Project2: K-means-Strategy Report
Student ID:1228855799

1.Clustering and K-mean

K-Means is a widely used clustering algorithm. Unlike classification or sequence labeling tasks, clustering involves dividing samples into several categories based on the intrinsic relationship between data without prior knowledge of sample labels, so that the similarity between samples in the same category is high and the similarity between different categories is low. K-Means is an unsupervised algorithm that focuses on similarity and uses distance as the standard for measuring similarity between data objects. The algorithm can discover k different clusters and each cluster's center is calculated by the average of its values. The number of clusters k is given by the user. K-Means is an algorithm that finds k clusters in a given dataset. Each cluster is described by its centroid, which is the center of all points in the cluster.

2.Part 1: K-means_algortihm_Strategy(K-mean)

1. Initialize the constant K and initialize k cluster centers.

Initialize the constant K	initialize k cluster centers
K= 3	[[3.81135136 5.98125361], [5.68845261 8.27229082], [1.20162248 7.68639714]]
K=5	[[4.75184863 4.20214023], [8.61947945 2.98598319], [2.46087695 6.86898874], [3.2492998 5.59125171],[3.75004647 4.90070114]]

2. Repeat the following process until the cluster centers no longer change:

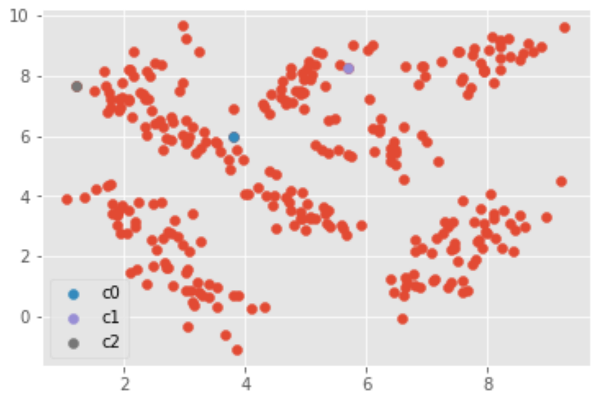
Calculate the distance between each sample and each cluster center and assign the sample to the closest center.

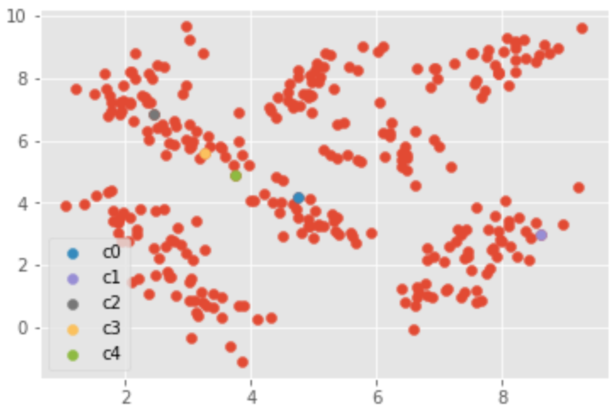
$$d(p,q) = \sqrt{（p1 - q1）^2 + (p2 - q2)^2}$$

Compute the mean of all sample features assigned to each category and use that mean as the new cluster center for each class.

3. Output the final cluster centers and the class to which each sample belongs.

Initialize the constant K	final cluster centers
K=3	[[5.477400,2.254981],[6.497250,7.522973],[2.561464,6.088613]]
K=5	[[2.68198633, 2.09461587],[6.7786424 , 8.07967641],[5.22321274, 4.22502829],[2.87490813, 7.01082281],[7.55616782, 2.23516796]]





4. μ1, μ2,……,μk， then the loss function of the cluster problem can be expressed as:

$$J = \sum_{i=1}^{k} \sum\nolimits_{x \in Di} || x - \mu i ||^2)$$

Initialize the constant K	sum-of-squared-error criterion
K=3	1293.7774523911348
K=5	598.5546443663114

3.party 2: K-means_algortihm_Strategy2(K-mean++)

1. Randomly select a sample point from the dataset as the first initialized cluster center.

Initialize the constant K	initialize k cluster centers
K= 4	[7.67406359 7.37819153]
K=6	[3.01047612 6.54286455]

Selecting the remaining cluster centers:

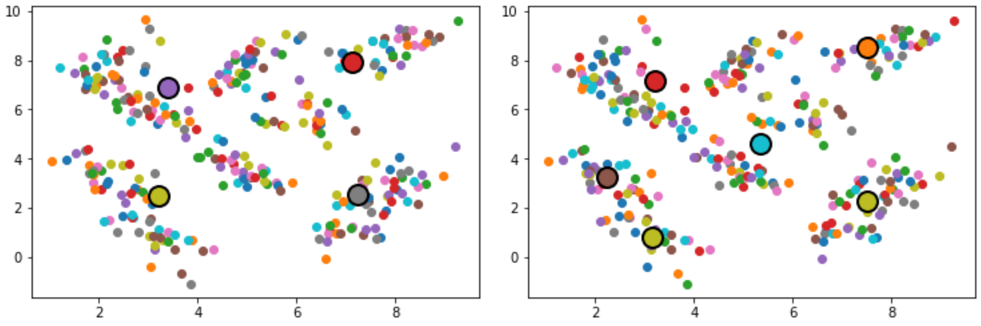
Calculate the distance between each sample point in the dataset and the initialized cluster centers and select the shortest distance, referred to as d_i

Select a new data point as a new cluster center with the principle of selecting a point with a larger distance to be chosen as the cluster center with a higher probability.

Repeat the above process until k cluster centers are determined.

2. Using the K-Means algorithm, calculate the final cluster centers using the k initial cluster centers.

Initialize the constant K	final cluster centers
K=4	[[7.227077, 2.522344],[3.212575 , 2.496581],[3.392621 , 6.892881],[7.135607 , 7.916517]
K=6	[5.335050,4.603777],[3.16906,0.814325],[7.493654,8.524180],[7.493152,2.258441],[3.214633,7.178712],[2.242048,3.251007]



3. sum-of-squared-error criterion

Initialize the constant K	sum-of-squared-error criterion
K= 4	797.9601840789946
K=6	493.03670009094003