Data Mining Semester Project

Predicting the Impact of COVID-19

Prakash Dhimal
William Austin
CS 584 GMU
March 24th, 2020

Why Investigate COVID-19?

Problem and Motivation:

- COVID-19 is a now a major pandemic and is having a major impact on people's lives around the globe.
- There is significant uncertainty about what the effects the pandemic will be going forward.
- Large amounts of data have been collected as the spread of Covid-19 has progressed. Using data mining techniques to attempt to predict likely outcomes will support the task of building an appropriate and effective response strategy.

Primary Project Goal:

For this project, we will construct a model showing how cases of COVID-19 will spread around the globe. Our main focus will be on accurately predicting the number of infections, fatalities, and recoveries we should expect, along with what the time frame is for these events to unfold.

A Few Possible Secondary Goals:

- How much variation has there been in the severity of the outbreak between different countries and regions? What are some possible reasons for this?
- How has public sentiment about Covid-19 changed over time?
- What are the likely future impacts on the economy?
- What effects has the pandemic had on consumer behavior?
- Which policies and responses have been most and least effective?

COVID-19 Data Sources

Primary Data Source:
Our primary data source for mapping the global spread of COVID-19 will be the data provided by the Johns Hopkins Center for Systems Science and Engineering (CSSE). This contains day-by-day time series data, broken down by country and region.

- Source: https://github.com/CSSEGISandData/COVID-19
- Kaggle Page: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
- Visualization: https://coronavirus.jhu.edu/map.html

Additional Data Sources:
We plan on integrating our main data repository with other sources to improve our results. Some of these may include:

- Demographic information for each country and region. For example, population density, average age, climate, and OECD scores may be helpful.
- Patient-level records for treatment, indicating when patients became ill, symptoms, and outcome.
- Sales information for common items, showing how purchase have changed over the timespan.
- Stock market information, which roughly measures general economic confidence and has some correlation with major events.
- Social media information, like tweets, which show what the public perception of COVID-19 is over time.
- Important dates, grouped by region, showing when major milestones occurred (schools close, etc.)

High Level Project Approach

Our main activity for the project will be a regression analysis task to try to predict how the virus will progress. We will begin by modeling the data after common activation functions and can be used to show how the infection rate begins slowly, gradually increases, and eventually flattens.

It may also be possible to use information about countries at different points in the recovery process to estimate what will happen in other countries. For example, there are currently very few new daily cases in China, while the US is still experiencing exponential growth. In addition, depending on what data we are able to use, we may be able to use methods of associative rule learning to improve our results.

We may also explore other methods of analyzing time series data. Two of the most common options are:

- RNNs (Recurrent Neural Networks)
- Markov Models

Lastly, we will apply standard data mining practices to our data sets as we proceed with the project. For example, data normalization, feature reduction, and parameter tuning will all be applicable.