

CS 254 Machine Learning

County Level COVID-19 Prediction

Matt Zahar, Parker Strawbridge, Isaac Padberg

1. Introduction

Our country's response to COVID-19 was varied and not organized. Each state had their own idea of what response was needed, and as a result, some states fared better than others. We plan to inspect New York Times COVID-19 data, mask data based on specific counties, and census data to get an idea of population density. This data will be used to help hospitals allow them to have an idea of their patient increase. Therefore allowing them enough time to order proper materials so they can maximize the number of patients that they can see.

2. Problem Definition and Algorithm

Our inputs will be: percentages of varying levels of mask use, county population, and county size.

Our output will be the predicted number of confirmed cases of COVID-19 for each county.

This is an interesting problem because if our country had access to this model, in the hands of a competent leader, it could have been used to create rules and regulations to greatly reduce the mortality rate and create lasting effects on our economy. Additionally, this model is important because it could inform hospitals, city officials, and the general public on the severity of COVID-19 in their county.

3. Dataset

We will be using three different datasets. The first is a regularly updated csv looking at county level COVID-19 information in the United States, this set includes the number of cases and deaths in each county. The second dataset is mask wearing data broken down to the county level, it includes estimated percentages of the population of how often people wear masks out in public. The third dataset that we will be using is United States census data for population estimates in 2019 by county. All of the datasets are labeled, the only changes needed to be made are removing “county” from the county column in the county data csv from the census.

4. Related Work

The problem is to predict the amount of cases in individual counties across the country. Our approach to the problem of predicting COVID-19 cases is different because we have broken it down to smaller regions. This will help us with the overall accuracy of our model and

hopefully result in a better outcome. One way in which we could improve our model would be to add data about the number of tests done in each county. This would allow us to better understand the state of each county and predict their covid numbers with better accuracy.

5. Bibliography

The New York Times. (2020, July 28). Nytimes/covid-19-data. Retrieved September 25, 2020, from

<https://github.com/nytimes/covid-19-data/blob/master/mask-use/mask-use-by-county.csv>

The New York Times. (2020, March 03). Covid in the U.S.: Latest Map and Case Count. Retrieved September 25, 2020, from

<https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html>

The New York Times. (2020, September 25). Nytimes/covid-19-data. Retrieved September 25, 2020, from

<https://github.com/nytimes/covid-19-data/blob/master/live/us-counties.csv>

United States Census. (2020, March 26). Retrieved September 25, 2020, from

<https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/>