Forecasting American COVID-19 Cases and Deaths through Machine Learning

Anaiy Somalwar[1]
[1]**BASIS Independent Silicon Valley, San Jose, CA**
anaiysomalwar@gmail.com

Abstract

While the most commonly used models for COVID-19 include epidemiological models and Gaussian curve-fitting models, recent literature has indicated that these models could be improved by incorporating machine learning. However, within this research on potential machine learning models for COVID-19 forecasting, there has been a large emphasis on providing an array of different types of machine learning models rather than optimizing a single one. The objective of this research is to examine whether a recursive, primarily autoregressive machine learning model could significantly improve or compare with current, state-of-the-art COVID-19 modelling and whether such machine learning models could be the optimal solution for COVID-19 forecasting.

To build our models, we chose to use the national data for the United States from the New York Times' GitHub under the covid-19-data repository as it was updated daily and contained no missing values. To predict the future total United States COVID-19 cases or deaths, we assumed that there existed functions f and g that respectively mapped the COVID-19 cases and deaths of some past constant i days to the COVID-19 cases and deaths of the next day, or the $i + 1th$ day. We experimented with the number of previous days inputs to use for the prediction of future cases and deaths, respectively represented using variables n and k. We split the first 80 percent of the chronologically sorted data to serve as training data where the model learns the function that mapped the inputs to the outputs, and we used the remaining 20 percent of my data to test how well the model would perform for data it had never seen. The models were trained using gradient descent and the mean squared error loss function, and we found that a simple linear regression model with two cross-validation sets and l2 penalization was performing the best after we set $n = 7$ and $k = 5$.

To predict the next l future cases or deaths from current time t, we recursively used our models' output of the cases and deaths for time $t + 1$ to serve as input for predicting the cases or deaths for time $t + 2$ and so forth. We chose l as 7 while making predictions as that was a number which was used in previous literature on this topic and made this model easily comparable. The Ridge Regression recursive model for predicting future United States COVID-19 total deaths performed with an average RMSE of 570.264 based on the average RMSE of the seven-day prediction task, and the Ridge Regression recursive model for predicting future United States total COVID-19 cases performed with an average RMSE of 4282.341 based on the average RMSE of its seven-day prediction task. The r^2 value for the recursive deaths predictor was 0.996, and the r^2 value for the recursive cases predictor was 0.992. By comparing these results to current state-of-the-art models, we conclude that a hybrid of a recursive machine learning model for shorter range predictions and a Gaussian curve-fitting model or an epidemiological model for longer range predictions could greatly improve the accuracy of COVID-19 forecasting.