CS 230 Milestone 2

Sean Roelofs – sroelofs@stanford.edu – 006205512

May 27, 2020

1 Project Description

On March 11, 2020, The World Health Organization delcared COVID-19 a global pandemic. In the US, as of may 18, there are over 1.3 million cases and 79 thousand deaths [1]. As the US enters the third 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 will 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 my X labels for each county. There are 79 pieces of information per county, and I normalize these inputs in a preprocessing stage.

The second is data compiled by the New York Times on coronavirus cases by county level [3]. This dataset includes the number of reported cases and deaths each day for every county. These two datasets can be meshed by using the County FIPS code, a unique ID assigned to each county in the US. I use this dataset to create my Y labels.

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 1156 out of 3007 counties with 50 or more cases. If I chose a larger number, I would have less data to work with.

3 Data Statistics

As of May 9, there are 1156 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.

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.

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

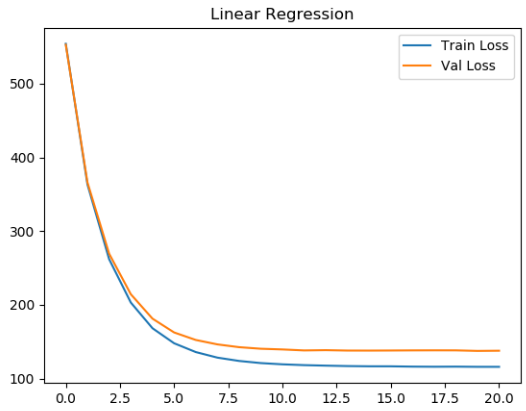


Figure 1: Training and Validation Loss per Epoch

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

4.2 Support Vector Regression

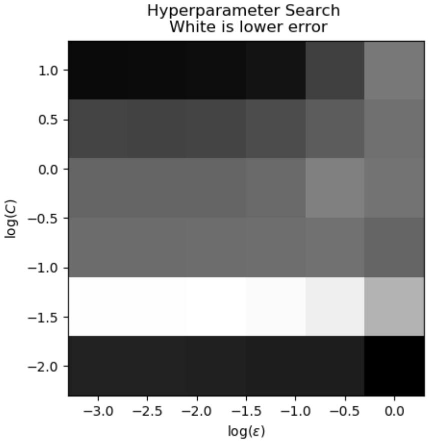
As a secondary baseline, I also tried Support Vector Regression. Support Vector Regression is the continuous version of SVM. It attempts to minimize the size of the learned parameters while insuring that most predictions are within a margin of the actual values.

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

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 $e - 2$).

5 Neural Network Model

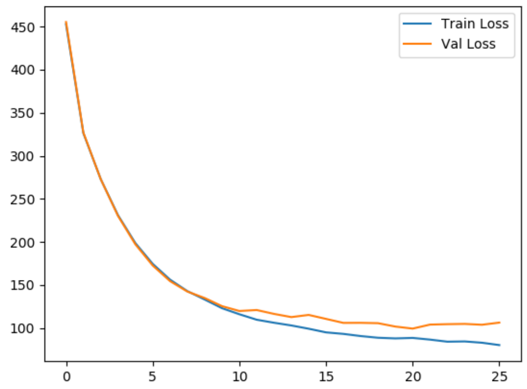
The data provided by the US Census is rich with detail for each county. There are fields that tell what percent of people commute to work, what types of jobs exist in the county, what the poverty levels are, and what the education levels are. These characteristics of a county likely interact in convoluted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

5.1 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(0.3)
FullyConnected(128→128)
BatchNorm()
ReLU
Dropout(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 and that Tanh activation function is the last nonlinearity in my network before the prediction.



	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, 1.5 days closer to the actual five to fifty number. 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. Although it is hard to tell where this line is, this neural network model is certaintly eeked out extra performance over the baseline.

6 Next Steps

For my next steps, I want to try to use my neural network to predict the results of counties that are soon to hit fifty cases. These counties may follow a different distribution then the counties in my current dataset because they are later in time. It will be interesting to see if my models can have predictive power in the real world, and potentially impacts on peoples and governments decisions.

7 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