Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 4

Due: 02/15/2024(100 points)

Dharun Suryaa Nagarajan
https://github.com/dharun4772/CS6220
nagarajan.dh@northeastern.edu

Parameter Estimation

It is well-known that light bulbs commonly go out according to a Poisson distribution, and are independent regardless of whether or not they're made in the same factory. An architect has outfitted a building with 32,000 of the same light bulb. Assuming the Poisson distribution has the form:

$$p(X|\lambda) = \frac{\exp^{-\lambda} * \lambda^{x_i}}{x_i!}$$

1. Derive Likelihood function:

$$L(\lambda) = \prod_{i=1}^{N} P(X|\lambda)$$
$$L(\lambda) = \prod_{i=1}^{N} \frac{\exp^{-\lambda} * \lambda^{x_i}}{x_i!}$$

Log Likelihood function:

$$\log(L(\lambda)) = \log(\prod_{i=1}^{N} \frac{\exp^{-\lambda} * \lambda^{x_i}}{x_i!})$$
$$\log(L(\lambda)) = \sum_{i=1}^{N} \log(\frac{\exp^{-\lambda} * \lambda^{x_i}}{x_i!})$$
$$\log(L(\lambda)) = \sum_{i=1}^{N} (\log(\exp^{-\lambda}) + \log(\lambda^{x_i}) - \log(x_i!))$$
$$\log(L(\lambda)) = \sum_{i=1}^{N} (-\lambda + x_i \log(\lambda) - \log(x_i!))$$

Derivative of Log Likelihood function and equating to 0:

$$0 = \sum_{i=1}^{N} (-1 + \frac{x_i}{\lambda})$$
$$0 = -N + \sum_{i=1}^{N} (\frac{x_i}{\lambda})$$
$$N = \frac{\sum_{i=1}^{N} (x_i)}{\lambda}$$

MLE of the parameter λ :

$$\lambda = \frac{\sum_{i=1}^{N} (x_i)}{N}$$

MLE for λ when N is 32000 :

$$\lambda = \frac{\sum_{i=1}^{N} (x_i)}{32000}$$

K-Means

Vanilla K-Means

In this part of the homework, we'll take a look at how we can identify patterns in this data despite not having the labels. We'll start with the simplest approach, the k-Means unsupervised clustering algorithm.

2. Implement a simple k-means algorithm in Python on Colab with the following initialization:
   Under normal K means we use the Euclidean distance to calculate. Since euclidean distance does not work well with all shapes and sizes the resulting output looks like this.

Figure 0.1: Euclidean Distance

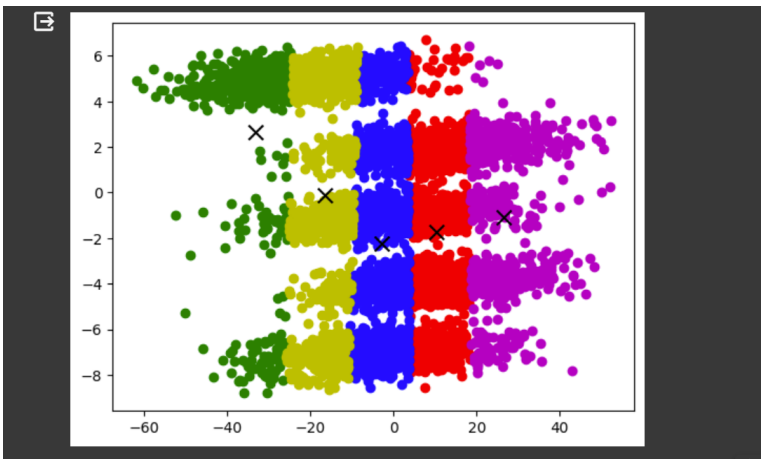

Figure 0.2: Cluster formation

3. You will notice that in the above, there are only five initialization clusters. Why is k = 5 a logical choice for this dataset? After plotting your resulting clusters and. What do you notice?
   The logical choice of k depends on the structure of the data. In this case, since we have information about the years of data (1996, 1999, 2006, 2015, and 2022), and each year may represent a distinct pattern or cluster, choosing k=5 is logical.
   We see that the clusters are not stratified properly because of their metric range variability and Euclidean distance gives more importance to larger distance than the shorter once.

This makes the cluster shape formation not in the required shape.

With Production Information

4. Implement a specialized k-means with the Mahalanobis Distance. Scatter the results with the different clusters as different colors. What do you notice?
   Since we are adding the production information matrix to the distance calculation, the smaller distance is scaled by 10 times and the larger one is made 0.25 times. This changes the importance and eventually results in five correct set of clusters.

```
[4]  def mahalanobis_distance(x,y, p_inv):
         return (x-y).T @ p_inv @ (x - y)
```

Figure 0.3: The New modified distance

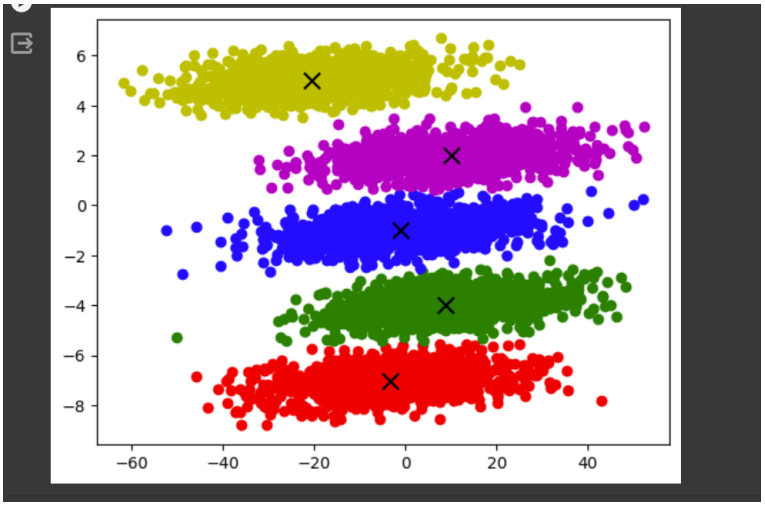

Figure 0.4: Cluster formation Part 2

5. Calculate and print out the first principle component of the aggregate data.
   The first component is shown below

6. Calculate and print out the first principle components of each cluster. Are they the same as the aggregate data? Are they the same as each other?
   They are not the same as aggregated data but they are same as each as shown below in the diagram. When calculated for the entire dataset, it captures the overall structure of the data, regardless of how it's partitioned into clusters. However, when calculated separately for each cluster, it captures the structure of that particular cluster.

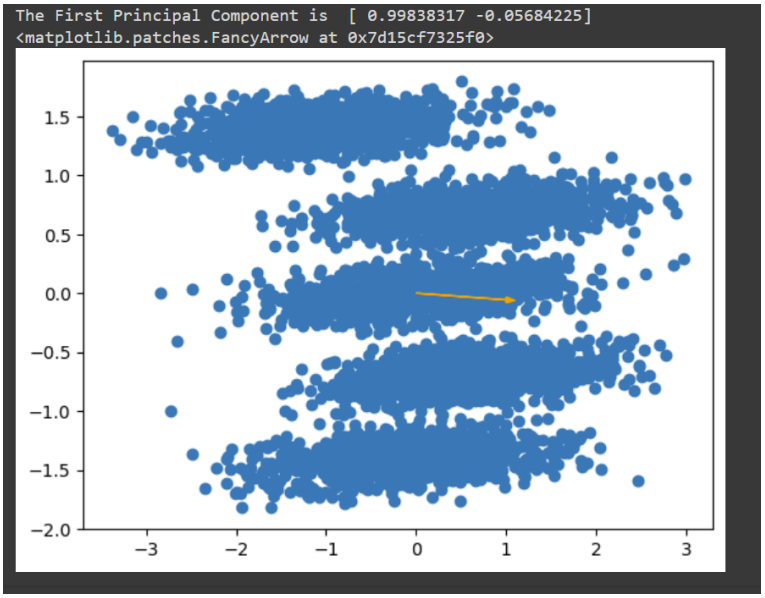

Figure 0.5: Aggregated First Component

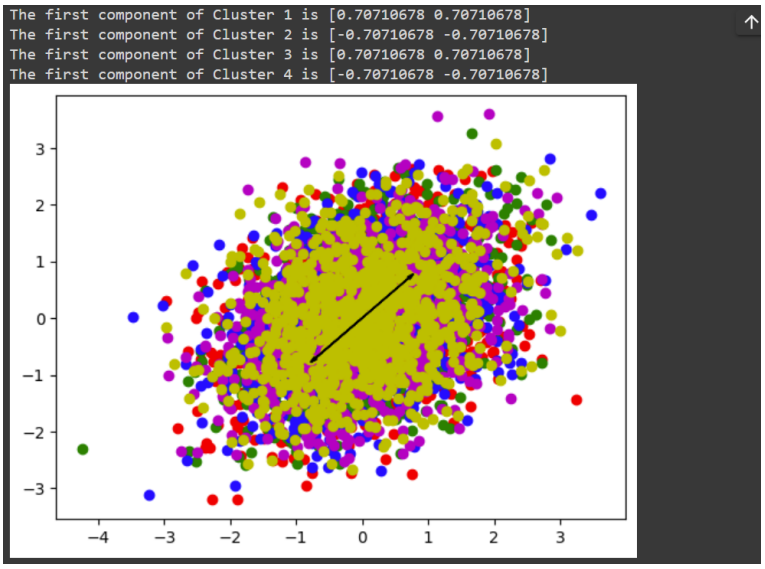

Figure 0.6: Each Cluster First Component