Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining | Assignment 4

Due: February 15, 2024 (100 points)

Yichen Sun
https://github.com/LAnselet/cs6220-datamining

Parameter Estimation

1. derive the maximum likelihood estimate of the parameter λ.

$$L(\theta) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} \qquad (likelihood\ function)$$

$$LL(\theta) = \sum_{i=1}^{n} -\lambda \log e + X_i \log \lambda - \log(X_i!) \qquad (log - likelihood\ function)$$

$$= -n\lambda + \log \lambda \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log{(X_i!)} \qquad (use \log with\ base\ e)$$

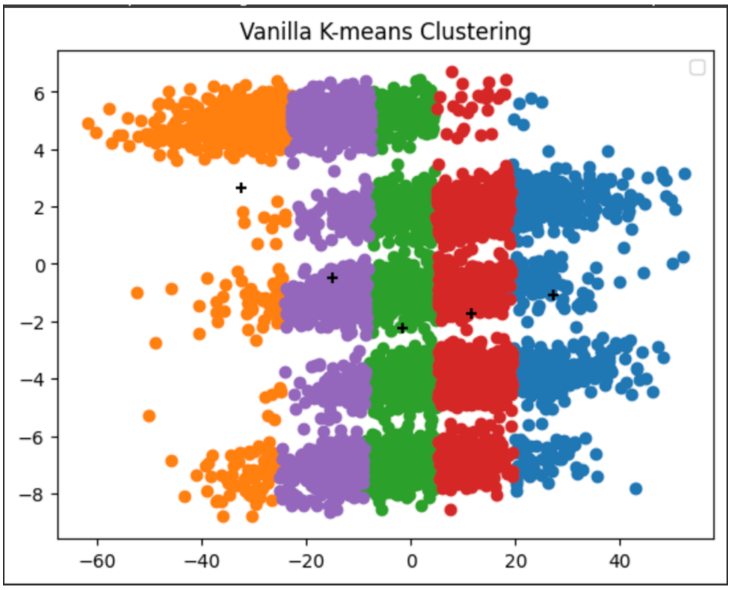
Then take the derivative with respect to our parameter λ and set it equal to 0.

$$\frac{\partial LL(\theta)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0$$

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} X_i$$

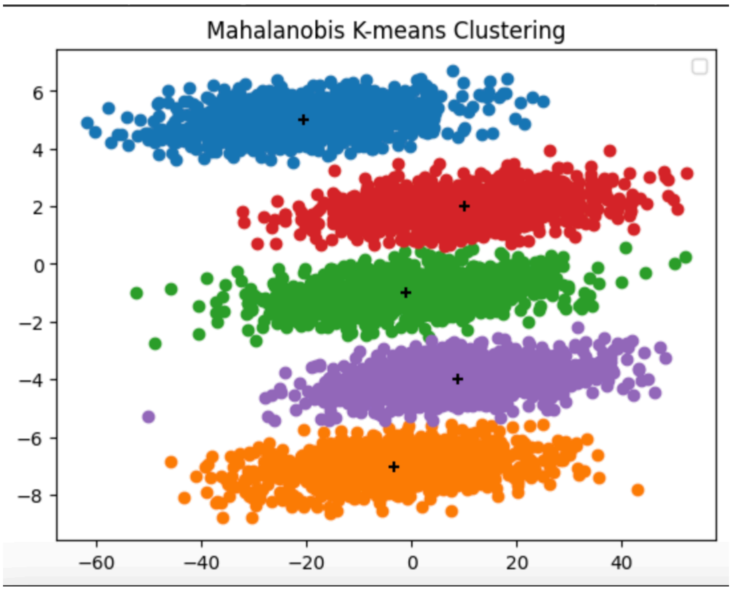
K-Means

Vanilla k-Means

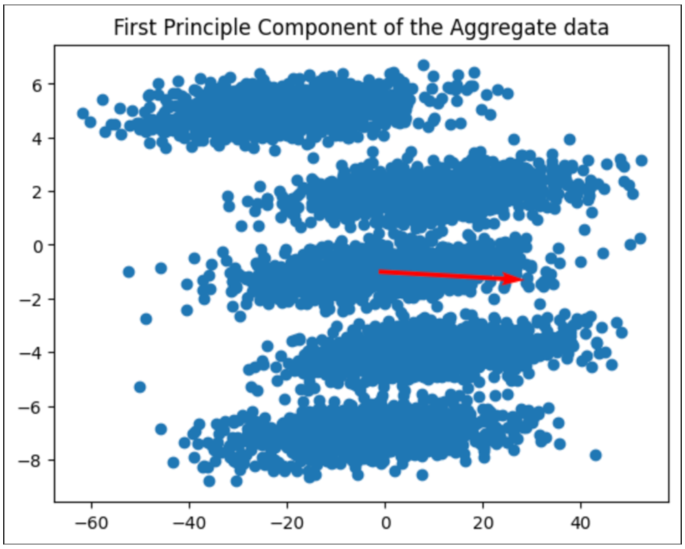
3. 

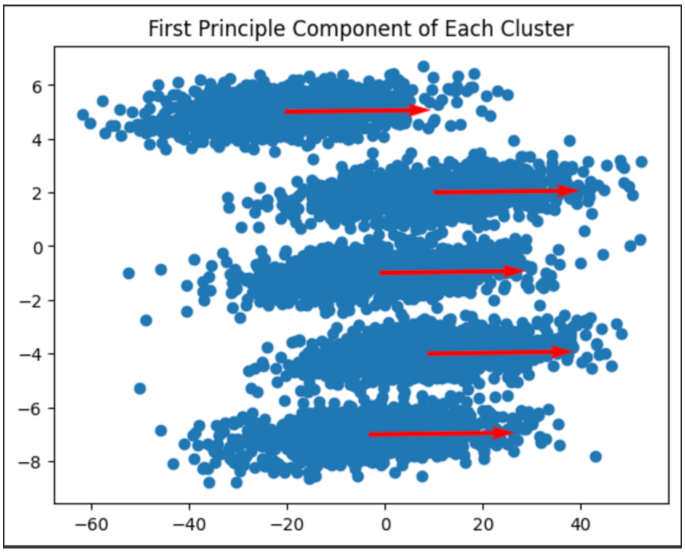
4. After plotting the resulting clusters, I noticed that the dataset naturally partitions into five groups, with each cluster showing clear separation from other clusters. Also, the centroids could represent the profile within each cluster.

With Production Information



5. I do notice that cluster each cluster is obviously separated with multiple colors and the centroids are clearly plotted in the center of each cluster.

6. 

7. 

They are not the same as the aggregate data. But PCA of each cluster are similar.