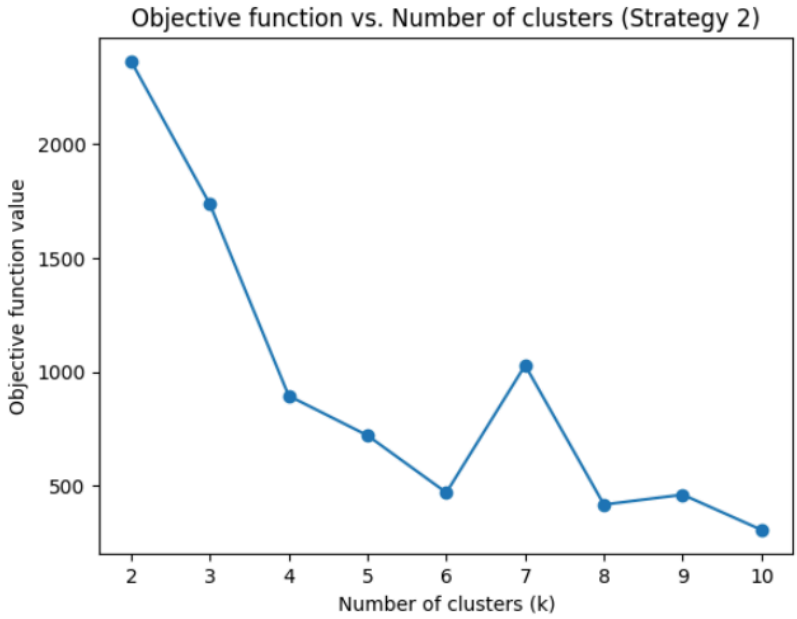
[Strategy1]

k	Objective Function Value
2	2887.672292
3	2151.746116
4	948.053818
5	918.471017
6	1048.875360
7	691.960933
8	835.906565
9	594.791195
10	318.438746



[Strategy2]

k	Objective Function Value
2	2364.765049
3	1736.520053
4	894.782278
5	720.451028
6	469.380813
7	1029.134617
8	416.722391
9	460.368877
10	305.573846



Conclusion

We successfully implemented the k-means algorithm using two different strategies for initializing cluster centers. The objective function values were computed, and we successfully plotted the results.