Analysis of COVID-19 Based on Several Machine Learning Techniques

Tonglin Sun\textsuperscript{1,\textit{a}}

\textsuperscript{1}School of Engineering, Tufts University, Medford, Massachusetts MA 02155, United States;
\textsuperscript{\textit{a}}tsun06@tufts.edu

\textbf{Abstract.} COVID-19 sweeps the world with high infection and high death rates. It is essential for researchers to find an effective model to predict the trend of epidemic. With the good performances of several traditional models in predicting and analyzing previous epidemics, we compare several popular machine learning methods, including multi polynomial regression, logistic growth model and Long Short-Term Memory (LSTM) in epidemic prediction. We use least squares method for feature selection to determine the most relevant features and we also scale the data according to different experiment environments. We measure the accuracy using Mean Squared Error (MSE) and R\textsuperscript{2}. We conclude that the LSTM model is the most effective model among all the competitors with the highest R\textsuperscript{2} (R\textsuperscript{2} = 0.97). We find that LSTM model is the most effective model among all the competitors. Our study gives a good example of feature and model selection for epidemic prediction and attempts to make a significant contribution to the government and hospital to supply the public resources and provide drugs to handle the incoming issues.

1. \textbf{Introduction}

The purpose of this article is to use machine learning and deep learning algorithms to explore the development trend of the COVID-19 epidemic in the United States and the world and also try to make predictions and analysis based on existing data. COVID-19 has become the biggest difficulty faced by the world. Applying machine learning algorithms combined with mathematical models to analyze and predict the COVID-19 epidemic will play a guiding role in fighting the epidemic [1]. In 2003, some scholars used the logistic growth model to simulating the development trend of the SARS epidemic and achieved relatively good results [2]. However, machine learning algorithms were not popular at the time so there were not many epidemic predictions using machine learning as an auxiliary tool. For the COVID-19 epidemic, infectious disease models such as SIR and SEIR were also used in the epidemic research [3]. Liu et al. used the SEIR model to predict the COVID-19 epidemic in three countries including South Korea [4]. Chen et al. developed a new mathematical model named Bats-Hosts-Reservoir-People transmission network model for simulating the potential transmission from the infection source (probable be bats) to the human infection [5]. This time we use several classic machine learning algorithms and models suitable for epidemiology to predict the development trend of the epidemic. We have obtained a total of 181 days of COVID-19 data from January 22nd to July 20th of the entire United States from John Hopkins Coronavirus Resource Center and use the least square method to determine the features to be selected. Then, we apply machine learning algorithm to analyze and predict the number of daily positive increased. We also obtain historical data on the Wuhan COVID-9 epidemic from the official website of the China Centers for Disease Control and Prevention and use the logistic growth
model to fit the data of Wuhan and the United States respectively to compare the development trend of the epidemic in the two places. Finally, we use the long and short-term memory (LSTM) neural network model to predict the daily positive increase on the seventh day based on that of the last six days throughout the United States. The results show that LSTM has the most stable and best prediction results. The results demonstrate that several other models can be interfered by the experimental environment and external factors, which will lead to unsatisfactory prediction results in certain situations.

2. Methods

2.1. Dataset
There are two datasets collected for this experiment: data of the U.S. and data of Wuhan. They correspond to different periods of time. We get the US dataset from John Hopkins Coronavirus Resource Center and it contains COVID-19 data of the United States for a total of 184 days from January 22nd to July 20th. It is our main research object. The Wuhan dataset is used to compare with the United States dataset; both are later fitted into the logistic growth model. We obtain this dataset from the official website of the China Centers for Disease Control and Prevention and it contains the COVID-19 data of Wuhan for a total of 122 days from December 1st to April 3rd.

2.2. Data Preprocessing

2.2.1 Data Segmentation
The very first thing is to separate the data into three portions. As shown in Figure 1 and Figure 2 below, we can see that the development of the epidemic has gone through three different stages during this period.

These three stages may due to different environmental conditions. So, we split it into 3 portions and use different machine learning models to fit data of each of these three portions. The first portion is from day 1 to day 74, the second portion is from day 75 to day 145, the last portion is from day 146 to day 184. Then we randomly select ten percent of the data as test set and ninety percent as training set.

![Figure 1. Curve of total number of confirmed cases and total number of deaths over time.](image-url)
2.2.2. Feature Selection
Since the dataset we obtain contains many different features, we need to select useful features from the entire dataset. Our fundamental purpose is to predict the number of confirmed cases so we need to determine which features are most relevant to it. We use least squares method for feature selection. At last we find that the two most relevant features are the total number of confirmed cases and the number of existing confirmed cases. Total number of confirmed cases can directly be found in the dataset. Existing confirmed cases means 'total confirmed number - number of recovered - number of deaths'.

2.3. Logistic Growth Model
Logistic growth model is also known as self-inhibition equation. It can be used to reflect the relationship between population size and time under the influence of environmental resistance [6]. In the case of the epidemic, we can regard human anti-epidemic behaviors such as wearing masks and government control as environmental resistance. The expression of the model is expressed as Eq. 1,

\[ P(t) = \frac{Kp_0 e^{rt}}{K + p_0 (e^{rt} - 1)} \]  

where \( P_0 \) is the initial capacity, \( r \) is the growth rate and \( K \) is the maximum environmental capacity [7].

We use the least square method to solve the parameters in the model. First, we need to define the initial value of the function. Here we set \( K = 300000, r = 0.8 \) and \( p_0 = 20 \). Note that the function will eventually converge as long as the initial value is not too outrageous.

Besides, the use of least squares method to find parameters also needs to define an error function. Here the error function is defined as the difference between the predicted value and the true value.

2.4. LSTM
LSTM is an artificial recurrent neural network (RNN) architecture used in the field of deep learning [8]. Unlike standard feedforward neural networks, LSTM has feedback connections [9] [10]. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video).
Figure 3 shows the structure of LSTM. There are three main stages inside LSTM [8]: 1) Forget stage; 2) Memory selection stage; and 3) Output stage.

During forget stage, LSTM will use the calculated $z^f$ (f means forget) as the forget gate. Control which parts of $c^{t-1}$ needs to be kept and which needs to be forgotten. And at memory selection stage, the input is selectively "memorized". The main job is to select and memorize the input $x^t$. The current input content is expressed by the $z$ calculated above. The gate control signal is controlled by $z^r$. As for output stage, it will determine which will be regarded as the output of the current state by using $z^o$. And the $c^o$ obtained in the previous stage is also scaled (by a tanh activation function).

First, we regularize the data to avoid interference caused by large differences in data value. In this experiment, we try to use data of first six days to predict the data of the seventh day using LSTM. So, we use ninety percent of the data as the training set and the remaining ten percent as the test set. Starting from the first day, take the data of the first six days as input and the corresponding data of the seventh day as output. And the batch size of the network should also be 6.

Due to the limited data size, we use a single-layer LSTM with 200 neurons. Use 'relu' as the activation function. The dropout ratio is set to 0.1 and 'adam' as optimizer. Finally, we let it iterate 500 times which is 'epochs = 500'.

3. Results and Discussion

3.1. Polynomial regression

| Date     | Positive | Positive increase | Recovered | Death |
|----------|----------|-------------------|-----------|-------|
| 20200401 | 225559   | 25773             | 7084      | 5204  |
| 20200331 | 199786   | 24694             | 5666      | 4216  |
| 20200330 | 175092   | 21897             | 4560      | 3355  |
| 20200329 | 153195   | 19655             | 4052      | 2778  |
| 20200328 | 133540   | 19724             | 3143      | 2262  |

We use the data of portion 1 to fit the polynomial regression. The results corresponding to the four
degrees from 1 to 4 are shown in Table 2.

| Degree | R^2         |
|--------|------------|
| 1      | -10.43     |
| 2      | -45.03     |
| 3      | -214655.87 |
| 4      | -432757701.33 |

Table 3. Features of portion 2

| Date     | Positive | Positive increase | Recovered | Death  |
|----------|----------|-------------------|-----------|--------|
| 20200610 | 1997360  | 20803             | 618125    | 106723 |
| 20200609 | 1976557  | 17175             | 609476    | 105838 |
| 20200608 | 1959382  | 17012             | 603143    | 104940 |
| 20200607 | 1942370  | 18823             | 590988    | 104280 |
| 20200606 | 1923547  | 23106             | 585470    | 103834 |

Table 4. Feature of portion 3

| Date     | Positive | Positive Increase | Recovered | Death  |
|----------|----------|-------------------|-----------|--------|
| 20200720 | 3813916  | 57948             | 1160087   | 133283 |
| 20200719 | 3755968  | 63907             | 1131121   | 132918 |
| 20200718 | 3692061  | 65180             | 1122720   | 132395 |
| 20200717 | 3626881  | 77233             | 1106844   | 131523 |
| 20200716 | 3549648  | 70953             | 1090645   | 130572 |

We find that degree = 1 performs best according to the cross-validation result. Then, we apply degree = 1 to do prediction and get R^2 = -93.74. This result looks quite ill-performing.

Similar to 1, we also use the same method to fit the function using the data of 2 and 3 respectively. As shown in Table 5 and Table 6, degree=1 is still the best for portion 2 and portion 3. The accuracy is improved when using degree=1 for polynomial regression: 0.82 for portion 2 and 0.95 for portion 3. Since R^2 reflects the relationship between the use of the regression model and the average, we can
see from the results that using the polynomial regression model is not much better or worse than taking the average for portion 1 and portion 2. The reason is that portion 1 and portion 2 do not have obvious regular changes. For example, for portion 2 we can see that the overall changes in the data are not significant and it is difficult for us to see obvious trends. But portion 3 shows an upward trend similar to a linear function as a whole.

3.2. Logistic growth model
As shown in Table 7, there are data of 181 days. We use the data of the first 140 days as the training set to train the model and then use the model to predict the data of the last 41 days.

The least squares method was used to solve the relevant parameters. Then we draw a comparison image of the predicted value and the true value.

Table 5. The result of cross-validation of portion 2

| Degree | R^2  |
|--------|------|
| 1      | 0.50 |
| 2      | -0.69|
| 3      | -0.76|
| 4      | -1.29|

Table 6. The result of cross-validation of portion 3.

| Degree | R^2  |
|--------|------|
| 1      | 0.70 |
| 2      | -0.06|
| 3      | -0.02|
| 4      | -0.08|

Figure 4. Comparison of the predicted value and the true value of the logistic growth model on U.S. COVID-19 data

Figure 4 shows the difference between the predicted value and the true value of applying the logistic growth model to predict the number of COVID-19 positive in the United States. The predicted values of the last 41 days are very different from the real values and even the growth trend is completely
different.

The result shows that the development of the epidemic in the United States does not conform to the logistic growth model after the transition period (P = K/2). In order to determine whether the model is applicable we collected historical data of the Wuhan epidemic and used the logistic growth model to fit the data.

Table 8 shows the data of the Wuhan epidemic for a total of 122 days from December 1st to April 3rd. This time we use the data of the first 84 days as the training set to retrain the logistic growth model and predict the number of positive in the remaining 38 days.

We can see in Figure 5 that the predicted value curve is very close to the true value curve, indicating that the development trend of the Wuhan epidemic is in line with the logistic growth model.

As shown below in Figure 6, we can see that the development of the epidemic in Wuhan and the United States has shown completely different trends after the transition period of the logistic growth model.

According to the definition of the logistic growth model we know that the reason of why P rises slowly after the transition period is environmental-resistance.

| Date   | Positive | Positive Increase | Recovered | Death  |
|--------|----------|-------------------|-----------|--------|
| 20200720 | 3813916   | 57948             | 1160087   | 133283 |
| 20200719 | 3755968   | 63907             | 1131121   | 132918 |
| 20200718 | 3692061   | 65180             | 1122720   | 132395 |
| 20200717 | 3626881   | 77233             | 1106844   | 131523 |
| 20200716 | 3549648   | 70953             | 1090645   | 130572 |

Figure 5. Comparison of the predicted value and the true value of the logistic growth model on Wuhan COVID-19 data
Figure 6: Comparison of logistic growth model for predicting the epidemic in the United States and Wuhan

| Date       | City     | Confirmed |
|------------|----------|-----------|
| 2020/2/12  | Wuhan    | 19558     |
| 2019/2/13  | Wuhan    | 32994     |
| 2019/2/14  | Wuhan    | 35991     |
| 2019/2/15  | Wuhan    | 37914     |
| 2019/2/16  | Wuhan    | 39462     |

Here environmental resistance should be the various measures taken by the government and individuals to fight the epidemic corresponding to the development of the epidemic.

At the beginning of the epidemic, the Chinese government took measures such as closing the city to prevent the spread of the epidemic. At the same time, citizens also follow the leadership to actively fight the epidemic. These have effectively prevented the further large-scale spread of the epidemic as an environmental resistance to the spread of the virus. However, the situation in the United States is not optimistic. This may be related to the government's failure to take effective measures to deal with the epidemic in a timely manner and the public's lack of attention to the epidemic.

3.3. LSTM

As illustrated previously, we did not get a stable and reliable accuracy rate from polynomial regression. So, we hope to try a more complex model. Given that we hope to predict the future data based on a period of time, LSTM is a suitable choice.

The predicted results are shown in Table 9. We get $R^2 = 0.974$. The results obtained by using LSTM are more accurate and stable than those previous models. What is more, there is no need for us to artificially consider environmental factors.

Table 9. Comparison table of predicted and actual values on the test set

| Predict     | Truth      |
|-------------|------------|
| 3023139.50  | 3042503.0  |
| 3085322.00  | 3101339.0  |
### Figure 7. Line chart of LSTM predicted value and true value comparison

#### 4. Conclusion
In this study, we compare several machine learning based methods such as Polynomial regression, logistic growth model and Long Short-Term Memory model. We also do feature selection and other pre-processing works to ensure that the experiment runs successfully. Although all several machine learning algorithms have a certain guiding role in epidemic prediction after preprocessing and feature selection, we find that LSTM is the most accurate and stable model among these three models. Both the logistic growth model and the polynomial regression model have certain limitations when used. If environmental factors are not suitable for them, their accuracy tend to be significantly lower. In contrast, LSTM has an accuracy rate of 0.95 or more which can provide a reference for predicting the COVID-19 epidemic. We believe that our findings will be helpful to infectious disease control and medical management of COVID-19 patients. In general, this article uses machine learning algorithms to achieve the goal of more accurately predicting the new diagnoses of the COVID-19 epidemic in a single day, which has a significant role in epidemic control and prevention.

#### References
[1] Santosh, K.C. COVID-19 Prediction Models and Unexploited Data. J Med Syst 44, 170 (2020). https://doi.org/10.1007/s10916-020-01645-z
[2] Zhou G, Yan G. SARS epidemiology modeling. Emerg Infect Dis. 2004 Jun. http://dx.doi.org/10.3201/eid1006.040173
[3] Ian Cooper, Argha Mondal, Chris G. Antonopoulos, A SIR model assumption for the spread of COVID-19 in different communities, Chaos, Solitons Fractals, Volume 139, 2020, 110057, ISSN 0960-0779, https://doi.org/10.1016/j.chaos.2020.110057.

[4] Liu Z, Magal P, Seydi O, et al. A model to predict COVID-19 epidemics with applications to South Korea, Italy, and Spain. SIAM News, 2020, 1: 1-2.

[5] Tianmu Chen, Jia Rui, Qiupeng Wang, Zeyu Zhao, Jing-An Cui, Ling Yin bioRxiv 2020.01.19.911669; doi: https://doi.org/10.1101/2020.01.19.911669

[6] Dmitry Kucharavy Roland De Guio. Application of Logistic Growth Curve. In Procedia Engineering, 2015.

[7] Wang Jin, Scott W. McCue, Matthew J. Simpson, Extended logistic growth model for heterogeneous populations, Journal of Theoretical Biology, Volume 445, 2018, Pages 51-61, ISSN 0022-5193, https://doi.org/10.1016/j.jtbi.2018.02.027.

[8] [Hochreiter and Schmidhuber1997] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. Neural Computation, 9(8):1735-1780.

[9] Alex Sherstinsky, Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network, Physica D: Nonlinear Phenomena, Volume 404, 2020, 132306, ISSN 0167-2789, https://doi.org/10.1016/j.physd.2019.132306.

[10] Roondiwala, Murtaza et al. Predicting Stock Prices Using LSTM. (2017).