John Marshall

Udacity Deep Learning Nanodegree Capstone Proposal

Domain Background:

In late 2019, an infectious disease called Coronavirus (COVID) was first diagnosed in Wuhan, China prior to spreading globally before being labeled a pandemic on March 11th, 2020 by the World Health Organization (Adhanom *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*). Given the grave consequences concerning the proliferation of this pandemic, I thought it would be prudent to try and tackle some forecasting on the expected spread of this disease. While an infectious disease can naively be modeled with either an exponential growth/decay or some sort of Markov chain, these tools lack the capability of capturing more rich details. As of this instant, there have been 5,991,102 confirmed cases and 366,875 deaths spread across 188 countries/regions (Hopkins *Covid Dashboard*).

Problem Statement:

Given an initial condition (total confirmed cases and total deaths as of date X), how many new cases/ new deaths can we expect over a specified interval. If we can properly predict the evolution of this pandemic, it could allow more effective resource allocation for all the independent entities that are trying to contain this pandemic. My aim is to predict the pandemic one week out at a time. I am guessing perturbations inherent in the data (given the uncertainty about the unconfirmed cases) will cause the projections to grow increasingly unstable for each additional timestep. If I find success at the weekly interval, I will try extending the timeframe to see what type of accuracy I can achieve.

Dataset and Inputs:

I am going to use these datasets I found from Kaggle:

1. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_deaths.csv (Srk *Novel Corona Virus 2019 Dataset*)
2. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_confirmed.csv (Srk *Novel Corona Virus 2019 Dataset*)

The datasets are broken into confirmed cases and death totals, and I plan on running two separate models for the two datasets. I might try to do some research into how I can incorporate so that deaths at time t can be projected using a given context length for both deaths from $t-k$ to $t-1$ and confirmed cases from $t-j$ to $t-1$ where j can be equivalent to k, but the two do not necessarily have to be equal.

In terms of the dataset, it provides information on the province/state, the country/region, and the latitude and longitude of the particular state and country. Then it provides the raw count of the quantity of confirmed cases and deaths for the given date range. The data is layed out so that the dates are represented in columns and states/countries acting as the rows, (which I am reversing). I am

dropping the latitude and longitude because I think that information will be contained in the categorical representation of the state/country.

For the state/country, I am dropping the 'NaN's from the province/state column and subsequently concatenating that column's strings with the state and country's strings into a single column. Next, I changed the new state/country column into a column of categorical variables. After creating the categorical column, I will append an array of binary variables to the data frame, each column will represent each unique specific country and region combination. This representation will serve my purpose for running a linear regression for one of my proposed baseline models. Then, I will assign an integer value for each region and stack the data on top of itself, so it will have the dates along the rows in a repeating sequence (essentially as the index) and the first row will be the confirmed cases (or deaths), and the second row will be the integer country representation. This will serve the purpose of passing it to AWS in JSON format as a categorical variable.

In terms of the actual time series data, the data is a running total of the confirmed/death counts. Thus, I will take a one day differential to convert this into a time series of daily observations. Among the 266 unique state and country combinations, 33 of these produce negative values in their time series after taking the differential. For these, I will do some investigations into the structure of the data and will either average the two adjacent data points (if there is only a few negative values for that specific time series) or completely discarding that state/country if I have to generate too many synthetic observations.

Below is a summary of my differential data for my confirmed cases (Spain on April 24th is the culprit for the one egregioius negative value). Of the 54 negative observations, 22 of those are just a trivial difference of -1, and of the 15 countries with this minor difference, 12 countries have this as their only negative value. On the other hand, France has seven negative observations, with six of those a difference of greater than 100. They seem to be a likely candidate that I may drop.

	Confirmed_Cases
count	33782.000000
mean	171.937452
std	1353.983287
min	-10034.000000
25%	0.000000
50%	0.000000
75%	10.000000
max	36188.000000

Solution Statement:

I intend to replicate the timeseries notebook from the previous module and use SageMaker's DeepAR forecasting algorithm.

Benchmark Model:

I want to try a few different levels of sophisticated benchmark models. The most simplest is the persistence model which is the hypothesis $x_t = x_{t-1}$. For a slightly more sophisticated algorithm I will use a simple linear regression of the same form of the above, but I will also include a binary categorical variable to represent each individual region. If these models are nontrivially insufficient, then I will try a more advanced ARIMA model for panel data. I will compare these baseline models to Amazon's deepAR model to evaluate the efficacy of the relative models.

Evaluation Metrics:

I intend to use root mean squared error as my basis for comparison across the three different models. I will also include the mean absolute percentage error and the mean absolute scaled error for two alternatives if the RMSE doesn't seem to adequately capture each models' efficacy. It is entirely plausible given that countries that don't adhere to social distancing policies and/or economic periods of

shutdown witness an explosion of cases which could skew the RMSE, so hopefully these alternative evaluation metrics will mitigate some of these nonlinear effects.

Project Design:

To summarize, my intention is to condense the province/state and region labels into one categorical variable and then convert the dates into a proper format. Once I clean the data, I will investigate at multiple levels whether there are any discernible patterns. One thing I might want to include, is grabbing and normalizing the countries 2019 GDP as an instrumental variable to control for countries/regions with less equipped institutions to ensure social distancing is followed. After that I will start building my models and comparing the results not only across the models, but also across the time series timeline.

Works Cited

Adhanom, Tedros. “WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020.” *World Health Organization*, World Health Organization, 11 Mar. 2020, www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

Hopkins, Johns. “Covid Dashboard.” *ArcGIS Dashboards*, 30 May 2020, www.arcgis.com/apps/opsdashboard/index.html.

Srk, SRK. “Novel Corona Virus 2019 Dataset.” *Kaggle*, 29 May 2020, www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_confirmed.csv.

Srk, SRK. “Novel Corona Virus 2019 Dataset.” *Kaggle*, 29 May 2020, www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_deaths.csv.