Learning Systems
Blueprint Report

Caltech CS 156b
April 10^{th}, 2020

---

1 Group

- Group Members: Daniel Echeverri, Alex Ogren, Josh Lassman
- Group Name: Tentin-Quarantino
- Website: http://fishpoopsoup.com/tentinquarantino.html

2 Data

So far, only a few data sets have been explored. Primarily, the three sets we have used to create our premliminary models are the two New York Times data sets and a mobility dataset:

- `./data/us/covid/nyt_us_counties.csv`
- `./data/us/covid/nyt_us_states.csv`
- `./data/us/mobility/DL-us-m50_index.csv`

In the future, we hope to add many more data sources to our model. On the top of our list is mobility and hospital data, as well as international data sets.

3 Models

We have tried two different types of models, Neural Networks and SEIIR-QD, which is one of the Epidemiological models detailed in the models.ipynb provided by the TAs.

Future algorithms that we look at may involve Gaussian Process Regression, Ensemble Kalman Inversion, or Bayesian Optimization.

3.1 Neural Network

Both a simple neural network and a recurrent neural network were implemented. Both

Simple Neural Network: This took in inputs of days since January 1^{st}, latitude, and longitude. The output was the number of deaths on that day. This model was trained on data from every day and every county with a unique fips code in the `nyt_us_counties.csv` file. We found this method to be incredibly ineffective at prediction.

Recurrent Neural Network: This model was trained only on sequences of death data. The input was a ten day sequence of deaths for a specific county, and the output was a prediction for the 11^{th} day. This was trained on all ten day data sequences for each unique county within the `nyt_us_counties.csv` file. This performed much better on the training data, with a final RMS training error of around 6.2. However, as

expected, this model severely overfit, as is shown in Figure 1, in which this model was used to predict the last two weeks of New York City data from the preceding ten days.

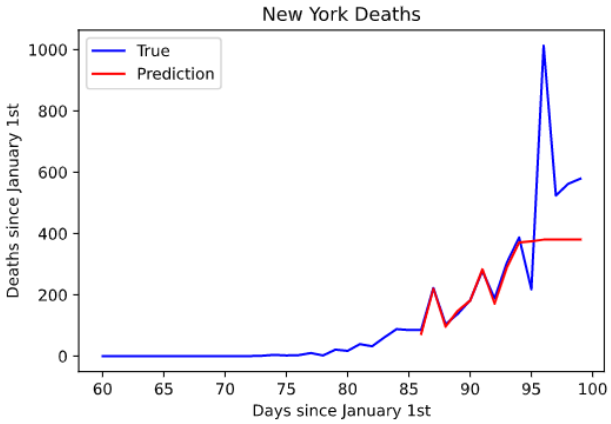

Figure 1: RNN New York City Predictions

3.2 SEIIR-QD

This model is an Epidemiological model that models flow rates between different portions of the population. This ultimately allows you to write a system of differential equations that can be solved to make future predictions based on the parameters that govern this dynamic system. In order to find the parameters, we use the Scipy least squares optimizer to find the best fit of parameters to the data. We fit both to the death data and the infectious symptomatic data. Below are a number of our own modifications that we made to the algorithm that was passed to us by the TAs:

1. In the error function that the optimizer calls, we have prescribed the *death* data to be 10 times more important than the *infectious symptomatic* data. This is because we don't have tons of faith in the latter.

2. In the error function that the optimizer calls, we have prescribed the *infectious symptomatic* data to be transformed by a LeakyReLU function. This is because we expect that if the data that we have on *infectious symptomatic* population has error, then that error is in a specific direction. We expect these values to be *under*reported.

3. In the error function that the optimizer calls, we have prescribed more recent data to be more important than data far in the past. If the epidemic dynamics are changing due to some unmodeled factors, it is most important that we capture the current epidemic dynamics rather than the past epidemic dynamics. Further, more recent data is more likely to be accurately modeled by the epidemiological models because these models rely on the Law of Large Numbers.

4 Future Work

In our next models, we think it would be very useful to incorporate mobility data. We have thought of several ways to do so:

- Include the raw mobility data given by Google, Descartes, and/or Unacast.
- Include raw data that describes mobility between states and counties.
- Implement a graph neural network in which the graph is formed by counties as nodes and edge data is informed by one of the above data sets.