Coronavirus by County

Sean Roelofs

1 Project Description

On March 11, 2020, The World Health Organization delcared COVID-19 a global pandemic. In the US, as of June 1st, there are over 1.8 million cases and 107 thousand deaths [1]. As the US enters the fourth month of pandemic, decisions to restart the economyy and other parts of society are largely being driven by state governments and individual citizens. Therefore, in order to help inform these local governments, businesses, and citizens, I plan to study how fast coronavirus outbreaks on a local level. Specifically, I will look on a county by county level.

2 Dataset

In this project I use two datasets. The first is the US Census Demographic Data from Kaggle [2]. This dataset contains the racial and economic profile by county as estimated by the 2015 American Community Survey. It includes information that may be relevant to predicting the spread of coronavirus including poverty rates, unemplyment rates, types of transportation used, job types, racial demographics, population size, and more. I use this dataset to create the X labels for each county. There are 79 pieces of information per county, and I normalize these inputs in a preprocessing stage.

The second dataset contains the number of coronavirus cases per county by day [3]. This dataset is compiled by the New York Times. It includes the number of reported cases and deaths each day for every county. I use this dataset to create the Y labels for each county as explained below.

The goal of this project is to use the census dataset to predict the spread of the coronavirus. Inorder to create a proxy for how fast the virus spreads, I look at how long it takes a county to go from 5 to 50 cases. I chose to start at 5 because this date is more likely to mean there is actually transmission within the county, as opposed to a lower number which may indicate a single individual entering the county. I chose to stop at 50 because there are only 1350 out of 3007 counties with 50 or more cases as of May 19. If I chose a larger number, I would have less data to work with.

I meshed these two datasets together be using the County FIPS code, a unique ID assigned to each county in the US.

I used a 60/20/20 train/val/test data split. This means there are 810 counties in my training dataset and 270 counties in each of my val and test datasets.

3 Data Statistics

As of May 19, there are 1350 counties reporting 50 or more coronavirus cases. The average time for a county to go from 5 to 50 cases is 17 days, with a standard deviation of 9.57. The minimum was 2 and the maximum was 46 days.

In order to evaluate my performance on different models I will use the average error (L_1 loss) between my models prediction of 5 to 50 cases and the actual numbers. The average error of my datasets is as follows:

	Train	Val	Test
Average Error	10.48	10.41	10.48

All of my models should be able to surpass this average error.

4 Baselines

4.1 Linear Regression

As an initial baseline, I implemented a simple linear model. The linear model takes the form:

$$\hat{y}^{(i)} = Wx^{(i)} + b \tag{1}$$

To train the model, I minimized mean squared error using stochastic gradient descent with weight decay. After searching over the weight decay hyperparameter space, I chose $\lambda = 0.05$ as it had the best validation set performance.

$$\mathcal{L}(\hat{y}, y) = \frac{1}{m}\Sigma_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2 \tag{2}$$

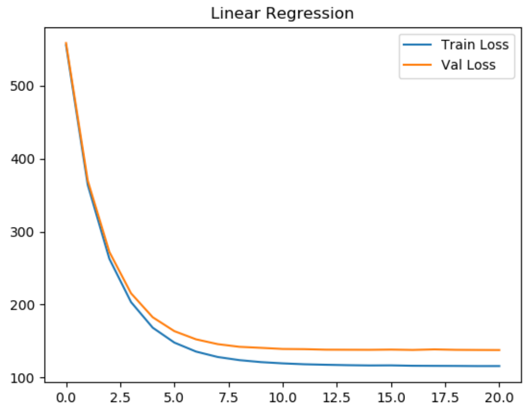


Figure 1: Training and Validation Loss per Epoch

	Train	Val	Test
Loss	115	137	140
Average Error	8.5	9.3	9.4

4.2 Support Vector Regression

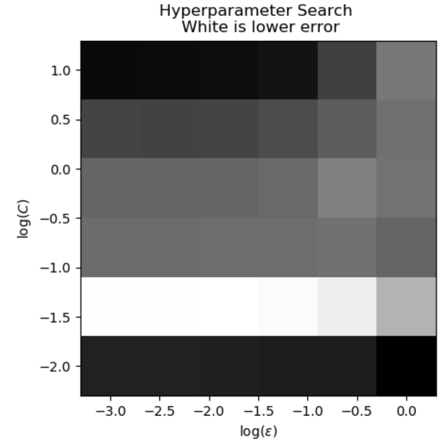
As a secondary baseline, I also tried Support Vector Regression. Support Vector Regression is the continuous version of SVM. It attempts to minimize the size of the learned parameters while insuring that most predictions are within a margin of the actual values.

$$\textbf{minimize}\frac{1}{2}||\theta||^2 + C\Sigma_{i=1}^{n}|\xi^{(i)}| \tag{3}$$

subject to

$$|y^{(i)} - \theta x^{(i)}| < \epsilon + |\xi^{(i)}| \tag{4}$$

I used Scikit-Learn's implementation of SVR. I also used its default radial basis kernel as it produced the lowest error. After searching over the hyperparameter space, I chose that $\epsilon = 0.016$ and $C = 0.04$.



	Train	Val	Test
Average Error	8.0	9.0	9.2

We see that the final test error is lower for SVR than for the Linear model. I hypothesize that most of this improvement is due to a closer match between the SVR's objective and my use of the L1 norm to evaluate the model's performance.

The best hyperparameter chosen was $\epsilon = 0.016$, which is much smaller than the average training error of 8.0. This means that most of the datapoints make use of their slack constraints, which means the objective of SVR becomes very close to a Linear model with L1 loss. Since I am eavluating my model on the L1 norm, it makes sense that this objective would return beter results. A relatively smaller percent of the increased performance is due to the radial basis kernel, as the error with this kernel is only slightly higher (on the order of $e - 2$).

Finally, we also see that the test average error of 9.2 is lower than 10.5, the average error of the test dataset itself. This shows that the SVR model does have some predictive power in this task.

5 Neural Network Model

5.1 Previous Literature

There is an emerging body of work that uses neural networks to model the dynamics of infectious diseases. In 2018, Chae, Kwon, and Lee published a paper titled ”Predicting Infectious Disease Using Deep Learning and Big Data” [5]. In this paper, they attempted to model the spread of Chicken Pox, Scarlet Fever, and Malaria using both a DNN and an LSTM. They found that the Deep Neural network outperformed the LSTM in predicting new cases per day.

In 2019, Akhtar, Kraemer, & Gardner published a paper titled ”A Dynamic Neural Network Model for Predicting Risk of Zika in Real Time”. In this paper, they attempt to identify countries at risk for the spread of the Zika by using a variety of data sources, including socioeconomic and human population data. The socioeconomic and human population data they use is similar to the census data this project uses. The authors claim that ”hhe proposed prediction problem is highly nonlinear and complex”, which motivates the use of neural networks for this task.

The data provided by the US Census is rich with detail for each county. There are fields that tell what percent of people commute to work, what types of jobs exist in the county, what the poverty levels are, and what the education levels are. These characteristics of a county likely interact in convoluted ways when assessing how a virus can spread through the county. This motivated me to try to model my data with a Neural Network.

5.2 Network Details

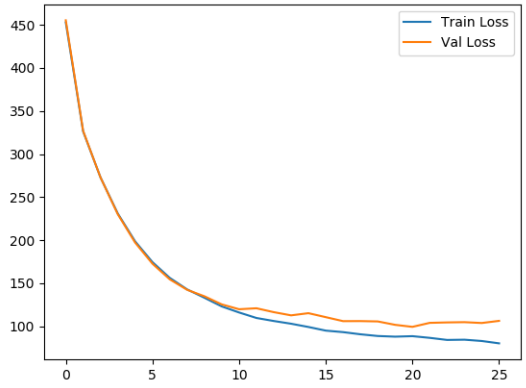
After experimenting with many different depths, layer types, and activation functions I settled upon the following architecture:

FullyConnected(78→128), BatchNorm, ReLU, Dropout(p = 0.3)
FullyConnected(128→128), BatchNorm, ReLU, Dropout(p = 0.3)
FullyConnected(128→32), BatchNorm, Tanh
FullyConnected(32→1)

I found that the second to last activation function being Tanh yields significant improvements in the results. I hypothesize that this is because I am attempting to model a continuous output. Having the Tanh activation function as the last nonlinearity in my network before the prediction allows my prediction to predict the last prediction.

I tried using a network architecture using soley sigmoid activation functions as proposed in [4]. However, this network architecture did not yield as good results. Both swapping in ReLUs for the initial layers and swapping in a Tanh for the final layer yielded better results. I hypothesize that the ReLUs allowed for better training, which increases model performance over the naive sigmoid implementation.

	Train	Val	Test
Loss	81	111	110
Average Error	6.9	7.9	7.8



These results are an substantial improvement over my baselines. On the test set, the neural network is, on average, 1.4 days closer to the actual five to fifty number. The baseline SVR is itself 1.2 days closer to the actual five to fifty number than the average difference in the dataset. Since improving the average error is harder the closer it is to 0, this shows that the neural network model is a significant improvement of the baselin SVR.

Given that there is a lot of noise in the five to fifty number because of inadequate and unequal testing throughout the country, it is highly likely that bayes minimum error for this task is not close to zero. It is hard to tell where this line is, however I predict that the inate difficulty in this problem is one reason my neural networks average error was not closer to zero.

5.3 Insights from Network Weights

Inorder to better understand how my neural network model is working, I looked at the weights in the first layer. All the input features of my data map to human understandable demographic data, so by looking at what data is weighted the most, I should be able to understand what data my model finds most valuable. I ranked the input features by the average of their squared contribution to each node in the first hidden layer. The five most important features, in order, are:

Total Population
Percentage White
Percentage Asian
Percentage over Eighteen
Percentage Working in an Office

It certainly makes sense how total population of a county would affect the five to fifty spread. It is likely harder to social distance in a more populated county, and a more populated county will have more people to potentially be infected. It also makes sense that age and job types would impact the spread of coronavirus. Young people are often asymptomatic and people who work in an office (as opposed to work at home, construction, etc...) are more likely to be in close contact with eachother. It is less obvious what impact race has on the spread of coronavirus, but perhaps race is just a latent feature for other salient properties of a county.

6 Sources

1. Coronavirus Disease 2019. Centers for Disease Control and Prevention. cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us

2. US Census Demographic Data. US Census Bureau. kaggle.com/muonneutrino/us-census-demographic-data

3. Coronavirus (Covid-19) Data in the United States. The New York Times. github.com/nytimes/covid-19-data

4. Chae, Snangwon eta al. ”Predicting Infectious Disease Using Deep Learning and Big Data.” International journal of enviromental research and public health. 27 Jul, 2018.

5. Akhtar, Kraemer, & Gardner. ”A Dynamic Neural Network Model for Predicting Risk of Zika in Real Time.” BMC Med. 02 Spet, 2019.