Sequential Models for COVID Peak Prediction

## 

## Introduction

# Problem Statement

To predict by Corona Virus Peak in a Country by Machine Learning Methods

1. **LSTM**
2. **Generalized Linear Mixed Models (Next Experiment)**

## Materials and Methods

# Data

We dynamically take daily data from Our World in Data repository

https://covid.ourworldindata.org/data/owid-covid-data.csv

The important usable features in the data per location are as follows

population, population\_density, median\_age, aged\_65\_older, aged\_70\_older, gdp\_per\_capita, extreme\_poverty, cvd\_death\_rate, diabetes\_prevalence, female\_smokers, male\_smokers, handwashing\_facilities, hospital\_beds\_per\_thousand

The following data is non-sparse and important

* **Location**
* **Date**
* **Total\_cases**
* **New\_cases**
* **Total\_cases\_per\_million**
* **New\_cases\_per\_million**
* **Gdp\_per\_capita**
* **Population**
* **Population\_density**
* **Median\_age**

The following features are important sequential data by date for each location

* **Total\_cases**
* **New\_cases**
* **Total\_cases\_per\_million**
* **New\_cases\_per\_million**

We filter data for each Location

* **With Population > 1,00,000**
* **Staring from Total\_cases\_per\_million > 1**
* **And had No of weeks since above > 10**

All data is aligned with day 0 as the day when Total\_cases\_per\_million crosses 1 for the country

Orignal data has 209 Countries, after filtering 163 Countries are left

Total of **151 Countries** are found among which **101 Countries have peaked and 50 Countries including India have yet to peak** one week before

Further, we generate 2 features, as **moving average with a rolling window of 7 days**

* **moving\_average(new\_cases)**
* **moving\_average(new\_cases\_per\_million)**

Further, to predict if the peak is n weeks after this week

Further, we find the peak of each country and take the last (n+1)\*7 to (n)7 days previous to peak as sequential features (of length 7, other lengths can also be used) for the LSTM model with the following six features:

* **Total\_cases**
* **New\_cases**
* **Total\_cases\_per\_million**
* **New\_cases\_per\_million**
* **moving\_average(new\_cases)**
* **moving\_average(new\_cases\_per\_million)**

So as to form,

n\_countries\*7\*6 features for training

### For n = 1 week

Total of **151 Countries** are found among which **101 Countries have peaked and 50 Countries including India have yet to peak**

### For n = 3 week

Total of **127 Countries** are found among which **74 Countries have peaked and 49 Countries including India have yet to peak**

### For n = 6 week

A total of **80 Countries** are found among which **31 Countries have peaked and 49 Countries including India have yet to peak.**

**Data is so less it’s completely upon chance, experimentally verified.**

# Model - LSTM

LSTM based Sequential model

LSTM with cell units= 28, activation layers='relu'

FC Dense Layer 1, activation='sigmoid'

loss='binary\_crossentropy', optimizer= stochastic gradient descent

Learning rate=1e-10, decay=1e-6, momentum=0.9

# 

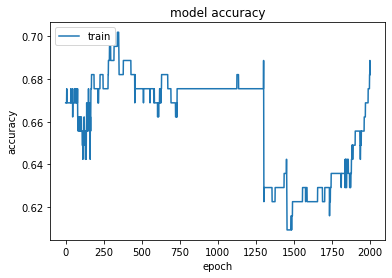
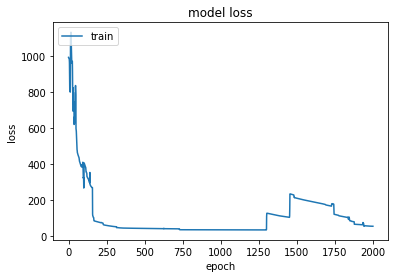
## Results

# LSTM Model Metrics

### For n = 1 week

There is no train-test split as to preserve data (which is not ideal), i.e. test set = train set

N = 151



### 

True Positive 83

True Negative 20

False Positive 30

False Negative 18

Accuracy 0.68

The probability that next week is the peak of new cases for India is 1.1e-11.

Therefore certainly next week is not the peak of new cases for India

### 

### Confusion Matrix

### [[20 30]

### [18 83]]

### Classification Report

### precision recall f1-score support

### 

### 0 0.53 0.40 0.45 50

### 1 0.73 0.82 0.78 101

### 

### accuracy 0.68 151

### macro avg 0.63 0.61 0.62 151

### weighted avg 0.67 0.68 0.67 151

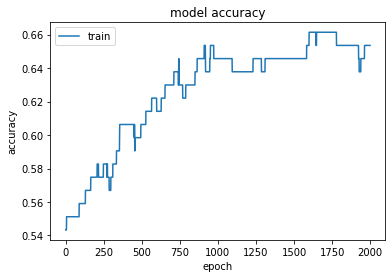
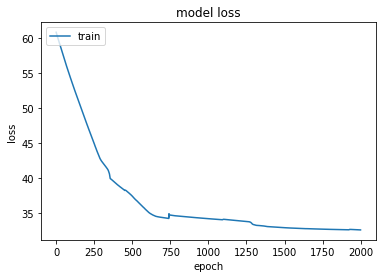
### 

### 

### For n = 3 week

There is no train-test split as to preserve data (which is not ideal), i.e. test set = train set

N = 127



True Positive 41

True Negative 42

False Positive 8

False Negative 36

Accuracy 0.65

Confusion Matrix

[[42 8]

[36 41]]

Classification Report

precision recall f1-score support

0 0.54 0.84 0.66 50

1 0.84 0.53 0.65 77

accuracy 0.65 127

macro avg 0.69 0.69 0.65 127

weighted avg 0.72 0.65 0.65 127

The probability that next week is the peak of new cases for India is 0.00010963.

Therefore three weeks after today has a very low probability of being the peak of new cases for India

### For n = 6 week

There is no train-test split as to preserve data (which is not ideal), i.e. test set = train set,

N = 80

Confusion Matrix

[[48 1]

[30 1]]

Classification Report

precision recall f1-score support

0 0.62 0.98 0.76 49

1 0.50 0.03 0.06 31

accuracy 0.61 80

macro avg 0.56 0.51 0.41 80

weighted avg 0.57 0.61 0.49 80

# Rough Conclusion

* We can roughly predict if the COVID new cases peak is exactly after 1 week with 68% accuracy, and after 3 weeks with 65% accuracy.
* At 6 weeks the data and learning are so insufficient that the prediction is almost completely chance and thus of not much use.
* With the majority of the difficulty in predicting given COVID peak is not in x weeks accurately predicting that it is not.
* It is not a fairly accurate prediction.

# Further Directions

1. Add features such as **Gdp\_per\_capita, Population, Population\_density, Median\_age** as covariates in the LSTM model possible as a logistic model.
2. Try out **Generalized Linear Mixed Models (Next Experiment)** with said data.
3. Try Further experimentation with LSTM based model.