Modelling a Pandemic: A Comparative Study

Intro

Hello. I am Ruoyu. Let me guide you through the comparative study on modelling a pandemic, conducted by Xingjian and yours truly (Ruoyu).

Background

In the United States, the ongoing COVID-19 pandemic presents an urgent challenge because of its contagious nature and frequently changing characteristics. To help prevent the spread of COVID-19, various models are applied in COVID-19 case predictions.

In our study, we will focus on two of them: the SIRD and the ARIMA. The SIRD, short for Susceptible-Infectious-Recovered-Deceased, is a classic compartmental model in epidemiology, whereas the ARIMA, or the Auto-Regressive Integrated Moving Average model, is a time series machine learning model for forecasting time series.

To find the most suitable model, here we propose our problem: *Which model is better for predicting infectious cases during the developing stage of a pandemic like COVID-19, SIRD or ARIMA?* In our study, we pay our attention to the developing stage, because this is often the time when people become aware of the pandemic. We believe that providing good predictions will greatly help policy makers to evaluate the pandemic situation.

Data Preparation

Here we choose to apply our models on the statistics for Wisconsin, an inland state of moderate population density and less population movement.

First, to prepare the data for our problem, we obtained the data with the three data sources, as listed here. With the help of pandas and NumPy, two very commonly used Python libraries, we applied data cleaning and extraction. Additionally, with the help of the data definitions provided by our major data source, we were able to extract the numbers of interest, so that we may proceed with our exploratory analysis.

Exploratory Data Analysis

We can plot out the data we prepared. Here we observe two critical points: at the end of April 2020, the statistics for recovered cases were made available; around the middle of November, the number of infectious cases experiences a turning point. We may therefore regard the dates in between as the developing stage. This is the stage where the number of infectious cases keeps increasing, and effective control needs to come into force as soon as possible.

Training the Models

We would like to limit the data available, because under real circumstances, no one can wait. With only 50% of the data as train data, we proceed with the model training.

We first train the SIRD model. The model, defined by a system of ordinary differential equations, is displayed here. With the help of statistical analysis libraries, we can integrate the system of ODEs, and apply Powell's method, an algorithm for minimising, to obtain the fitted model. Here is the result – the model fits reasonably well, despite some uncertainties observed.

We then tackle the ARIMA model, by grid searching for the best parameters, initialising a machine learning model with the best parameters, and training the model with the readily available machine learning libraries. In this part, the data tackling is more straightforward, as we only utilise the infectious cases. Here is the fitting result, which fits very well.

Predicting Test Data

When it comes to predicting test data, we can observe that the SIRD seems to fit better for the long run, while ARIMA, with a smaller root-mean-square error, provides better short term predictions. Another key observation is that the ARIMA predictions are concaving downwards, which should not be the case in the developing stage.

These observations seem to give the following conclusion: the SIRD model, which describes the virus transmission, comes up with more reliable results, whereas the ARIMA model, though not making use of the traits of the virus, can also give good short term predictions, possibly because of its dependence on previous values.

Another Case Study

Before we draw a solid conclusion, it would be better to have a study on another state. This time, we choose Kentucky, another inland state with similar conditions. Having followed the same procedure, we observed the following results. This time, the result is obvious – the SIRD model predicts much more accurately than the ARIMA model.

Having a closer look, we can observe that the SIRD model predicts quite accurately for the first few weeks, though it tends to give an overestimate for later dates. If we have enough confidence in our model, this would mean that the efforts to hold back the pandemic finally paid off, and hence, signal the transition from the developing stage to a new phase where things are getting back in control. The root-mean-square error for the prediction results is around 7,000, which is thought to be attributed to the overestimation on later dates.

The ARIMA model, however, has an RMSE about twice as much – the predictions do not seem close to the actual numbers, and they keep going away. This reveals the limitation of the ARIMA model – as mentioned earlier, it does not consider the characteristics of the disease, and may interpret the trend in an unexpected way.

Conclusion

Taking all the observations above into consideration, we would like to conclude our study as follows. The SIRD model, a mathematical one, tends to give more reliable predictions for pandemic statistics during its developing stage, because it well describes the virus transmission. The ARIMA model can give good short term predictions in certain cases, which can be attributed to the fact that the machine learning model focuses more on the pattern of historical values. It is, however, not so good at making long term predictions, because it only takes the infectious cases into consideration, and only treats them as general numbers with no special meanings. Things may go wrong in that way. Therefore, the SIRD model, giving more meaning to the data acquired, is better for predicting infectious cases during the developing stage of a pandemic. Thus, the SIRD model, on the basis of its simulation on the pandemic, may be the one that can be utilised by people, including policy makers, to make predictions, so that more meaningful results can be observed.

Here is the list of our references, and you may visit the links in the slides attached. In this project, while I (Ruoyu) am in charge of the technical details and parts of the project presentation, Xingjian takes responsibility for the overall idea, the overall structure and the preparation of this presentation.

We have come to the end of the project presentation. Thank you, and please feel free to reach us for any queries.