**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

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

**手动实施 K-Means**

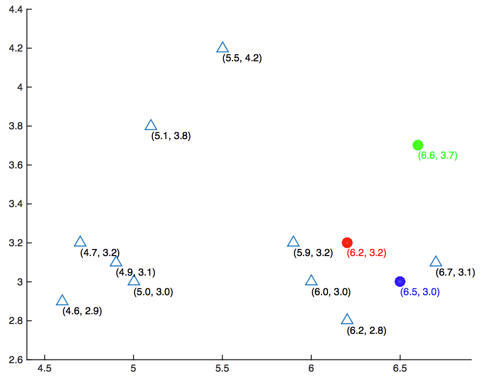
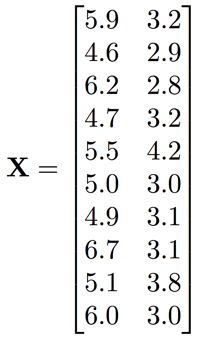


Figure 1: Scatter plot of datasets and the initialized centers of 3 clusters

图 1：数据集的散点图和 3 个集群的初始化中心

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).

给定矩阵 X（见下面的矩阵），其行代表不同的数据点，您需要使用欧几里德距离作为距离函数对此数据集执行 k 均值聚类。 这里 k 被选为 3。中的向量 x 和向量 y 之间的欧几里得距离 d 定义为 ../../../../../Library/Containers/com.tencent.qq/Data/Library/Application%20Support/QQ/Users/617589496/QQ/Temp.db/FDF4C5E4-ABD9-433B-A314-4298A065A281.png，其中 p 是每个数据点的维数。 X 中的所有数据都绘制在图 1 中。3 个簇的中心初始化为 μ1 = (6.2, 3.2)（红色），μ2 = (6.6, 3.7)（绿色），μ3 = (6.5, 3.0)（蓝色） . 回答下列问题 (a) 至 (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)

(a).一次迭代后第一个簇（红色）的中心是什么？ （以[x1, x2]的格式回答，将结果四舍五入到小数点后三位，同问题2和3）

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

(b).两次迭代后，第二个集群（绿色）的中心是什么？

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

(c).当聚类收敛时，第三个聚类（蓝色）的中心是什么？

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

(d).集群需要多少次迭代才能收敛？

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

**K-Means 的应用**

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).

有 6 个不同的数据集，分别标记为 A、B、C、D、E 和 F。每个数据集使用两种不同的方法进行聚类，其中一种是 K-means。 所有结果如图 2 所示。您需要确定哪种结果更有可能通过 K-means 方法生成。 这里使用的距离度量是欧几里得距离。 回答下列问题 (a) 至 (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))

对于数据集 A，K-means 方法更可能生成哪个结果？ （写A1或A2，下题（b）至（f）同）

(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).

提供您对问题 (a) 至 (f) 的答案的原因/原则。

(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?

对于数据集 F，您认为 k-means 表现良好吗？ 为什么？ 是否有其他更好的聚类算法可用于对数据集 F 中的数据分布的数据进行聚类？

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

**聚类技术在 IR 和 DM 中的应用**

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.

在信息检索和数据挖掘中，有没有可以应用聚类算法来提高性能的应用？ 解释聚类算法如何提高此类应用程序的性能。