Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining | Assignment 1

Due: Some Date in 2024 (100 points)

Wenbo Qian

https://github.com/wenboqian/CS-6220/tree/main/A4

K-Means

Vanilla k-Means

Question 2

Codes in A4.ipynb file.

Question 3

Codes in A4.ipynb file.

Question 4

The dataset contains 5 years of data that have stripped out the labels, so k=5 satisfy the goal of observing 5 years individually. I observed the points are grouped around the 5 cluster.

Question 5

I noticed the location of each point is closer to the center of cluster compared with Vanilla k-Means.

Question 6

[[0.99838317 -0.05684225]] Codes in A4.ipynb file.

Question 7

first principle component of the aggregate data: [[0.99838317 -0.05684225]]

first principle component of cluster 1: [[0.99993527 0.01137789]]

first principle component of cluster 2: [[0.99992533 0.01222027]]

first principle component of cluster 3: [[0.99990986 0.01342629]]

first principle component of cluster 4: [[0.99993306 0.01157047]]

first principle component of cluster 5: [[0.99989374 0.01457781]]

The first principle components of each cluster is not the same as the aggregate data, and they are not the same as each other.