Machine Learning
K-Means Clustering
Muhammad Husain | 1301153626

1. Problem Analysis

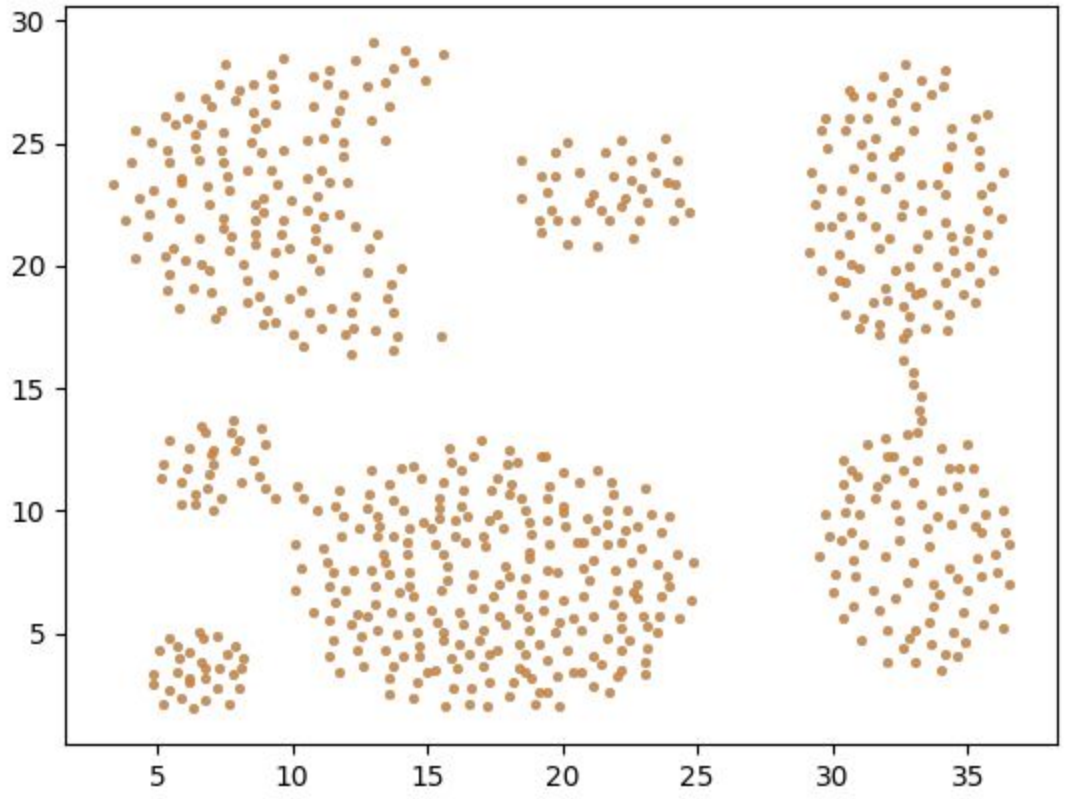
Given two set of data, training and test, we have to create a model for clustering data in test set using K-Means algorithm.

2. Design

Before we are able to cluster the data in data test, first we need to train a model by using training data set. Below is the method of creating a model for k means

a. Initialization

Start with **determining** the number of **k** or cluster. I do this by visualizing the training set into a 2 dimensional scatter plot and guessing the number of cluster by the number of area populated with most data. Also, initializing each centroids initial position, each centroid is positioned by choosing random data as its initial position.

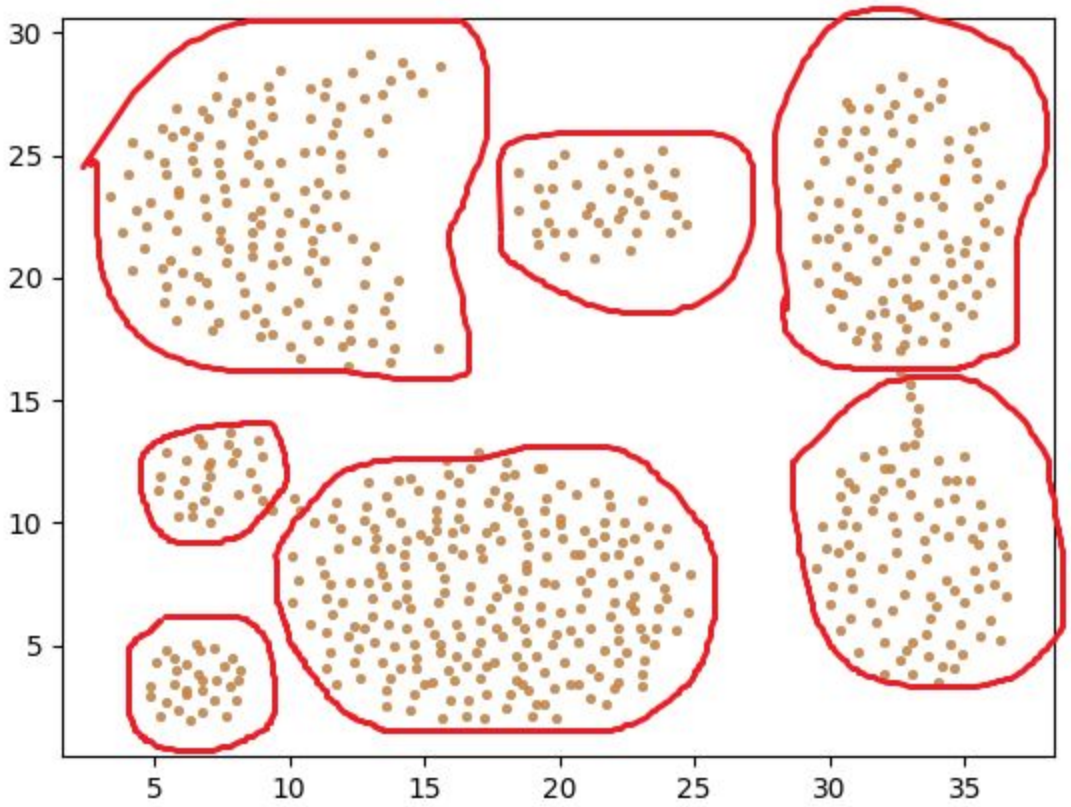


b. Training process

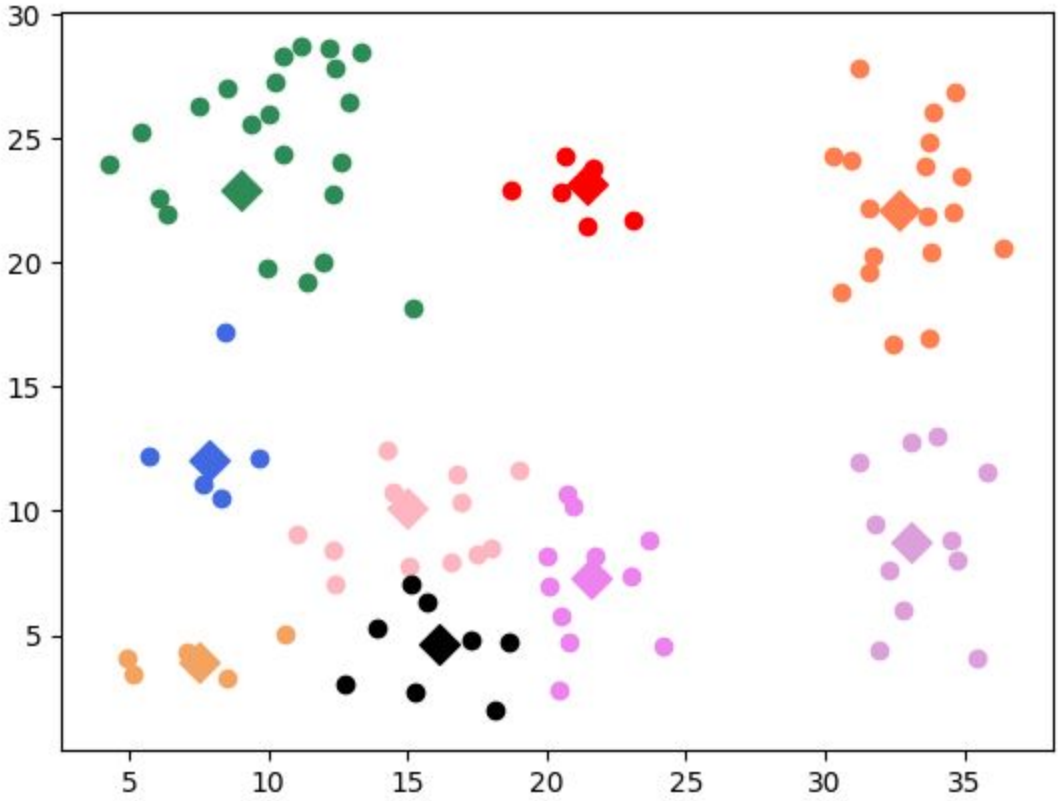
Training process is done by **iterating** through **all** data, on every iteration we will **count** the **euclidean** distance between the data and **each** centroids, this means a data will be computed k times and resulting an array containing k elements of number. Done with computing euclidean distance, next we will pick the index of the minimum value of euclidean distance. This index represents the id of a centroid, in other words the data that we just compute is belong to this cluster of centroid.

Iterate through all data to find each data's cluster, next is to **compute** the **average x** and **average y** of a data that belong to **same** cluster. The average x and average y is the new centroid of that cluster and the old centroid will be **replace** by this new centroids value.

All process above will be repeated until the old centroid and new centroid have the same value. Meaning, no changes in average x and average y of its cluster member. The final centroids are our model to be used on data test set.



3. Experiment Result Evaluation



Based on the experiment result on data test. Every data has grouped to its own cluster. With bare eyes, it looks like all of the test data has been clustered based on our first assumption cluster.