Chien BMI 500 Week 9 Write-Up

flchien

October 2021

1 Introduction

The intention of the Week 9 homework assignment was to use the MIMIC III database as an exploratory assignment for clinical informatics. The MIMIC III database is a large single center database with information on multiple aspects of clinical care in the intensive care unit. Using their data tables, patients with acute kidney injury was evaluated for systemic inflammation and lactic acidosis using the labs C-reactive protein (CRP) and lactate. Unsupervised clustering algorithm was then applied to the data

2 Methods

First, access to MIMIC III was obtained requiring completion of the additional CITI modules from MIT. Once access to the data was obtained, the csv file labelled labevents.csv was downloaded. In brief, labevents.csv is a table which includes the laboratory values obtained from the intensive care admission for each patient in the MIMIC III registry, including the numeric lab ID, lab value, unit, a timestamp corresponding to when it was collected, and a flag for whether the lab was identified as abnormal.

Attention was directed to another table available in MIMIC III labelled LABITEMS, which identified the clinical lab from its 5-digit numeric key. From this, the 5-digit keys for three labs of interest were obtained. Specifically, these being Creatinine (50912), C-reactive Protein (50889), and Lactate (50813). Creatinine was chosen due to it's role in defining acute kidney injury according to the Kidney Disease Improving Global Outcomes (KDIGO) guidelines. C-reactive protein is a sensitive marker for systemic inflammation, and Lactate correlates with clinical acidosis. Both elevated lactate and elevated C-reactive protein tends to signify more severe disease.

R was then used to read labevents.csv, selecting only the columns of interest. These columns include subject ID, lab ID, chart times, and lab values. To identify patients who developed acute kidney injury, rows were filtered based on the criteria of creatinine being greater than 1.5x upper limit of normal. For the purposes of the assignment, a cutoff of 1.5 was used, assuming creatinine of 1.0 being the upper limit. (In clinical practice, actual upper limit of normal will be dependent on body muscle mass and male versus female sex). After the filter was applied, the remaining rows were sorted within groupings by chart time, and only the first instance of the creatinine being greater than the cutoff was obtained. This represents when the patient first reached clinical criteria for acute kidney injury. The chart time for this event along with the patient's corresponding subject ID was saved in a table labelled firstAKI.

Two additional tables were obtained, with each table's construction including two columns: a unique patient identifier, and a corresponding lactate in one table or C-reactive protein in the other table. These two tables were constructed using a similar method to the above table detailing how acute kidney injury was identified, with the additional conditional filter that the chart time of any particular subject IDs first lactate or C-reactive protein must have been obtained prior to the onset of acute kidney injury. Thus, for each of C-reactive protein and lactate, a table was saved that includes the patients subject ID as well as their first lab value if and only if it preceded the onset of acute kidney injury.

Finally, the lactate table and c-reactive protein table was joined by subject ID. On joining, only the patients who had both a lactate and c-reactive protein drawn prior to onset of acute kidney injury were included in the analysis. NA, if arisen, were removed. The resulting data was written to an output CSV file.

To apply an unsupervised clustering algorithm to the data, the output csv was read into python. Sklearn's library was imported to use the k-means algorithm to group the data. Pyplotlib was imported and used to visualize the data in a scatter plot, with lactate on the x-axis and c-reactive protein on the y, color coded by k-means grouping. 2-groups were used in the clustering algorithm.

3 Results

Of the entire labevents table, 12825 patients were identified as having acute kidney injury based on the creatinine cutoff of 1.5. Of these patients, 4192 had lactates drawn prior to AKI onset, and 389 had C-reactive proteins drawn prior to AKI onset. 302 were identified as having both Creatinine and C-reactive proteins drawn, and these were the patients included in analysis. The 302 patients underwent K-mean grouping and the results are graphically represented below.

4 Discussion

At this point,the k-means algorithm appears to only identify C-reactive protein as important for clustering. This is likely due to the difference in units between the two labs, giving C-reactive protein higher numerical values and increased spread. Scaling both axis may correct this.

The overall appearance of the figure does not show obvious distinct populations of patients. Most patients who ended up developing acute kidney in-

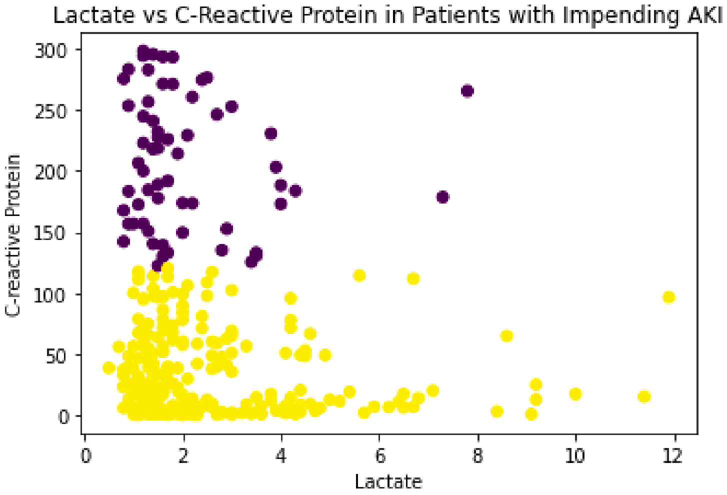

Figure 1: An image of a galaxy

jury still had normal ranges for C-reactive protein or lactase (bottom left of the graphic). Both markers being low may correlate with those who have mild grade acute kidney injury, which would be much more prevalent than high grade overt kidney failure. Patients presented on a spectrum of low to high C-reactive protein and low to high lactase, with the spread of lactase much more skewed towards lower values. Very few patients with impending acute kidney injury had both high values of lactase and high c-reactive protein. This is likely because severe systemic inflammation, acidosis, and with known impending kidney injury may not be physiologically compatible with life.