Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 4

Due: Feb 15, 2024(100 points)

Yu Wang
Git User Name: titojojo
Email: wang.yu25@northeastern.edu
https://github.com/Titojojo/CS6220-Data-Mining

1 Parameter Estimation

Q1: derive the maximum likelihood estimate of the parameter λ.

Solution:

Likelihood Function: Given n i.i.d observations $x_1, x_2, ..., x_n$, from a Poisson distribution, the likelihood is:

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

Log-Likelihood Function:

$$logL(\lambda) = \sum_{i=1}^{n}(-\lambda + x_i * log\lambda - logx_i!)$$

Derivative of Log-Likelihood:

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

Solve for λ

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

2 K-Means

2.1 Vanilla k-Means

Q4: 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?
A4: After visualizing the points, I can clearly see that the points can be grouped into 5 groups. So set the number of initialization clusters as 5 is a intuitive choice. But after the vanilla k-Means, the cluster is not well-generated.

2.2 With Production Information

Q5: What do you notice?
A5: The generated clusters are better than the vanilla k-means after visualization

Q7: 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?
A7: The first principle components of *each cluster* are not the same as each other and are not the same as the aggregate data.