Northeastern University, Khoury College of Computer Science

A red and white logo

Description automatically generated

CS 6220 Data Mining | Assignment 4

Due: February 15, 2023(100 points)

Chang Liu

<https://github.com/Chang-Liu-Harry/6220DataMining>

1.

Since the observations of Poisson distribution are independent, the likelihood function for a data set of n observations is to be multiplying the probability of X equals to x1 to xn:

A black text on a white background

Description automatically generated

After we have the likelihood function, we want to know at what lambda value, we are most likely to get this n observations. We take the log of likelihood function and like it’s derivative to be zero to get the result:

A black and white math symbol

Description automatically generated

A math equations and numbers

Description automatically generated with medium confidence

A math symbols with numbers

Description automatically generated

And the result:

A black symbols on a white background

Description automatically generated

2.

Colab link: <https://colab.research.google.com/drive/1jb1ImqPOaUxSYq1jJ39Cfh-S8AJKmwmH?usp=sharing>

When we don’t use P, it’s just a vanilla k-means algorithm, because the default p: dot(P,P’) = 1.

3. A computer code with text

Description automatically generated

A diagram of a data point

Description automatically generated with medium confidence

4.

Picking 5 clusters is reasonable since our data production F-150 trucks over 5 years.

And as can be seen in the graph, the data is scattered into 5 main parts, but our clustering is not getting the 5 groups but 5 stripes, which is not exactly what we want.

5.

I implemented a k-means clustering function, and defined the Mahalanobis distance function as well as updating centroids for readability of the code. Setting the p to default, which is not using p, the now we are getting the correct clusters:  
A diagram of different colors

Description automatically generated with medium confidence

6.

First principal component for all data : [ 0.99838317 -0.05684225]

7.

First principal component for cluster 0: [0.99993527 0.01137789]

First principal component for cluster 1: [0.99992533 0.01222027]

First principal component for cluster 2: [0.99990986 0.01342629]

First principal component for cluster 3: [0.99993306 0.01157047]

First principal component for cluster 4: [0.99989374 0.01457781]

No they are all different.