Predicting the Impact of Covid-19

CS 584 Semester Project Proposal
Submitted by Prakash Dhimal and William Austin
Due March 31st 2020

1. What is the problem your team is solving?

The coronavirus outbreak has become a global pandemic, with nearly a million people infected, and tens of thousands of deaths around the globe. Countries across the globe are taking action to stop the spread, but there is significant uncertainty about what the future holds. Our project aims to use principles of data mining to predict how Covid-19 will spread, how long it will take to subside, and identify what actions we can take to reduce the impact.

2. What are the challenges of the problem?

One of the main issues with our project is that the pandemic is still in progress, so we will be using partial data sets. However, we will still be able to build and test models based on data collected up to this point.

Another difference from the typical problems that we have covered in class is that we will be using time series data. Many of the techniques that apply to a standard data set are not applicable for this analysis. Rather, we will need to use other methods to extract knowledge from the data set.

Lastly, the spread of the coronavirus is an extremely complicated phenomenon and depends on many more variables than we can account for in our model. Therefore, we will be merging our main data sets with additional sources to try to find the most significant influences.

3. What data will you use?

Our main data source will be the Novel Coronavirus COVID-19 Data Repository from Johns Hopkins University's Center for Systems Science and Engineering (CSSE). This data set provides total counts of how many cases have been confirmed, broken down by country and region/province. In addition, totals for how many individuals have recovered or died from the virus are also provided. So far, there are about 70 days of data points for each region. This data set is compiled based on numbers reported by the World Health Organization (WHO) as well as numbers reported by individual nations.

As we mentioned, we may be able to provide a more meaningful interpretation of the raw counts in our main data set by including additional data sources. Therefore, we will also supplement our dataset with various publicly available datasets on platforms like Kaggle.

4. What's the state of art?

In the recents weeks, academic institutions, government, and industry researchers have put in a lot of effort into using Artificial Intelligence, Machine learning, and Data mining techniques to model the spread of the Novel Coronavirus. There is a large body of research and data around Covid-19 posted on Kaggle and other platforms to help scientists collaborate on this problem at unprecedented levels. We will be monitoring all of these advances and especially to those related to data mining techniques to help us get to our goals of predicting how Covid-19 will spread, how long it will take to subside, and identifying what actions we can take to reduce the impact at regional and national level.

5. How will you solve the problem?

Our main goal for this project will be predicting the spread of the Novel Coronavirus (COVID-19) using regression analysis. The virus is spreading at different rates at different geographical regions, therefore we will focus our regression on a specific geographical region, the United States, for example. We begin by using a typical SIR (Susceptible, Infected, Removed) model that shows how the infection rate begins slowly, increases at exponential rate and eventually flattens. One of our main calculations will be to see how R0 (reproduction number) varies over time and across regions. In addition, we will analyze responses such as mandatory social distancing orders, and closing schools to see how effective these policies are at reducing the intensity and duration of the outbreak.

6. How will you evaluate your method?

Given that we are undertaking a regression problem rather than a classification problem, we cannot use traditional metrics based on the confusion matrix generated by the test set. Rather, we will use R^2 and mean square error (MSE) scores. We will compute these values for each model that we generate.

In addition to these scores, we can also use our models to generate estimates for future counts and then evaluate how close our predictions are once these counts become available. This is a fortunate consequence of analyzing time series data for an event that is currently in progress.

References

- 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE https://github.com/CSSEGISandData/COVID-19
- Novel Corona Virus 2019 Dataset https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
- Mathematical modelling of infectious disease https://en.wikipedia.org/wiki/Mathematical_modelling_of_infectious_disease

- Sample Epidemic Calculator http://gabgoh.github.io/COVID/index.html
- Additional Coronavirus Dataset https://coronadatascraper.com/
- Coronavirus model from University of Washington, Institute for Health Metrics and Evaluation http://www.healthdata.org/covid/
- Additional Kaggle Resources for Covid-19 https://www.kaggle.com/covid-19-contributions