Predict the Spreading of Coronavirus

Team: Toolman
Team member: Shanli Ding, Qihang Mao

Covid-19 is probably the most important event in the world, we want to build a model based on this event as our final project. We found the data set of this event, and then we built a model to predict the spread of covid-19, to see how covid-19 spread around the world. We can draw some conclusions from our model and know how serious covid-19 is and what we can do to stop the virus.This can be a data collect,cleaning then use different models to train then get the related predictions task. We think our project is creative, because we want to do a project for a real event, and can learn something from our project. As the data keeps updating, our model prediction can be more accurate and meaningful. Shanli mainly works on model building and training, and shaping output results into the format we want. Besides, Qihang mainly works on data analysis and cleaning, and making visualizations for cases study.

Our data from Kaggle

Kaggle dataset:

https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

`covid_19_data.csv`

- Sno - Serial number
- ObservationDate - Date of the observation in MM/DD/YYYY
- Province/State - Province or state of the observation (Could be empty when missing)
- Country/Region - Country of observation
- Last Update - Time in UTC at which the row is updated for the given province or country. (Not standardised and so please clean before using it)
- Confirmed - Cumulative number of confirmed cases till that date
- Deaths - Cumulative number of of deaths till that date
- Recovered - Cumulative number of recovered cases till that date

And files of time series for confirmed cases, deaths and recovers. And there are around 17.4k samples for COVID-19. And there are 8 features excluding the serial number column in the COVID-19 csv file.

In our project, we first load the data from Kaggle, and then find the data keys to determine how to clean and use them. Let's take the global confirmed cases as the first chart to see how the growth trend of covid-19 is on a global scale. We can see the number of confirmed cases grows slowly at the beginning, and then erupts 60 days later. The trend of death curve is consistent with that of recovery cases, which shows the correlation of these three curves. Then

we combine the three curves into a single graph to make it easier for readers to find the relationships. As can be seen from the summary chart, we can see that the recovery curve is higher than the death curve, which means that the number of recovered cases is far more than the number of dead cases, proving that our response to the epidemic is effective and becoming more mature. Then we try to see how covid-19 in different countries has responded to this incident. We split the case data by country / region, counted the total number of confirmed cases in the time series, and sorted them, and found the top 5 countries with the most confirmed cases: "United States", "Spain", "Italy", "France", "Britain" and our home: "China" and another big country "Canada" as 7 country cases to generate a new chart. According to the new confirmed case chart, we find that the curve trend of different countries is consistent with the world trend. China has an earlier outbreak than the other six countries, and the situation of the United States is more serious than other countries. Then we followed their death curves and found the same results.But the Italy death case can be higher than other countries at the same level. For the recovery curve can be different, unlike other countries, China's recovery rate is much higher than other countries. I think this may be the result of China's medical treatment and covid-19 free treatment policy. Spain has a high recovery rate after 60 days, and the US's recovery rate has increased a lot even over Spain at around 90 days, which shows the US has concentrated on this big event and works hard.

After using the existing data, we try to build models to predict: simple linear regression, polynomial regression, ridge regression, SVR and SGD regression. The auxiliary function of training test segmentation is used to extract the test data set and training data set according to the ratio of 1:3. However, after testing all models, we found that simple linear regression and SGD regression are not suitable for predicting the future confirmed cases worldwide. Because linear regression and SGD regression both output a line similar to linear, we can't use such a line model to predict a strange curve. Finally, we agree to use SVR as our main research model. After that, we tried to make predictions for China, the United States and Canada, but the prediction curve was not ideal. We have considered many reasons, the most important is our lack of data records. However, we still hope that we will not receive a lot of records about coronavirus. So when we explain our project, we conclude that the number of confirmed cases of coronavirus will still grow rapidly. Finally, stay healthy and fight for the coronavirus.

Github: https://github.com/Ding3LI/covid-19_pred (presentation video is in github)

Data resources: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset