Conrad Niewienda
Marcos Sanson

December 15, 2024

Lab 3: Point Processes

Instructions: In this lab, cluster the patient according to the features and then adjust a Hawkes process per cluster. You can use the Hawkeslib library. Discuss the cluster differences between the parameters of the Hawkess and what these differences imply.

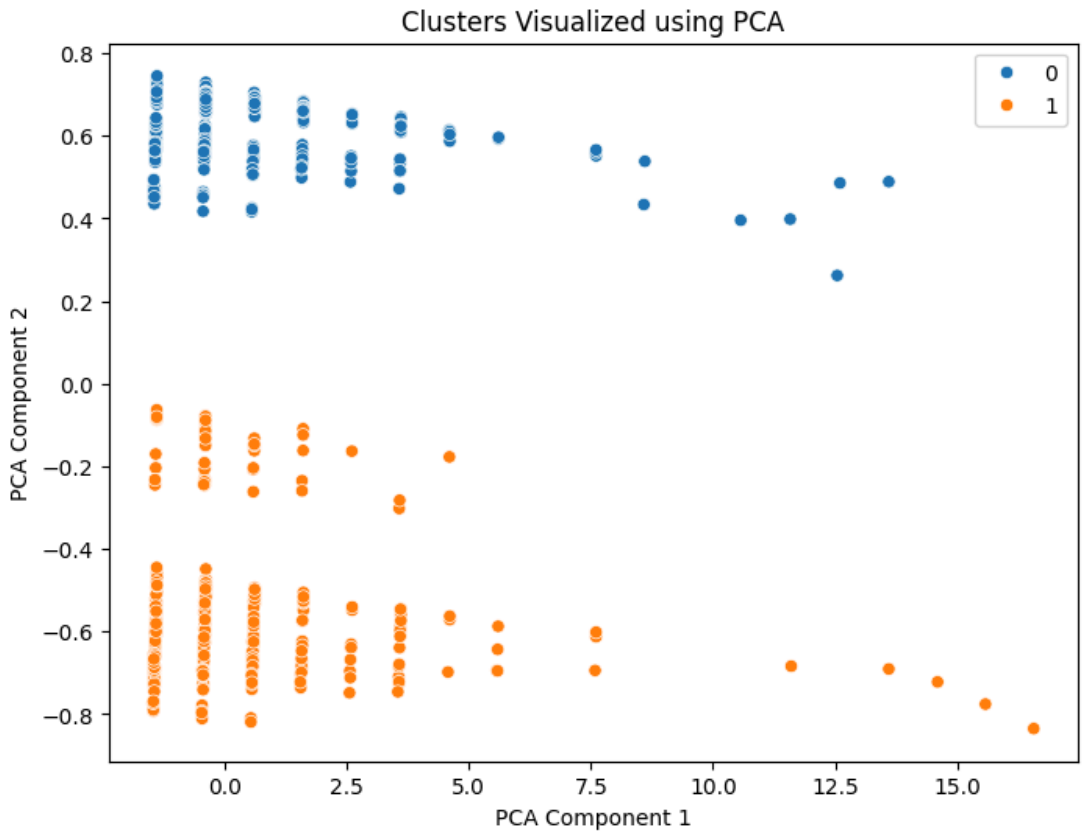
Objective

The goal of this lab was to cluster patients based on their anonymized features and fit a Hawkes process to each cluster using the provided timeline data for each patient. We began by preprocessing the data, clustering the patients into groups, and determining the parameters of the underlying Hawkes process. Finally, we interpreted the differences in these parameters between clusters to understand the event dynamics within each patient group.

Methodology

Clustering:

- A MinMaxScaler was used to scale the continuous feature to a similar scale as the binary categorical features.
- K-Medoids clustering in combination with the Gower distance was chosen since it is suitable for clustering features with different data types (categorical and numerical).
- To find the optimal amount of clusters, the silhouette score was calculated.
- The clustering was visualized using Principal Component Analysis (PCA).



Conrad Niewienda
Marcos Sanson

December 15, 2024

Fitting the Hawkes Process:

- For each member of a cluster, the absolute timeline data was calculated and added to a list.
- In order to fit the Hawkes Process, the list was sorted in ascending order and then processed using a univariate Hawkes process which is implemented in the hawkeslib library (see code for visualizations).

Key Insights

1. **Description of clusters**
    - Using the clustering method from above, two clusters were identified.
    - The clustering is relatively balanced: Cluster 0 contains 532 members and Cluster 1 contains 580 members.
    - Cluster 0 contains a total number of 425 events and Cluster 1 contains 501 events (about 15% more).
2. **Comparison of the two clusters**
    - **Parameters:**
        - **Cluster 0:** $\mu = 5.08$, $\alpha = 0.86$, $\beta = 7.33$
        - **Cluster 1:** $\mu = 1.38$, $\alpha = 0.96$, $\beta = 9.14$
        - **μ:** The baseline intensity for Cluster 0 is moderately high, meaning events happen spontaneously at a moderate rate. In comparison, the one of Cluster 1 is relatively low, meaning that Cluster 0 has a higher spontaneous event rate than Cluster 1
        - **α:** The excitation factor of both clusters is high, meaning each event significantly increases the chance of subsequent events.
        - **β:** Cluster 1 has a higher decay rate than Cluster 0; in Cluster 1, the intensity of future events drops off faster than in Cluster 0.
    - **Intensity Plots:**
        - In **Cluster 0**, the intensity decreases gradually and continues to show fluctuations for a longer time.
        - In **Cluster 1**, the intensity decays more rapidly and returns to baseline faster, with less sustained fluctuation.

Conclusion

This lab demonstrates the importance of understanding the parameters of Hawkes processes to accurately interpret and predict event dynamics in self-exciting systems.

In summary, we found that Cluster 0 represents patients with moderate spontaneous activity, while patients in Cluster 1 are characterized by infrequent spontaneous events but stronger and shorter-lived bursts of activity. These differences can have significant implications depending on the research that was done with these patients.

Link to Code and Visualizations:

Google Colab:
https://colab.research.google.com/drive/1UFzj0Q7ZQdXDehTrvaL26A7xVxbTdVHk?usp=sharing
GitHub:
https://github.com/Marcos-Sanson/UC3M-ML/blob/main/LAB03.ipynb