CAPSTONE PROJECT PROPOSAL

UDACITY MACHINE LEARNING ENGINEER NANODEGREE

1. Project Overview

1.1. Problem Statement

On December 31st, 2019, the outbreak of "pneumonia of unknown cause" started in Wuhan City, Hubei Province, China. The virus was named novel Coronavirus, or 2019-nCoV in short. At January 23rd, there had been over 800 cases of 2019-nCoV confirmed globally, and it is still raising despite many effort from governments.

Amid the outbreak of COVID-19, it is a critical need to analyze and forecast the progress of the pandemic and the effective of current policy. An important task is to predict the number of confirmed cases and fatalities globally, and in each country in order to have appropriate preparation and policy. Many data scientist are trying to use AI to forecast these numbers by studying mass social interaction, analyzing patients travel records, and clustering high risk areas.

The goal of this project is to use time series record by days to forecast the number of cases (confirmed and deceased) in the next 14 days.

The method I use for modelling is inspired from this Kaggle notebook:

https://www.kaggle.com/saga21/covid-global-forecast-sir-model-ml-regressions

1.2. Metrics

In the context of this project, because of limited resource and data, I will only conduct Exploratory Data Analysis and use basic log-linear and logistics functions model to predict the number of cases based on time series data.

The metrics that I will use to evaluate the model is Root mean squared error (RMSE)

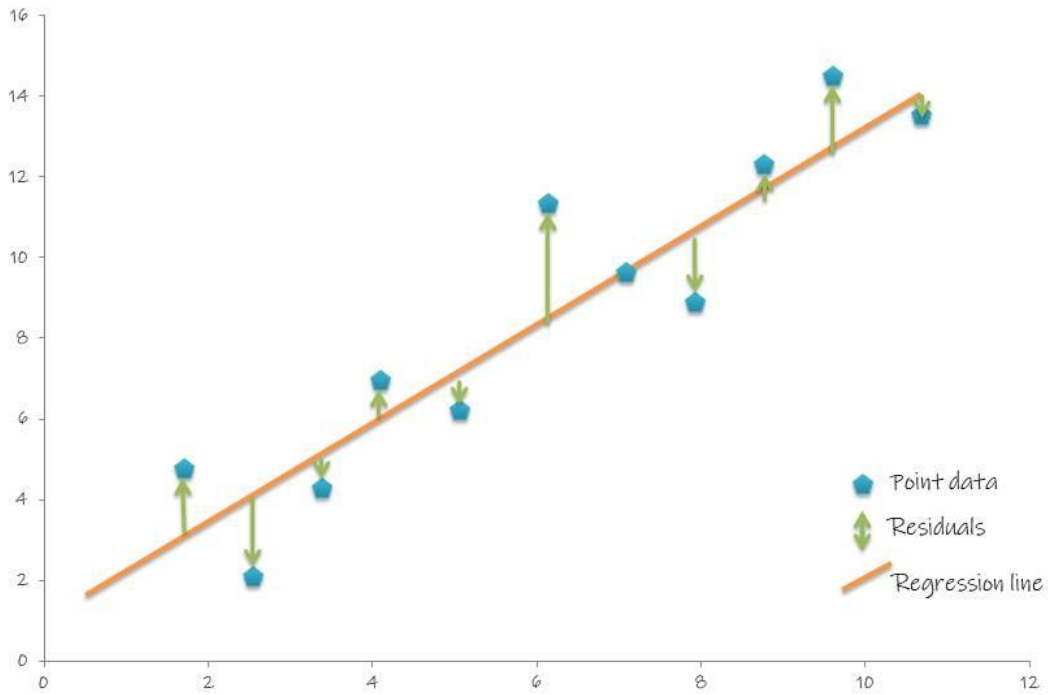

Figure 1: Root mean square error. Source: https://www.hatarilabs.com

To calculate the RMSE, this equation is used:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2}$$

With:

- n: number of samples
- p: predicted value
- o: observed value

Unlike other classification problem where we can use metrics like accuracy, precision, recall, because this is a regression problem, the RMSE should be effective enough to evaluate the performance of the model.

2. Exploratory Data Analysis

2.1. Dataset

There are many data sources available online, providing many insight and data in various aspect of the diseases. I will use the time series dataset collected and updated daily by Center for Systems Science and Engineering, John Hopkins University. The dataset can be downloaded from this github repository: https://github.com/CSSEGISandData/COVID-19

3. Methodology

The method I use for modelling is inspired from this Kaggle notebook:

https://www.kaggle.com/saga21/covid-global-forecast-sir-model-ml-regressions

I will use 3 models to forecast the number of cases

3.1. SIR model

SIR model is widely used to analyze the progress of transmitted diseases, including COVID-19. There are many versions of this model, some of which pay more attention to demography and external factors like lock down policy. I can only implement a simple model that monitor a population in 3 states:

- **Susceptible (S)**: The individual has not contracted the disease, but might be infected due to transmission from infected people. Initially, all of the population will be in S, and for each time step, some will be infected and is counted to I
- **Infected (I)**: The person has contracted the disease. The group initializes with 0, and is increased by the number of infected people and decreased by recovered people
- **Recovered/Deceased (R)**: The person who had been infected but has either recovered or deceased.

Figure 2: SIR model. Source: lewuathe.com

This model can be expressed by these differential equations:

$$\frac{dS}{dT} = -\frac{\beta IS}{N}$$

$$\frac{dI}{dT} = \frac{\beta IS}{N} - \gamma I$$

$$\frac{dR}{dT} = -\gamma I$$

With β the transmission rate and γ the recovery rate and N is the total size of the population.

SIR model can be implemented in many ways. I will simply run a numerical method Runge-Kutta to solve the differential equations system and fit the real data with the theory.

3.2. Log – Linear Regression

One thing we can observe when doing the EDA is that the transmission rate of the disease is exponential at the early stage (for about 10 - 15 days starting from the outbreak). I will try plotting the number of cases for that period and try applying the log function to see if we can use log-linear model to observe the spreading of this virus at the early stage.

3.3. Logistics function

In reality, the number of cases should look similar to a logistic function, with lower rate at the beginning and late stages of the outbreak, while higher at the middle when the disease starts to transmit.

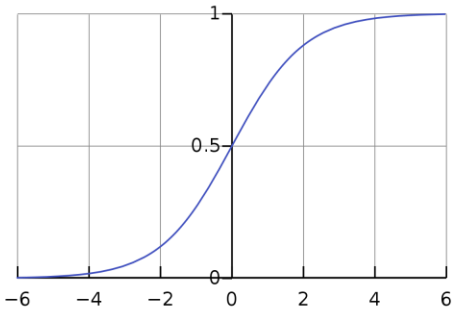

Figure 3: Logistic function. Source: Wikipedia

Logistic function can be expressed in the below equation:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

May 1st, 2020

Nguyen Manh Tuan

tuan.ngmanh@gmail.com