107062381 楊孝偉

Machine Learning HW1 Report

In this homework 1, I used Autoregression model to predict the COVID-19 cases for all countries.

At first, I tried to use the linear regression model for this homework, but after reconsidering that cases curve in some countries have flatten (Example : Taiwan), it is not really suitable to use linear regression model. Then I searched on the internet to look for other models that suits this homework more. Then I found Autoregression model.

Autoregression is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step. Based on this definition, it already gave me the sense that this model suits this homework as some cases are fluctuative. Therefore, predicting what are the cases for tomorrow needs to be based on the pattern of past cases.

First, I split the observations into training and test sets, with the test sets to be the last 7 days in the dataset and the training sets to be the remaining cases. The training sets of data are used to train the model to predict for the 7 days kept in test sets. After the model has been trained well, it will try to predict those 7 days in test sets and I compared the the values predicted by the model and the real test dataset to ensure that this model suits for this assignment . the statsmodel library provides an autoregression model where we must specify an appropriate lag value and trains the model. We can use this model by first creating the model AR and then calling the fit function to train it on the training data sets that I have provided. After it is fit, we can use the model to make prediction. The predictions are made using a walk-forward validation model so that we can persist the most recent observations for the next day. After everything is done, plot the final results of the predictions to a graph using pyplot and export results to excel with panda dataframe.

The model file is in .py file. It takes “data.csv” as the input dataset to the model. By simply running the whole code, it will automatically look for the file “data.csv” and then output the final prediction results for the 5 countries asked.