**Project Proposal: Sepsis Distribution across the United States Machine Learning Extension**

Raehash Shah

2023-11-28

**Quick Overview:**

Sepsis is a life-threatening condition that affects 1.7 million adults in the United States every year. Sepsis is typically a result of an infection, which are commonly bacterial, triggering a cascade of reactions throughout the body. These infections, which often begin in the lungs, urinary tract, skin, or the gastrointestinal tract, can cause tissue damage, organ failure or death if not properly treated. Given that so many Americans are affected by Sepsis, we would like to observe what the distribution of Sepsis is across the U.S. states in 2023.

In the previous report, two models were created as estimates for sepsis distribution. The first model included looking at bacterial infections that have been shown to lead to Sepsis and using statistics on published statistics on what portion of Sepsis are caused by which bacterial infection. The second model created estimated the distribution of Sepsis by looking at diseases that have been shown to be associated with Sepsis. However, with each of these models, there is no way of assessing how well our model performed. Therefore, we want to use machine learning techniques to train a model which we can assess. Thus, this proposal is for a project that can be performed to create a distribution of Sepsis across the U.S. states.

**Methodology:**

In the second version of the previous model, the data used was from a study called the 500 cities project which provided the number of patients diagnosed with each of the associated diseases in the 500 cities in that given year. From this an estimate on the number of patients diagnosed with Sepsis was derived using the risk of Sepsis from those diseases. For this machine learning extension, this project would use that same dataset but source more data from the CDC across multiple years. This is so we have more data points which we can use as training and testing data to generate distributions of the correlated diseases in the U.S. which would be a metric for Sepsis. In addition to the number of patients diagnosed with each disease in each year, we will use the longitude and latitude as other metrics upon which we will rely spatial clustering.

Using these data points, we will calculate the number of patients with Sepsis and have a longitude and latitude (i.e. each point is 3 dimensions). Now the way we will consider this is that our x,y dimensions is the position of the point sampled and the z dimension is the number of patients in that city. Therefore, if we were to consider a map of the U.S., we would have a bunch of points with certain peaks. Now, we can split our points into our training and testing set.

Using our training data, we can focus only on the x,y dimensions or the longitude and latitude to fit a Gaussian mixture Model to estimate how our trained points are spread across the U.S. We should be able to calculate the mean and the variance positionally so that for each start point we have an associated variance which corresponds to the rate of dissipation of Sepsis among that cluster of points. An important note is that the number of points we choose (i.e. 3 in the example provided) can be a hyperparameter tuned. Some initial ideas that are relevant to the question we are trying to ask is to create 50 distributions for each state or it could be based on more regional basis in terms of the number of counties. An important caveat to that point is that it must be less than the number of points we have in our training data or else we will end up with each point in our training data with a normal (0,1) distribution. Therefore, this is something that can be tuned later to see what leads to the best result.

|  |  |
| --- | --- |
| A graph of a diagram of a galaxy  Description automatically generated with medium confidence |  |
| An example of Gaussian Mixture Models results from some experiments I performed in class | The associated Log Likelihood for the plot after 8 iterations |

As we can see above, which was opportunity I had in class to learn more about Gaussian Mixture Models, we would get an associated log-likelihood which would be reflection of how well the clusters we formed fit the datapoints or the cities. This can be the first way we can validate our model as it will tell us if the clusters, we created are any good and if we have the right number of starting points for Sepsis (red points) and if we have good estimations for the spread of Sepsis across the U.S. An important thing that comes up here is that we must cover every town in the U.S. which we can ensure by deciding how many rings (or standard deviations away from the center) we want to use as a reflection of the spread from that point.

Now once we have our distributions created with the circles as demonstrated above, we need to translate this to what the spread of Sepsis is in each area encompassed by the variance of each cluster. Therefore, we can compute the mean and the standard deviation for all the points in the cluster and use that to estimate values. For example, the city at the center (the red circle) will be the peak so it would be estimated to have mean number of individuals with Sepsis. In the towns between the red circle and the first ring would be the mean number of individuals – 1 standard deviation value and in the second ring it would the mean – 2 standard deviations, etc. Note the number of rings that we want to consider would be another hyperparameter we would want to tune but note that we still must cover every town in the U.S. at the same time, so this would be a little tricky to work out. Another thing to consider is that it is intrinsic that we will have overlaps in our circles. Therefore, we need to consistently add the number of individuals infected in each city every time we create this estimate.

Finally, with our testing data, which will be different cities across the U.S. with known number of individuals infected with Sepsis, we can compute a mean squared error loss to see how well we performed in estimating those cities which would be a reflection on how well our model estimated the number of people infected with Sepsis in all the U.S.

**Benefits and Limitations of this Model:**

I wanted to point out some benefits and limitations of this model. The benefit of this model is that it gives context to a spread of Sepsis patients which seems the most similar to how individuals move and where they might live. In other words, not all people that are estimated to be infected with Sepsis in New York City live in NYC - they may just have been diagnosed there but they may be in the suburbs and estimation in this way allows for that. In addition, this model allows for a distribution to be implemented locally and can be tuned so that it can focus on very specific counties or even regions or it can generalize to state level, whichever provides the best results.

Some limitations are that it estimates that everything is Gaussian Normal. But this isn’t always the case, even though it makes estimation easier. The rate may be different and even among a cluster the distribution isn’t necessarily gaussian. In addition, this region relies on the sampling data to be evenly distributed across the U.S. which it often isn’t and therefore, should be considered when considering the validity of the estimation. Finally, our estimations still on Sepsis are still reliant on the assumptions made by the papers and the limitations presented in the prior version of the model.

**Discussion Questions:**

Below are some questions that I came up with when writing the proposal which if we get a chance to call, we may want to discuss before I get started. They include model considerations that might be relevant.

1. Is there a necessity to plot this on a U.S. map especially if I am considering to use longitude and latitude as features? This can get quite messy to visualize.
2. How might we choose the hyperparameters (number of start points, number of rings to use)? Some discussion provided in proposal already.
3. Do the evaluation metrics make sense given that it is a probabilistic and regression-based task?
4. Does it make sense to estimate every individual longitude and latitude relevant to the U.S. or does something like treating it like boxes of 5-10 longitude/latitude make a bit more sense? More a question of computational power and accuracy I assume.
5. A counter idea I thought of after the fact is if I just treat each point as uniform gaussian. Does that perform better results? I think I would have to test that to try.
6. I wonder how this type of model might perform compared to the Random Forest approach you mentioned to me. To me it’s interesting that used decision trees when it is a spectral problem which goes to my next question.
7. I wonder if you created a graph of cities where a node meant they were adjacent to one another on a map, how might one’s approach change then. This was a big thing I found a challenge which is how to translate cities spatially in terms of neighbors.