COVID-19: Coronavirus Modeling with Prediction

A Project by Ronald DeLuca & Lulu Melhem

Introduction

With the global pandemic of COVID-19, a novel occasion to utilize data science methods and applications could be used to model the data provided in these uncertain stages. Much is currently unknown about COVID-19, including the aspects of severity, transmission, and actual penetration into society. This uncertainty causes fear and angst amongst a populace that seeks reassurance. Thus, data science can be used to model input data, such as confirmed cases, active, recovered, death, and more with reliable models that can be evaluated to show that prediction and forecasting of these numbers can be made with high accuracy.

The process of undertaking this project was based on important sections, such as pulling latest input data, pre-processing, manipulation, dealing with noisy data, separating the data for training and testing, inputting data into models, tuning the models and their parameters, graphing and visualizing the results, and comparing the evaluation metrics of each model to find one with the highest level of prediction accuracy. Each section of this project was optimized to ensure scalability and accuracy throughout. In each stage, the sections were optimized to take into account multiple reliable data science procedures to best handle the eventual desired output results. Each of these sections will be discussed in further details throughout this paper.

The continuous growth of the John Hopkins CSSE data set on COVID-19 has been taken into account in our implementation. With this growth in mind, much planning was performed to ensure that each stage held up its responsibility from the beginning of the project's undertaking until completion. A total of 13 prediction models including SVM, Bayesian Ridges, and Linear and Polynomial regression, have been implemented, visualized, and evaluated.

Furthermore, evaluation metrics were important to understand which model would provide the "best" results for forecasting features of the dataset. Many metrics were employed to evaluate the accuracy of each model and to conclude which model performed more accurately given current input data. As the input dataset grew, the models that performed well early in the COVID-19 outbreak, continued to perform well over time and showed more accurate predictions in the short to long term forecasting range compared to other models.

Improvements Since the Last Presentation

Since our last presentation we have finished our implementation of the SIR model and fine-tuned all other models. Using estimated parameters, we have visualized the SIR model in the United States and globally. In addition to the classic SIR model implementation, we have added a social distancing parameter, to visualize the effect such a precaution has on the model's predictions. More details on this will be discussed in the implementation section below.

Problem Statement

The world is now more than 4 months into the COVID-19 outbreak and we, as people, are still wary of what will happen. As governments across the world contemplate on whether or not to reopen businesses and return to normalcy, prediction modeling such as linear and nonlinear models are used to forecast and visualize the impact and potential of the virus, COVID-19. COVID-19 is a respiratory disease spreading from person to person caused by a novel coronavirus [1]. This virus can infect anyone, and individuals, such as the elderly and those with underlying health issues are at higher risk for severe illness [2]. Infections and deaths are increasing and without a cure or preventative measures, this virus could have a catastrophic impact on the world. In the meantime, data scientists use current data and data science models to predict the effect this virus will have on our health.

Our project attempts to address this problem. Using prediction models such as XGBoost, Facebook's Prophet, and the SIR model, we attempt to predict how infectious and fatal this virus will be around the world. Our principle data set comes from the John Hopkins CSSE data set [3]. Using pre-processing tools, we have consolidated daily reports into one data set. Additionally, we have used a Kaggle data set [4] where the population for each country around the world can be found. Due to how novel the virus is, it is challenging to find many reliable, complete, and accurate data sets at this time.

Approach

To understand the growth of COVID-19, data science techniques could be employed to attempt to model features such as confirmed cases, active, recoveries, deaths, etc. Beginning with our input data, the Johns Hopkins CSSE COVID-19 dataset was used to provide accurate case counts with features of locations, emergence date, and growth each day given in their Daily Report CSVs located on GitHub. Over the course of the project, the number of input datasets grew but continued to evolve around the daily reports that appeared on GitHub. An end-to-end overview of the process for each model is seen in Figure 1.

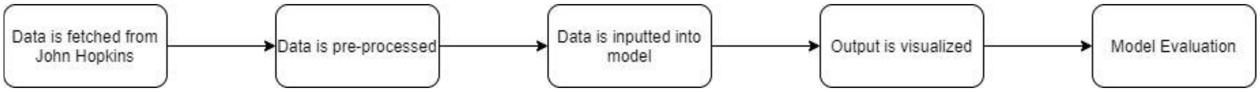
Data is fetched from John Hopkins
Data is pre-processed
Data is inputted into model
Output is visualized
Model Evaluation

Figure 1. Application process

Our Jupyter Notebook begins by extracting the data from all the John Hopkins CSSE CSVs locally stored in a folder into a Pandas DataFrame and outputting a time-ordered CSV that combines all of the data into one useful set. The features of the early CSVs, those produced before March, were limited in attributes compared to recent Daily Reports. To account for this difference, a complex pre-processing algorithm was custom made to maintain consistency amongst all input CSVs and to scan the similar locations with exact longitudes and latitudes and fill in missing data to attempt to make the DataFrames more complete.

The complexity of the data pre-processing step was accounted for, as the memory arrangement and scanning the DataFrame for matches while filling in missing values could be cumbersome on simpler machines. Therefore, the algorithm was optimized to avoid excess time and memory usage by ensuring that DataFrames are not built on top of each other. This custom process felt more appropriate given the dynamic nature of the input data and data analysis performed, however, it could be optimized even further depending on the hardware used.

Once the input data was pre-processed, filled in, and manipulated into the desired DataFrame and intermittent output CSV, the DataFrame was split into testing and training sets. Several different set sizes were tested to verify results against our evaluation metrics, but a ⅔ training and ⅓ testing split seemed most appropriate for most of the models used. With the data being time-series format, the split was based on how many days since January 22nd, 2020; the date of the first reported case in The Johns Hopkins COVID-19 dataset. Additional arrays and datasets, each of different timespans, lengths, etc., were prepared for features such as date, location, confirmed cases, deaths, and. All DataFrames created served a specific purpose at different parts of the project.

The DataFrames were prepared, knowing that our different models would each require somewhat different inputs depending on their parameters and specific model requirements. Some of our models attempted to use regression modeling in formats that would attempt to fit training data to a line with a simple linear format or in some cases, more broadly using polynomial features. Many different models were considered on the regression format that would fit to a line, such as linear regression, polynomial, Bayesian Ridge, Lasso, ElasticNet and others. Additional models were considered that use regressor algorithms which employ other tactics, such as how SVMs would use more intricate parameters when tuning the input data for understanding features in a higher dimensional plane. The other considered models were SVMs, XGBoost, ARIMA, Random Forests, KNN, Facebook's Prophet and many others that seemed to suit a time-series prediction well. The final chosen models that are given in our Jupyter Notebook represent those which all perform decent at time-series forecasting beyond the latest given date in the dataset while maintaining decent evaluation metrics and reasonable results that were 'close' to the current data.

Each model was tuned with specific parameters and compared the testing data set against a multitude of metrics that were visualized in multiple tables, for each model and then against all other models. Beyond the evaluation metrics of each model, they all output around a ten-day range forecast of features such as confirmed cases ten days from the last date given date. The entire timespan of data was output in individual graphs and compared against a prediction graph. Further, numerical predictions were given in an output table that was offered for the model itself and then again compared against all models. This approach per model was helpful to understand the individual performance of each model when looking to understand more information about that particular model.

Lastly, we consider the SIR model, an epidemiological model that computes the number of individuals infected in a closed population over time [5]. Different variations of the model exist but we have implemented a classic SIR model where we examine and predict the number of susceptible, infected, and removed (recovered + deceased) individuals in a population. The SIR model has two parameters, the rate of infection, β, and the rate of recovery, γ. The challenge with this model is that these two epidemiological metrics are is still unknown [6]. Estimates of these parameters, $\beta = 1.75$ and $\gamma = 0.5$ are calculated with information provided by two recent studies that were conducted on the COVID-19 [7] [8].

Three differential equations, seen figure 1, describe this model. Using the scipython library functions, each of these equations is integrated by the parameters β and γ on the population of interest. In addition, in order to model the effect social distancing can have on the spread of the virus, we introduce a new constant term, ρ, between 0 and 1 where 0 indicates everyone is locked down. We set ρ to 0.5 and 0.8 and evaluate the effect it has on the spread of the disease.

Once we had multiple reasonable models, we combined results of evaluation and visualization to compare all models against each other to determine which one was performing the "best". The model comparison section serves to gather results about the whole project in one glance, such that one could understand the accuracy of all models employed, the resulting predictions by the graph of all models, and the actual forecasted numbers of a specific date in the short-term range.

Evaluation

To provide reliable results, the output predictions of our models needed to be evaluated and compared as to show which model provides the "best" results for forecasting. In undertaking linear and other regression model types, standard metric calculations and packages were used from providers such as Scikit-learn. Each model was evaluated with Mean Absolute, Mean Squared, Mean Squared Log, Root Mean Squared, Median Absolute, Mean Absolute Percentage Error, R2 Score, Mean Poisson, Mean Gamma, Mean Tweedie Deviance and more. The multitude of metrics gave opportunities for individual models to shine while others seemed to have large errors in all categories. Besides giving these individual metrics for each model, a final collected table gave the metrics of all models of importance used in the project with the opportunity to visually understand the strengths of a particular model.

Analysis

With the computed predictions and evaluation metrics of each model, the results tend to speak for themselves. Models that had more finely tuned parameters and took into account more features of importance in understanding the specific potentially exponential nature of COVID-19 cases served to better predict eventual findings. More advanced models, such as Facebook's Prophet can predict short and longer-term cases more accurately. SVMs and Bayesian Ridge Regressors could be tuned with parameters that account for COVID-19's actual nature and how it was not maintaining strictly linear growth on some features. The polynomial nature can better predict features from our training and testing splits, which showed up well in minimizing error for the evaluation metrics. Finally, as expected, the SIR model was a weak predictor. Because so much is unknown about the virus, the parameters are hard to estimate at the moment and an accurate SIR modeling will be seen after the outbreak ends.

Conclusion

As the project progressed, much was learned about the applicability of data science techniques to novel approaches with the understanding of COVID-19's progression. The initial data gathering proved to be difficult in some cases as data may have been sparse, unreliable, or have not been thoroughly shared due to privacy concerns or other issues. The format of data and reporting has changed several times since the beginning of this project, causing the need to iterate our process of data pre-processing while planning for the potential for future change. Additional details and information were learned about specific models, and the benefits over other models were further understood as performance metrics were evaluated. The nature of COVID-19 presents opportunities to more thoroughly understand the need for modeling and how best to apply certain models to continue to provide reliable results that can be trusted or used for helpful situations. Overcoming the troubles presented by COVID-19 can best be performed by applying similar methods that apply to data science, finding reliable data, processing it to understand useful knowledge, and evaluating that knowledge to know the truth hidden behind the data.

References:

[1] "Situation Summary," *Centers for Disease Control and Prevention*, 19-Apr-2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/summary.html#emergence.

[2] "Frequently Asked Questions," *Centers for Disease Control and Prevention*, 22-Apr-2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/faq.html.

[3] "CSSEGISandData/COVID-19," *GitHub*, 28-Apr-2020. [Online]. Available: https://github.com/CSSEGISandData/COVID-19.

[4] T. N. Prabhu, "Population by Country - 2020," *Kaggle*, 18-Apr-2020. [Online]. Available: https://www.kaggle.com/tanuprabhu/population-by-country-2020.

[5] "The SIR Model for Spread of Disease - The Differential Equation Model," *The SIR Model for Spread of Disease - The Differential Equation Model | Mathematical Association of America*. [Online]. Available: https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model.

[6] G. Yeghikyan, “Modelling the coronavirus epidemic spreading in a city with Python,” *Medium*, 04-Feb-2020. [Online]. Available: https://towardsdatascience.com/modelling-the-coronavirus-epidemic-spreading-in-a-city-with-python-babd14d82fa2.

[7] J. Hellewell et. al, “Feasibility of controlling 2019-nCoV outbreaks by isolation of cases and contacts,” *The Lancet Global Health*, vol. 8, no. 4, Feb. 2020. Available: https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(20)30074-7/fulltext

[8] L. Peng, W. Yang, D. Zhang, C. Zhuge, and L. Hong, “Epidemic analysis of COVID-19 in China by dynamical modeling,” Feb. 2020. Available: https://arxiv.org/pdf/2002.06563.pdf