ML2 DSEB62 W3- KMEAN - NGUYEN TUAN DUY

NGUYEN TUAN DUY

February 2023

1 Problem 3

How to choose K:
Method 1: Elbow Curve Method
Steps:

1. Select a range of values of K.
2. Perform K-means clustering with all number of K. For each value of K, we calculate the average distances to the centroid across the data points.
3. Plot these points and find the point where the average distance from the centroid falls suddenly.

Method 2: Silhoutte analysis
Silhoutte score is a measure of how similar a data point is within-cluster compared to other clusters.

- The value of the silhouette coefficient is between [-1, 1]
- A score of 1 denotes the best meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.
- The worst value is -1. Values near 0 denote overlapping clusters.

Steps:

1. Select a range of values of K
2. Calculate Silhoutte score for each data point

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{1}$$

   - S(i) is the silhouette coefficient of the data point i.
   - a(i) is the average distance between i and all the other data points in the cluster to which i belongs.

- b(i) is the average distance from i to all clusters to which i does not belong.

3. Calculate average silhouette coefficient.

$$Average\_Silhouette = mean\{S(i)\}$$

4. Plot all average silhouette for different K and compare.

2 Problem 4

Limitations of kmean:

1. Choosing K manually.
2. Being dependent on initial values.
3. Clustering data of varying sizes and density.
4. Clustering outliers.
5. Scaling with number of dimensions.

Example:
(1) Choosing K manually:

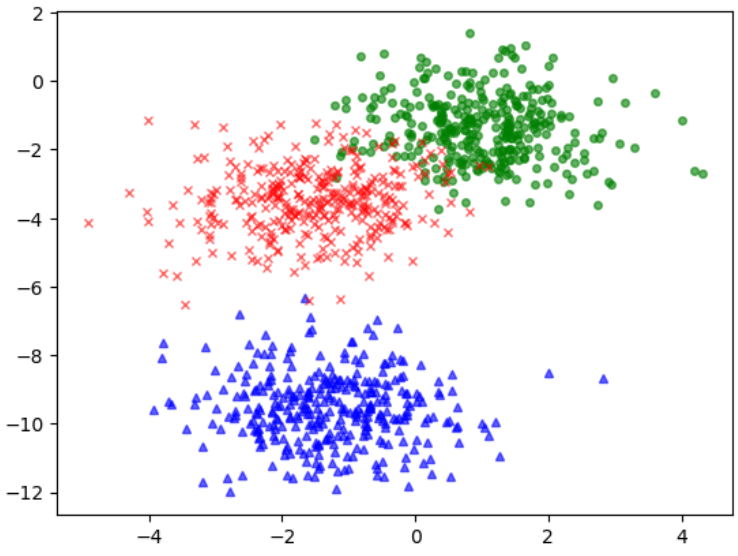

Figure 1: Actual dataset.

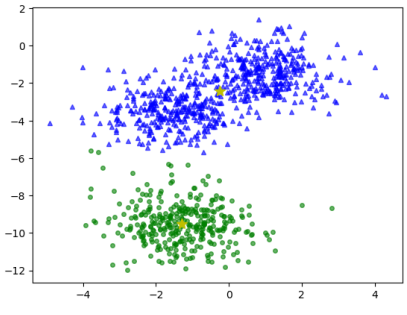

Figure 2: 2 clusters

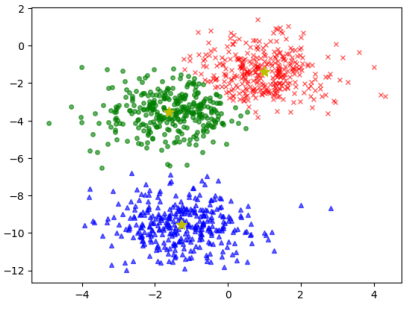

Figure 3: 3 clusters

For this dataset, when fitting Kmean with both 2 and 3 clusters, it looks reasonable in both cases, it may be hard to choose optimal K manually.

(2) Being dependent on initial values:

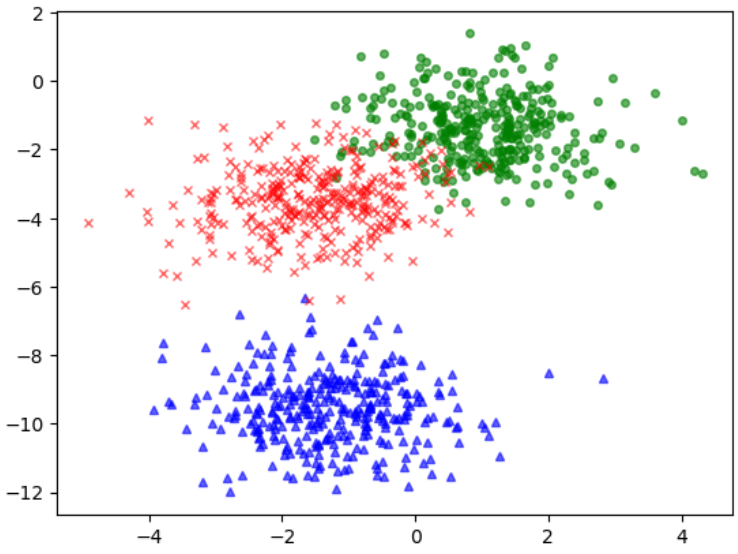

Figure 4: Actual dataset.

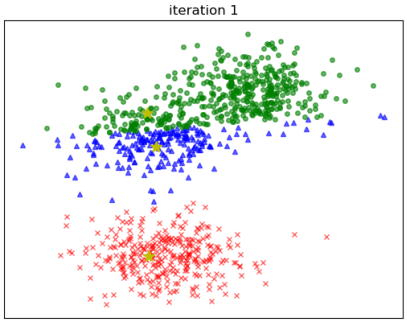

Figure 5: first try: initial centers.

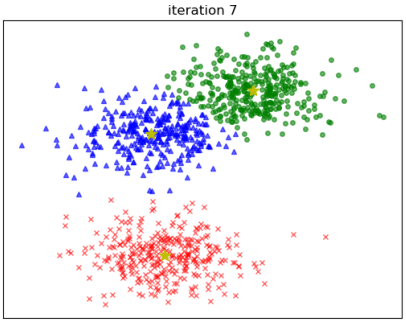

Figure 6: first try: result.

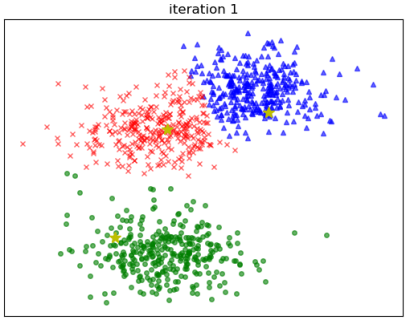

Figure 7: second try: initial centers.

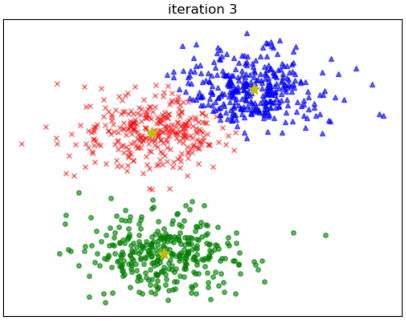

Figure 8: second try: result.

Different choice of initial centers leads to different number of iterations required to achieve convergence on the center (on first try it took 7 iterations whereas on the second attempt it only took 3). For larger amount of centers it takes more iterations to complete and takes more resources
(3) Clustering data of varying sizes and density:

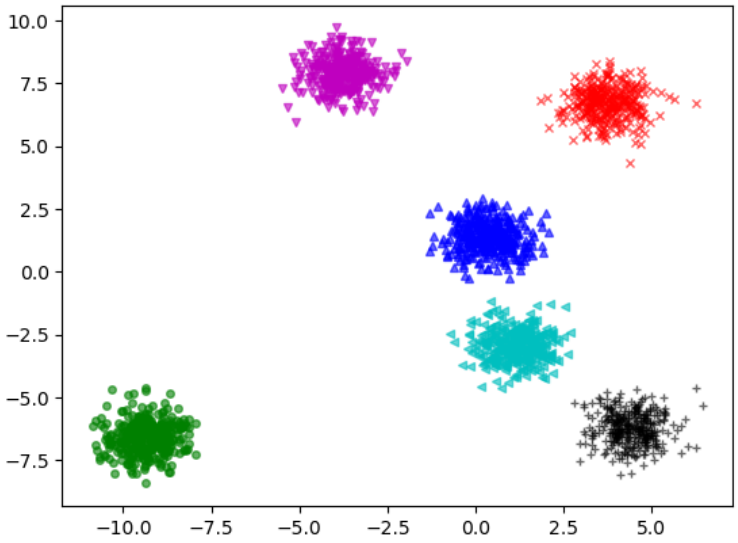

Figure 9: Actual dataset.

The densities between clusters is varied (blue cluster is very close to cyan, green cluster is very far from other clusters)

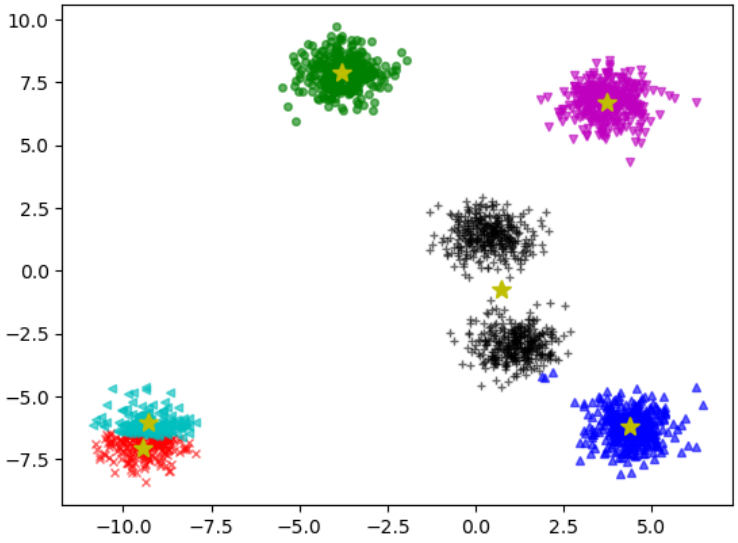

Figure 10: varying size result.

Result: original green cluster is divided into two clusters while original blue

and cyan clusters are wrapped into one
(4) Clustering outliers:

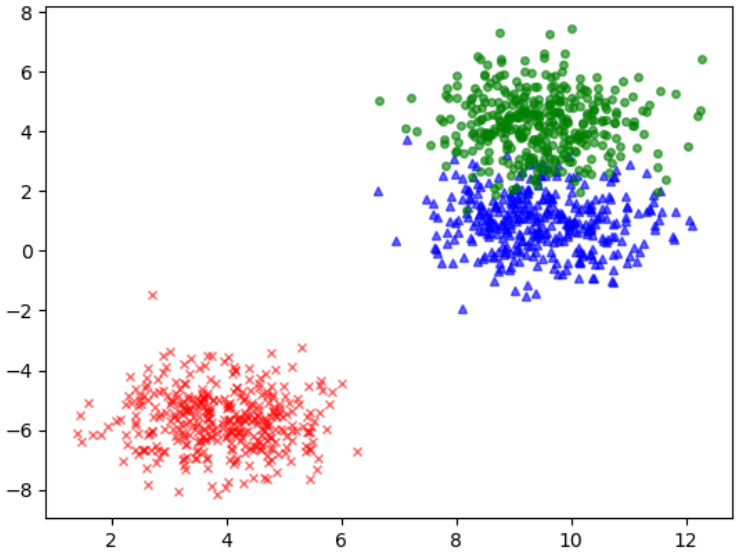

Figure 11: Actual dataset.

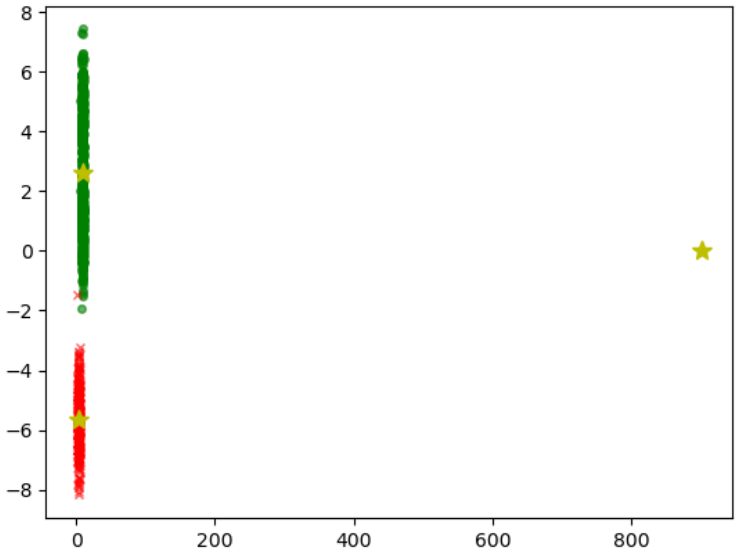

Figure 12: Dataset with one outlier.

When added one outlier, the point becomes its own cluster