

Aech Loar, Erik Tarango, Josh Cordova, Kobe Lin

Math 189Z

5/15/20

Research Question:

It is easy to simplify risk of COVID-19 infection into mere population density, but to do so ignores the effect of a number of other factors. As such, we decided to look into how a neural network trained on other countries classified the risk level of US counties. Obviously, all areas are at risk for coronavirus, but some are less likely to end up widespread than others, which can be identified through other health statistics of the region. Identifying which populations are most at risk for COVID-19 is a popular task at the moment--the dataset that we used had 26 such submissions, including one looking specifically at the county level as we did. However, this submission was looking primarily at the risk of a second wave based on antibody development, and thus is answering a very different question than our project. Johns Hopkins updates a map daily showing the fatality rate of the coronavirus, which shows a similar trend to our results, though it is looking at pre-existing data, rather than being a predictive model.

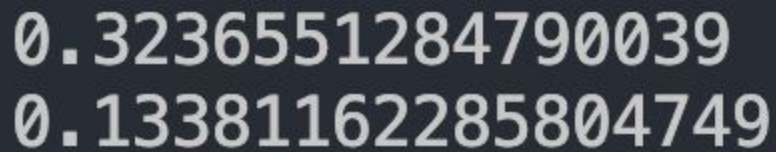
Methods:

Using two data sources (US Public Health Rankings and recent covid-19 states), we created a dataset that contained important features of a county's well being (obesity, smoking, age distribution, social association rate) as well as confirmed coronavirus cases and death toll. Using the coronavirus data, we calculated the median death rate for counties that have been significantly impacted already (confirmed cases > 100). We then labeled each county either "at

risk” (death rate > median) or not. We encoded that column into binary vectors, and converted it into tensors for our network to use as the training labels.

Now for the independent variables. We had 63 columns in our dataframe, and some of them were just plain useless. For example, the column ‘percent_insufficient_sleep_white_ci_95_high’ was not likely to be useful as a predictor variable. We gathered the 23 features that we thought would be most relevant to the fatality rate of the virus, such as those discussed above. After converting them into tensors, we thought we were pretty much done, as all we had to do was train the network. This was not exactly the case, as the first couple of training attempts resulted in almost zero change to the MSE. Our matplotlib loss curves were stubbornly horizontal. Our neural network was acting like a class of rowdy teens refusing to listen to the well meaning but inexperienced substitute teacher. We had to find a way to reach these kids. That meant iterating through different learning rates, numbers of layers, and sizes of those layers to get this to work. We originally made a goal of minimizing the MSE and had found plenty of models to do that, but when we tested it on our validation set we were getting really bad results. After a certain level of detail we reached a point where any more complexity would result in overtraining. For every .01 shaved off of our training mse, we’d be adding it right to our validation mse.

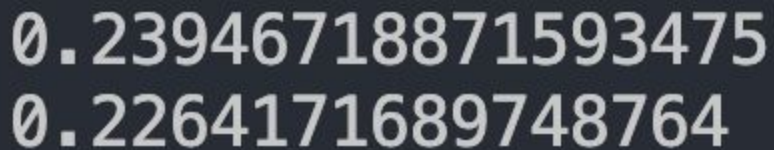
Here’s the MSE for the first epoch of our validation set (top) compared to that of the last epoch of our training set. Quite good for our training data, quite bad for our validation set. (Actually worse than the MSE for an untrained network.)



0.3236551284790039
0.13381162285804749

It turns out that getting these numbers to be similar is very hard.

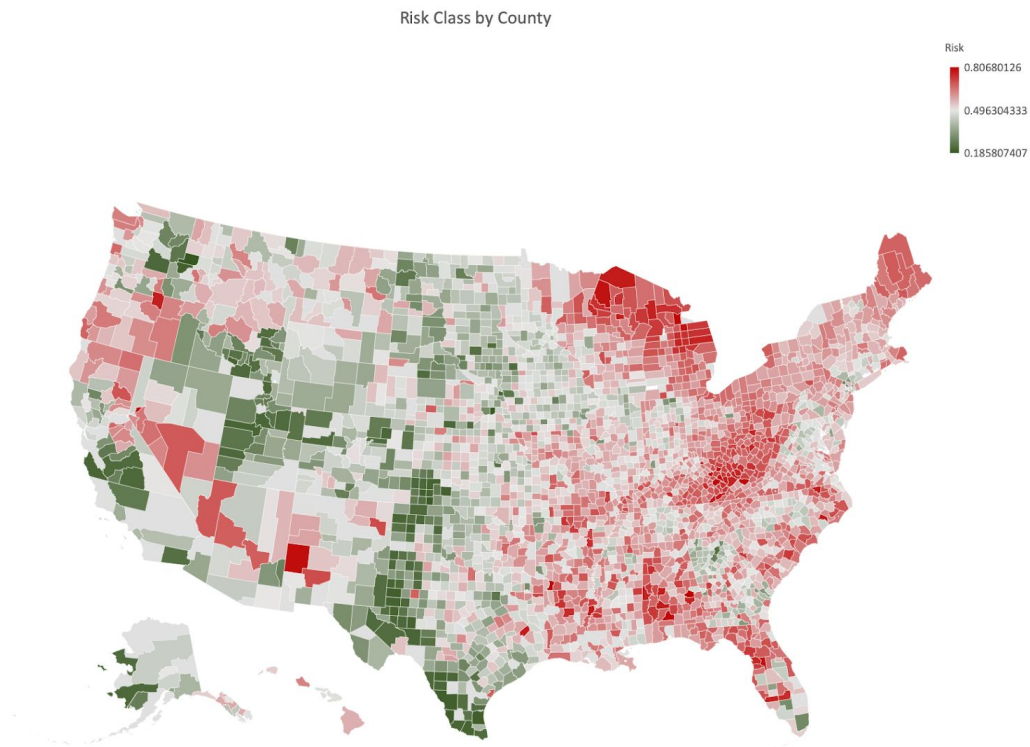
After modifying the features yet again (cutting them down to 18), adding layers, and changing the optimizer to use Stochastic Gradient Descent since Adam wasn't doing enough regularization, we got this:



0.23946718871593475
0.2264171689748764

This seems ideal. Once we made sure that the model was solid, we ran it on every county in the US (except for like 4 cause the data was missing).

Here's what we found:



To clarify: This map is not supposed to suggest that some counties are safer than others from the virus. Whether you live in Juneau or Detroit, you have to do your part to slow the spread, as the consequences will be large even for the greenest counties. This is projecting the large scale fatality rates should Covid-19 spread throughout each of these counties. This is not purely a safety metric. However, if you were to catch the virus, you might wanna take a look at if you're in a red county.

Discussion:

In the map above, we can see that many counties on the East Coast are at high risk of fatality. This information is valuable to have because it indicates the level of need of resources for different counties so that these supplies can be distributed appropriately. While it is still important that all counties are given resources, those with higher risk factors should be considered more because of the fear of overwhelming the medical capacity of that area. Moreover, through the MPE per Epoch curve we can see that we have relatively reliable results.

We believe we gained more valuable information by looking into risks of counties rather than states. This gives information for resources to be more specifically distributed and though all counties need to be especially careful, can emplace greater measures or orders in order to prevent the high spread of COVID-19. These steps are obviously essential since there is not a vaccine for the virus available yet.