**Homework 4: Clustering Techniques**

Student ID: Student Name:

Lectured by: Shangsong Liang

Machine Learning and Data Mining

Sun Yat-sen University

Deadline for Submission: April 5, 2022

 4月5号上课前提交

**Exercise 1: Implement K-Means Manually**

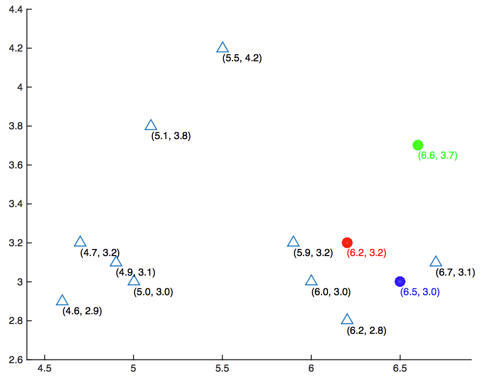
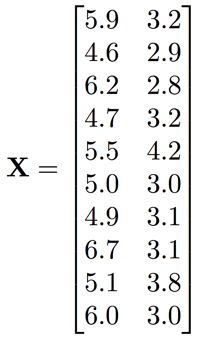


Figure 1: Scatter plot of datasets and the initialized centers of 3 clusters

Given the matrix **X** (see the matrix below) whose rows represent different data points, you are asked to perform a k-means clustering on this dataset using the Euclidean distance as the distance function. Here k is chosen as 3. The Euclidean distance d between a vector *x* and a vector *y* both in  is defined as ../../../../../Library/Containers/com.tencent.qq/Data/Library/Application%20Support/QQ/Users/617589496/QQ/Temp.db/FDF4C5E4-ABD9-433B-A314-4298A065A281.png , where *p* is the number of dimensions of each data point. All data in **X** were plotted in Figure 1. The centers of 3 clusters were initialized as μ1 = (6.2, 3.2) (red), μ2 = (6.6, 3.7) (green), μ3 = (6.5, 3.0) (blue). Answer the following questions (a) to (d).



(a). What’s the center of the first cluster (red) after one iteration? (Answer in the format of [x1, x2], round your results to three decimal places, same as problems 2 and 3)

(b). What’s the center of the second cluster (green) after two iterations?

(c). What’s the center of the third cluster (blue) when the clustering converges?

(d). How many iterations are required for the clusters to converge?

**Exercise 2: Application of K-Means**

There are 6 different datasets noted as A, B, C, D, E, and F. Each dataset is clustered using two different methods, and one of them is K-means. All results are shown in Figure 2. You are required to determine which result is more likely to be generated by K-means method. The distance measure used here is the Euclidean distance. Answer the following questions (a) to (h).

hw3.pdf

Figure 2: Clustered results for 6 datasets

(a). For dataset A, which result is more likely to be generated by K-means method? (write A1 or A2, same in the following questions (b) to (f))

(b). Dataset B (B1 or B2?)

(c). Dataset C (C1 or C2?)

(d). Dataset D (D1 or D2?)

(e). Dataset E (E1 or E2?)

(f). Dataset F (F1 or F2?)

(g). Provide the reasons/principles that draw your answers to the questions (a) to (f).

(h). For dataset F, do you think k-means perform well? Why? Are there other better clustering algorithms to be used to cluster data distributing like the data in the dataset F?

**Exercise 3: Applications of Clustering Techniques in IR and DM**

In information retrieval and data mining, are there any applications where we can apply clustering algorithms to improve the performance? Explain how clustering algorithms can improve the performance of such applications.