California State University, Fullerton

Computational Statistics and Data Analysis

MATH 502B

Using Mobility Data to Predict the Spread of COVID-19

Author:

Cameron Abrams

Professor:

Dr. Sam Behseta

Summer 2020

July 30, 2020

Abstract

This project uses freely available mobility data and COVID-19 case information to predict how many new reported cases of COVID-19 there will be in a U.S. county on a given day. I applied ridge regression, lasso, and neural network models for COVID-19 case prediction. Additionally I used the data to predict which county the data originates from. This was implemented using LDA, QDA, multinomial regression and neural network models. It was found that we can predict the trend of reported COVID-19 cases. Furthermore, we can predict which county the mobility and COVID-19 data comes from with high accuracy.

Introduction

There have been over 4.4 million confirmed cases of COVID-19 in the US. Public health officials notified companies such as Google and Apple that their anonymized mobility data could be useful in helping to make important decisions about combating the spread of the virus and reopening the economy [4].

Previous work has been done to predict the number of confirmed cases, suspected cases, recovered cases, and deaths resulting from COVID-19 in Wuhan, China. This was done by using an SEIR model as well as training neural networks on COVID-19 case information and migration data[9]. In this project I looked at multiple U.S. counties and attempted to learn from each region's mobility data and reported COVID-19 cases in order to build a model which can predict the number of new cases in *any* county. I also wanted to see if we could use the mobility data and COVID-19 case information to predict which region the data came from.

In total, I collected data for nine counties across the US: Orange, CA, Alameda, CA, Los Angeles, CA, New York, NY, Jefferson, AL, Duval, FL, Miami-Dade, FL, Milwaukee, WI, and Cook, IL.
The number of reported COVID-19 cases in Orange County was collected from the Orange County health department website[1]. The COVID-19 case data for Alameda and Los Angeles was collected from the California health department website[5]. The reported cases for New York county was collected from a github repository hosted by the New York health department[6]. The reported cases of COVID-19 for the remaining counties were collected from USAFacts.org[8].
Mobility data was collected directly from Google's mobility data website and Apple's mobility website [4, 3]. The Google mobility data represents "movement trends by region, across different categories of places" [4]. The data is a percentage change in visits (or time spent in) certain categories of locations from a baseline. That baseline is the median value for each category from the five week period between January 3rd and February 6th, 2020. The categories are parks, retail, grocery stores, transit stations, workplace and residential. The Apple mobility data represents "relative volume of directions requests per country/region, sub-region or city compared to a baseline volume on January 13th, 2020" [3].

Data Processing

Missing Data

For each county, Apple mobility data from May 11^{th} and 12^{th} is missing for unknown reasons. The missing values for each county were replaced with the mean of the apple mobility values which were present for that county.

Forming A Data Set To Predict Cases Today

The time range from exposure to development of symptoms of COVID-19 is 3 to 14 days [7]. The mean incubation period (time until symptoms develop) from COVID-19 is 5.1 days [2].Recent research suggests that people are most contagious in the 48 hours before symptoms develop and they remain contagious for up to 10 days after onset of symptoms[7].This means the number of cases reported today are related to past interactions of infected with non infected in some previous, but recent, time period.

I had to find a way to consider some previous number of COVID-19 cases and use that as a predictor for the number of cases on a given day. A naive approach was taken. We have n data points for n different days which tell us how many new cases were reported on each of the n days:

$$X_{cases,1}, \ldots, X_{cases,n}.$$

A new variable $Y_{cases,i}$ can be created, where

$$Y_{cases,i} = \sum_{j=1}^{7} X_{cases,i-j-2}.$$

This is the sum of cases for the 7 day period ending 2 days ago. The goal is to capture the number of people who were contagious last week. Two additional days are added since once someone is infected it takes, *usually*, at least two days to develop symptoms[7]. In the example below the X_{cases} variable represents the original data and each value is the number new cases reported for that day. We then create a new variable Y_{cases}. In this case, the Y_{cases} variable for 4/10 will be the sum of the new cases from 4/1 to 4/7. The goal is to then use that information to predict the number of new cases on 4/10, which is 6.

$X_{cases} =$

4/1	4/2	4/3	4/4	4/5	4/6	4/7	4/8	4/9	4/10
1	2	3	4	5	2	1	3	4	6

$Y_{cases} =$

4/10	4/11	4/12	4/13
18	20	22	28

Then we have n data points which tell us the mobility patterns for 7 different categories on n different days. Since I do not know which days people were infected, who was infected, and where they travelled before they developed symptoms I again used a naive approach. I created a parameter from the average of the mobility values from the 7 day period ending two days ago. In other words, I took

$$X_{mobility,1}, \ldots, X_{mobility,n},$$

and created a new variable $Y_{mobility,i}$, where

$$Y_{mobility,i} = \frac{1}{7} \sum_{j=1}^{7} X_{moblity,i-j-2}.$$

I did this for all 6 of the the Google mobility parameters as well as the Apple mobility parameter. An example is given below and follows the same logic as the example above except we are now averaging the data instead of simply summing.

$X_{mobility} =$

4/1	4/2	4/3	4/4	4/5	4/6	4/7	4/8	4/9	4/10
1	2	3	4	5	2	1	3	4	6

$Y_{mobility} =$

4/10	4/11	4/12	4/13
2.57	2.86	3.14	3.57

The final set of data had 9 predictors. One predictor was the sum of new cases over a one week period ending 2 days prior to the date we are predicting for. Seven predictors were the averages of each of the mobility categories over the one week period ending 2 days before the date we are predicting for. And the ninth predictor was the county from which the other 8 predictors were averaged and summed from.

Principal Component Analysis

I performed principal component analysis on the reconstructed data set. The first component accounted for 65.1% of the variability in the data set. In the figure below we can see there is a strong relationship between the first component and all of the mobility predictors. County and the previous cases are also related to this first component but to a lessor extent, which still makes them significant since this single component accounts for so much of the variability in the data set. The second component accounts for 11.2% of the variability in the data set and it is most strongly related to the previous weeks total cases. The third component which accounted for 10.2% of the variability was most strongly related to the county parameter. It appears all of the constructed parameters are strongly related to the total variability of the data set.

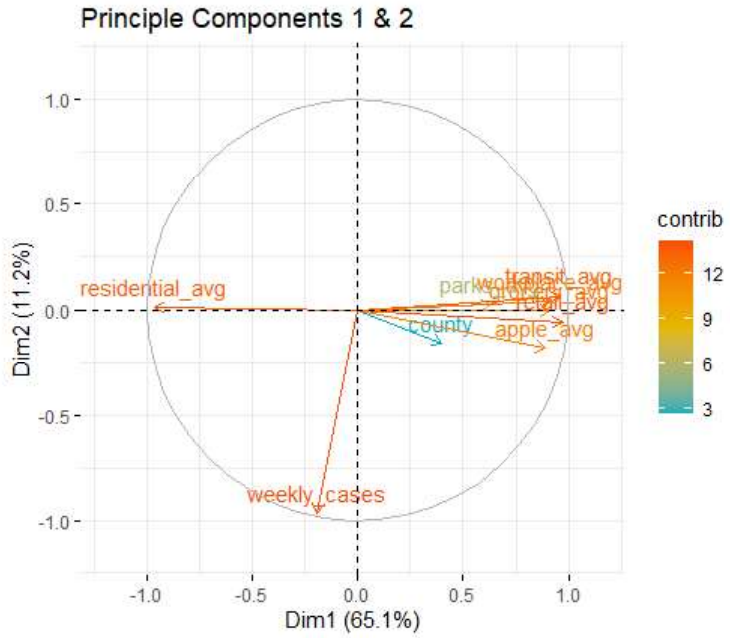

Figure 1: The first and second principal components visualized using a bi-plot.

Methodology

Regression Methodology

Ridge regression, lasso, and neural network models were trained on the data including the county parameter. I also trained the models on the data with the county parameter removed and then used those models to predict cases for three new counties for time periods ranging from February 2020 to July 2020. Cross validation was performed on the ridge regression and lasso models to determine the optimal value of λ. Fifteen neural network models were built and 5-fold cross validation was performed on the data to determine which would be the best to use to predict cases in the new counties. The neural network which performed best was a 4 layer fully-connected network with the three hidden layers having 8,8, and 5 nodes respectively. The hidden layers applied the activation functions ELU, ReLU, and ELU respectively. The final output had ReLU applied and the mean squared error served as the loss function. Of the 15 neural networks trained, the 9 best performing were trained on the full data set. Afterwards, their predictions for the three counties were averaged and used as a set of predictions.

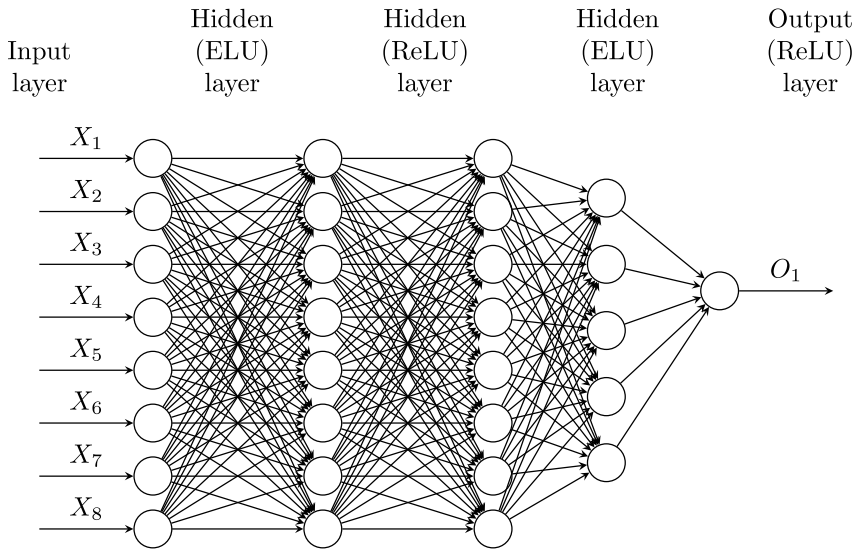

Figure 2: A diagram of the neural network that was found to be optimal for predicting the number of new cases out of the 15 models which were trained.

Classification Methodology

LDA, QDA, multinomial regression, and neural network models were used to classify the data points to their respective counties. Five-fold cross validation was performed on 10 neural network models to select the final model. The final model used 3 layers, 1 hidden, each having 9 neurons since there were 9 predictors and 9 counties. The activation functions for the layers were linear, ReLU and sigmoid functions, respectively. The sparse categorical cross entropy loss function from the keras library was applied to train the model.

Results

Regression Results

When the data including the county parameter was broken up into a training and validation set ridge regression performed the best on the validation set with a validation MSE of 44825.67. This was followed by the single neural network with an MSE of 48381.90, then the ensemble of neural networks with an MSE of 49386.63, and finally lasso had the lowest performance with an MSE of 51743.21.

Models	**Validation MSE**
Ridge	44825.67
Lasso	51743.21
NN (Best)	48381.90
NN (Ensemble)	49386.63

Table 1: For the ensemble of neural networks the median MSE is provided.

The models were trained on the entire data set with the county parameter removed and were then used to predict new cases over a contiguous time period. The counties were San Diego County (CA), Bexar County (TX), and Maricopa County (AZ). No single model performed particularly well on *all* of the counties. Instead, three different models performed best on different counties. The ridge regression model performed the best on Maricopa county, AZ (which includes the city of Phoenix) from 2/25/2020 to 7/21/2020. The single best neural network most accurately predicted the number of new cases for San Diego, CA from 3/28/2020 to

7/12/2020. The ensemble of the best 9 neural net models performed the best at predicting new cases for Bexar county, TX (which includes the city of San Antonio) from 2/25/2020 to 7/10/2020.

Regression Models	**SD MSE**	**TX MSE**	**AZ MSE**
Ridge	7444.48	18047.07	160346.20
Lasso	5676.24	16492.61	162077.80
NN	5218.07	16133.21	163157.08
NN (Ensemble)	5363.91	16084.09	162319.20

Table 2: NN was 4 layer full-connected network with a neuron shape of [8,8,8,5] with respective activation functions [ELU, ReLU, ELU, ReLU] trained for 50 epochs. The ensemble was a total of 9 of the best performing neural network models and the predictions are their mean predictions.

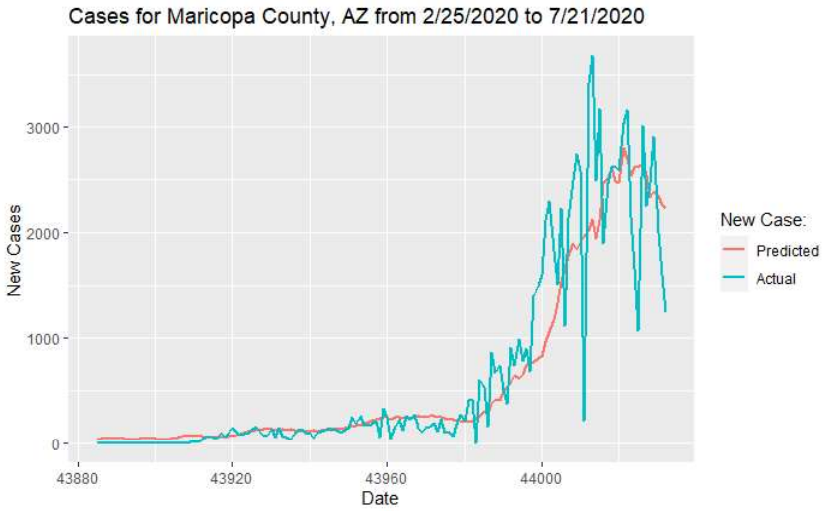

Figure 3: Ridge Regression Predictions for new COVID-19 Cases in Maricopa County, AZ

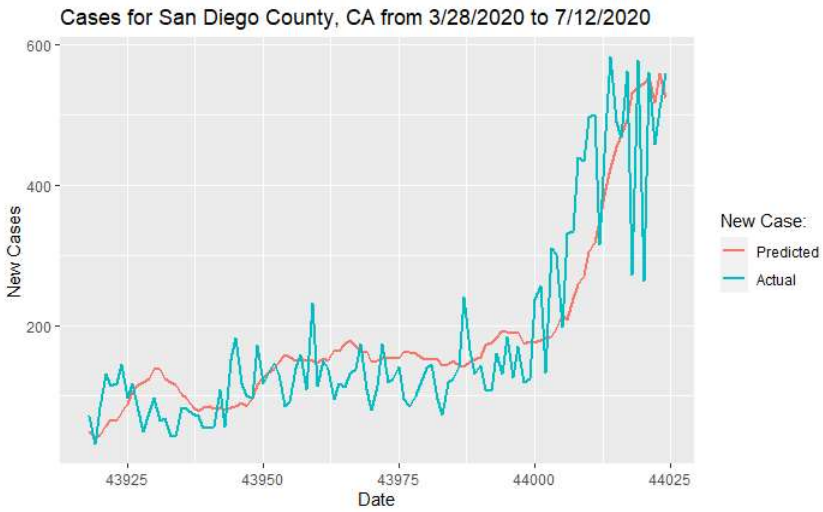

Figure 4: Neural Network Predictions for new COVID-19 Cases in San Diego County, CA

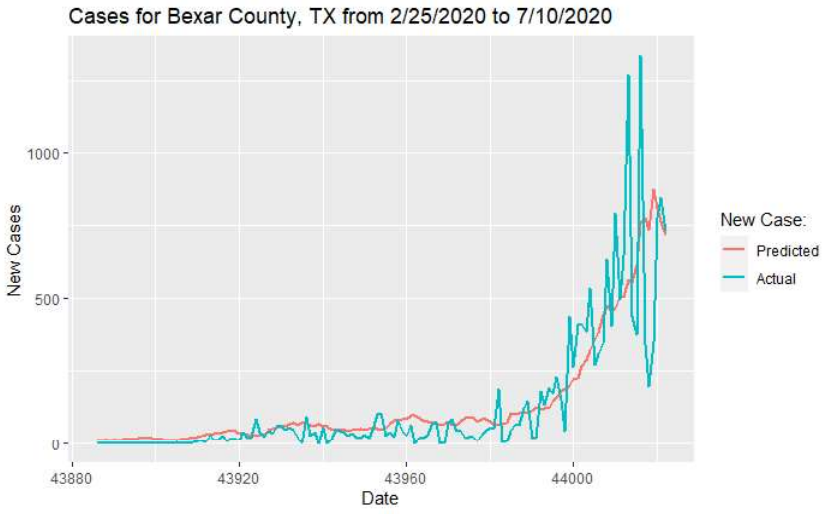

Figure 5: Ensemble Neural Network Predictions for new COVID-19 Cases in Bexar County, TX

Classification Results

20% of the data was held out and used as a validation set for the classification models. The neural network performed the best with a validation accuracy of 95.3%, followed closely by QDA (94.5%), multinomial regression (92.9%), and lastly LDA (86.3%) which performed the worst, but still admirably. What's nice about this is, given the right data, simple techniques such as QDA and multinomial regression perform very well at determining which county the data originates from. It's also very promising that this very basic neural network achieved an accuracy of over 95%.

Classification Models	**Accuracy**
Multinomial Regression	0.9294118
LDA	0.8627451
QDA	0.945098
Neural Network	0.9529412

Table 3: The neural network was 3 layer full-connected network with a neuron shape of [9,9,9] with respective activation functions [linear, ReLU, sigmoid] trained for 5000 epochs using the sparse categorical cross entropy loss function.

Future Work

Trend and Seasonality Removal

Removing trends and seasonality from the mobility data will make the predictions more ubiquitous across regions. For example, parks are filled year around in California. This may not be true in colder regions such as Wisconsin and Michigan. Seasonality issues such as this could cause problems in predicting cases in counties for which the models do not have any seasonal information.

Aiding Healthcare Administration Decisions

The ability to predict future cases could be extremely useful not just for policy makers but for hospital and health department administrations. Being able to take the mobility and number of new cases this week and then determining how many people will be symptomatic next week would give hospitals and medical teams time to prepare adequately for an influx of patients, possibly preventing a catastrophe.

Population Density Estimation

Replacing the county parameter used in this project with the population density would give us a new parameter, that is likely informative and thus useful for predictive models. Furthermore, we could build more robust models that can be tested on new counties since population densities are regularly estimated. Perhaps we could even use this data set to build models which can accurately predict population density. This could be useful for determining true population sizes in counties across the US.

Behavior Inference

The data could possibly be used to infer interaction behavior amongst different populations. Suppose we notice two regions have similar density and similar mobility patterns, yet have vastly different infection rates. This could give us a clue as to which populations are following measures to reduce COVID-19 exposure, such as wearing masks and social distancing. Perhaps we can even determine which counties have low exposure rates given their mobility and density and investigate the factors which lead to the lower rates of infection.

References

[1] Orange County Health Care Agency. *COVID-19 Case Counts and Testing Figures*. https://occovid19.ochealthinfo.com/coronavirus-in-oc. Accessed on 2020-07-25. 2020.

[2] Stephen A. Lauer et. al. *The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application*. https://www.acpjournals.org/doi/10.7326/M20-0504. Accessed on 2020-07-25.

[3] Apple. *Apple Mobility Data*. https://www.apple.com/covid19/mobility. Accessed on 2020-07-25. 2020.

[4] Google. *Google Mobility Data*. https://www.google.com/covid19/mobility/. Accessed on 2020-07-25. July 2020.

[5] California Department of Health. *COVID-19 Cases*. https://data.ca.gov/dataset/covid-19-cases. Accessed on 2020-07-25.

[6] New York Department of Health. *nychealth/coronavirus-data*. https://github.com/nychealth/coronavirus-data. Accessed on 2020-07-28.

[7] Harvard Health Publishing. *If you've been exposed to the coronavirus*. https://www.health.harvard.edu/diseases-and-conditions/if-youve-been-exposed-to-the-coronavirus.. Accessed on 2020-07-28. 2020.

[8] *USAFacts.org*. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/. Accessed on 2020-07-25. 2020.

[9] Shuai Huang Zeye Liu and Wenlong Lu et al. *Modeling the trend of coronavirus disease 2019 and restoration of operational capability of metropolitan medical service in China: a machine learning and mathematical model-based analysis*. https://ghrp.biomedcentral.com/articles/10.1186/s41256-020-00145-4. Accessed on 2020-07-25.