Accurate prediction and review of COVID-19 development trend

Robin Chiang

January 01, 2021

1. Introduction

1.1 Background

Coronavirus is a family of viruses that are named after their spiky crown. The novel coronavirus, also known as SARS-CoV-2, is a contagious respiratory virus that first reported in Wuhan, China. On 2/11/2020, the World Health Organization designated the name COVID-19 for the disease caused by the novel coronavirus. Now the virus is sweeping the world, a serious threat to human health and well-being and life. As of 2 January 2021, more than 83.9 million cases have been confirmed, with more than 1.82 million deaths attributed to COVID-19. Therefore, it is beneficial for all human beings to accurately predict the trend of covid-19 in the future and make appropriate prevention. This report aims at exploring COVID-19 through data analysis and projections.

1.2 Problem

Symptoms of COVID-19 are highly variable, ranging from none to severe illness. The virus spreads mainly through the air when people are near each other. It leaves an infected person as they breathe, cough, sneeze, or speak and enters another person via their mouth, nose, or eyes. It may also spread via contaminated surfaces. People remain infectious for up to two weeks and can spread the virus even if they do not show symptoms. Data that might contribute to determining COVID-19 development trend might include it performance in different countries, daily confirmed cases and daily deaths cases. This project aims to predict whether COVID-19 will slow down or intensify in the future based on these data.

1.3 Interest

Obviously, this virus is attacking people all over the world. Therefore, the World Health Organization, government authorities, public health experts and ordinary citizens are very interested in whether they can accurately predict the development trend of COVID-19 and effectively prevent the spread of the epidemic.

2. Data acquisition and cleaning

2.1 Data sources

This is a daily updating version of COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). The data updates every day at 6am UTC, which updates just after the raw JHU data typically updates.

2.2 Data cleaning

The data in the data table is very simple and does not need to be cleaned up too much. However, there are several problems with the datasets. First, the new ***Date*** column are all string with ***mm/dd/yy*** format, therefore we have to convert it to datetime values. Second, replacing missing value ***NaN***. We can find a lot NaN in the ***Province/State*** by running the test, and that makes sense as many countries only report the ***Country/Region*** data. However, there are 1,602 ***NaN*****s** in ***Recovered*** and let's replace them with 0. Third, there are COVID-19 cases reported from 3 cruise ships: Grand Princess, Diamond Princess and MS Zaandam. These data need to be extracted and treated differently due to ***Province/State*** and ***Country/Region*** mismatch over time.

2.3 Feature selection

After data cleaning, there were 95,472 samples and 9 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features such as, ***Province/State, Lat*** and ***Long***. Let's aggregate data from ***Province/State*** into the total number of those countries, and remove ***Lat*** and ***Long*** of non-critical data. Next, aggregate data

into ***Country/Region*** wise and group them by ***Date*** and ***Country/Region***, and the total count of ***Confirmed***, ***Deaths***, ***Recovered***, ***Active*** for the given ***Date*** and ***Country/Region*** will be summarized one by one. Now add day wise ***New cases***, ***New deaths*** and ***New recovered*** by deducting the corresponding accumulative data on the previous day.

Table 1. Simple feature selection during data cleaning.

Kept features	Dropped features	New features
Country/Region	Province/State	
Confirmed, Deaths, Recovered, Active	Lat, Long	
Date		New cases, New deaths, New recovered