K-means Clustering

Jeong Hoon Shin

Abstract. In this assignment, I implement K-means clustering algorithm to cluster given three different datasets.

1 Datasets

The dataset files contain features (in 2D) and class labels. In this assignment, I didn't use class labels since K-means is an unsupervised algorithm and does not need class labels. Scatter plot of the datasets given in Figure 1.

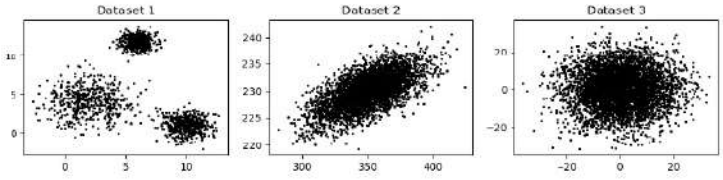


Fig. 1: Three datasets.

2 K-means Algorithm

K-means clustering is a simple and popular type of unsupervised machine learning algorithm, which is used on unlabeled data. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups according to provided features similarity. Algorithm steps for K-means given in Algorithm 1.

Algorithm 1 K-means Algorithm

1. Randomly pick K observations and set them as initial cluster centers
2. Iterate until cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

When N is number of samples and K is number of clusters, K-means algorithm try to minimize objective function which given as following.

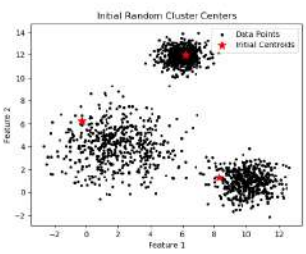
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_i - \mu_k||^2 \qquad \text{Objective Function}$$

3 Clustering Results

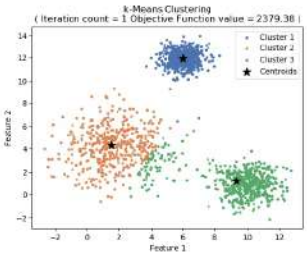
I implement K-Means class for algorithm and cluster three given dataset with following configurations:

1. Dataset1: k=3, k=7
2. Dataset2: k=2, k=5
3. Dataset3: k=3, k=8

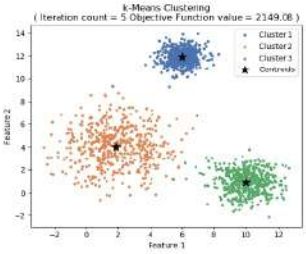
Plot of the initial cluster centers, final cluster centers, iteration count and objective function values for each configuration given in following figures.



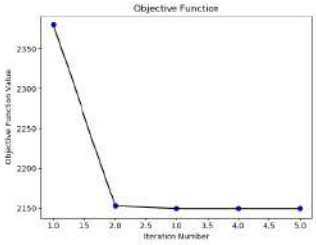
(a) Initial Cluster Centers



(b) First Iteration

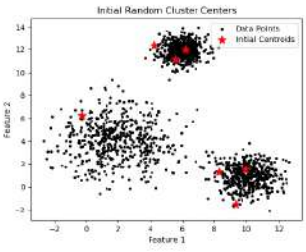


(c) Final Cluster Centers

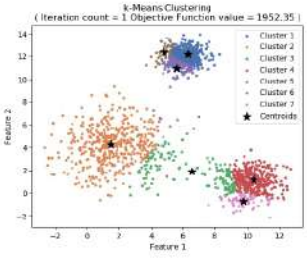


(d) Objective function value

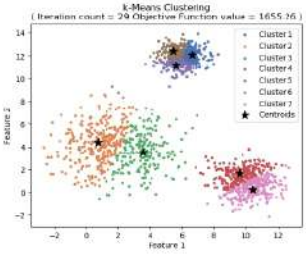
Fig. 2: Result for first dataset when $k = 3$



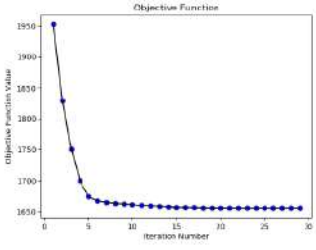
(a) Initial Cluster Centers



(b) First Iteration

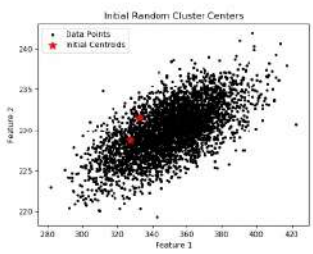


(c) Final Cluster Centers

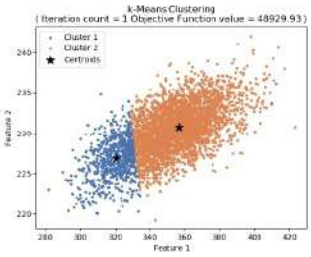


(d) Objective function value

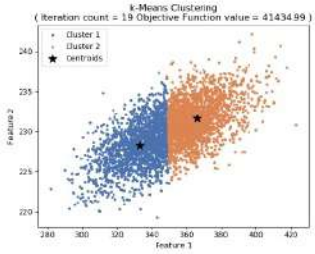
Fig. 3: Result for first dataset when $k = 7$



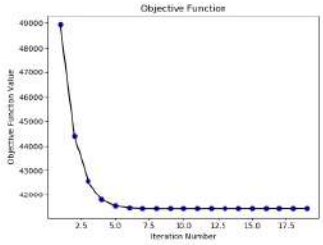
(a) Initial Cluster Centers



(b) First Iteration

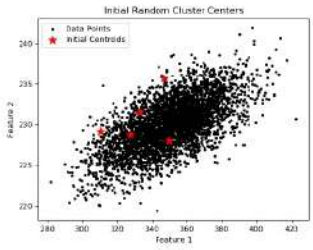


(c) Final Cluster Centers

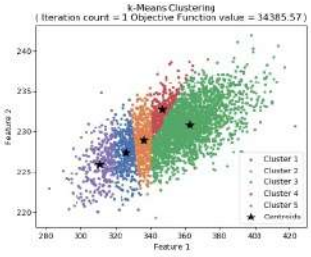


(d) Objective function value

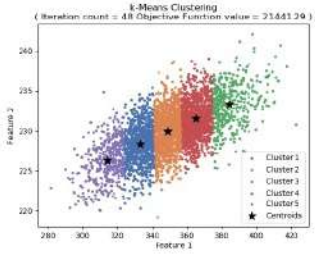
Fig. 4: Result for second dataset when $k = 2$



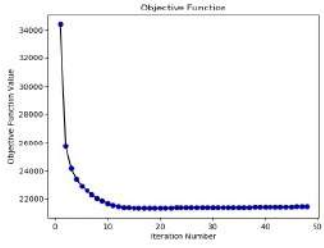
(a) Initial Cluster Centers



(b) First Iteration

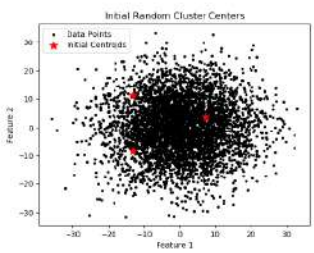


(c) Final Cluster Centers

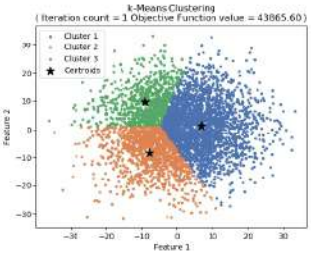


(d) Objective function value

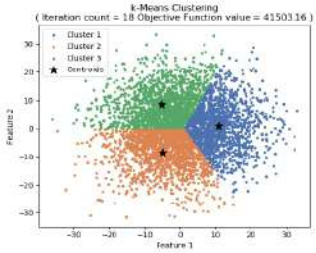
Fig. 5: Result for second dataset when $k = 5$



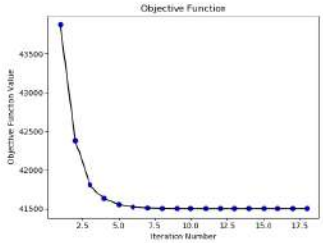
(a) Initial Cluster Centers



(b) First Iteration

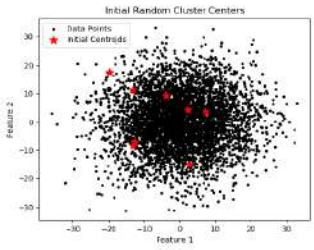


(c) Final Cluster Centers

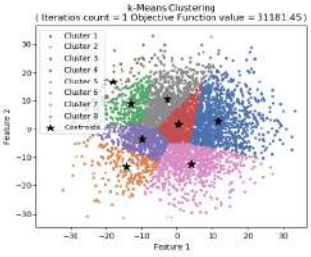


(d) Objective function value

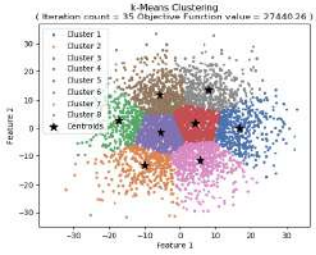
Fig. 6: Result for third dataset when $k = 3$



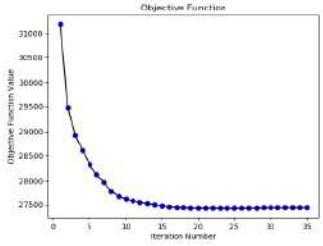
(a) Initial Cluster Centers



(b) First Iteration



(c) Final Cluster Centers



(d) Objective function value

Fig. 7: Result for third dataset when $k = 8$