Machine Learning

Assignment 1: COVID-19 Forecast Report

106070038 杜葳葳

一、Model and Features

（一）前處理

1. 將其他不相關的欄位用 drop 移除，**僅留下日期、國家名稱和確診人數，並將日期設為 index**。
2. 將所有資料**順序倒反**，因為本來時間的排序是由現在到年初，為了預測方便，因此將所有資料的順序改為由年初到現在。
3. 有些國家的確診人數為負值，故將確診人數**小於 0 的改為 0。**
4. 執行完以上三個步驟產生 clean_data.csv 檔（如左下圖），產生一個國家名稱列表與各個國家資料筆數的列表 countries_size.csv （如右下圖）

clean_data.csv

dateRep	cases	countriesAndTerritories
21/03/2020	1	Zimbabwe
22/03/2020	1	Zimbabwe
23/03/2020	0	Zimbabwe
24/03/2020	0	Zimbabwe
25/03/2020	0	Zimbabwe
...	...	...
04/10/2020	7	Afghanistan
05/10/2020	44	Afghanistan
06/10/2020	145	Afghanistan
07/10/2020	62	Afghanistan
08/10/2020	68	Afghanistan

[47689 rows x 2 columns]

countries_size.csv

countriesAndTerritories	0
Afghanistan	273
Albania	214
Algeria	278
Andorra	209
Angola	201
...	...
Vietnam	279
Western_Sahara	166
Yemen	182
Zambia	204
Zimbabwe	202

210 rows × 1 columns

（二）模型

最初分析題目認為應該使用「**時間序列**」相關的演算法，因此有嘗試使用線性迴歸、ARIMA、LSTM，以下分別列出三種方法嘗試的過程：

1. **線性迴歸**：計畫找出**與 COVID-19 相關的外部因素**，然而對原始資料做了些視覺化分析與評估，個別差異大、難以找到共同點，且在其他開放資料平台很難找到一項數據 210 個國家皆有資料，除非使用自己訂定的平等，例如：國民防疫意識指標，由於對於傳染病學沒有太多研究，故放棄這個方法。
2. **LSTM**：需要花較長的時間做訓練，且感覺要**更大量的數據**做訓練效果會更好，嘗試手動對參數做些調整後，MAPE 仍然很大，於是後來沒有使用這個方法。
3. **ARIMA**：剛開始使用手動調整參數，後來發現 **pyramid-arima** 和 **pmdarima auto-arima** 的套件，可自動產生最 fit 的參數，但仍有幾個參數可以自行調整的彈性，最後採用 **pmdarima auto-arima** 套件。

- 經實驗後發現，**調整判斷模型好壞的統計指標**（default 是 AIC、手動改為 BIC），在有些國家會對 MAPE 造成蠻顯著的影響。
- 另外，**調整 p、d、q 三個參數的範圍**也會影響到 MAPE，故將三個參數的最大值與最小值皆設為 0 到 10，希望讓模型能更加 fit 原始資料，然而部分國家再將範圍加大時模型並不會變好，所以還是有些模型保留 Default 設定。
- 綜合以上，嘗試了**三組模型參數**，以該國家「**所有資料扣除 10/2-10/8 的資料**」作為 **Training Set**、「**10/2-10/8 的資料**」作為 **Validation Set**，跑完全部 210 個國家，**分別計算每個國家的 MAPE，選擇最小的作為該國家的模型參數**。但因為每個國家的資料筆數不同、資料分布與特性也不同，模型的準確度差異蠻大的。

  下圖為各國 MAPE

MAPE (Mean absolute percentage error)

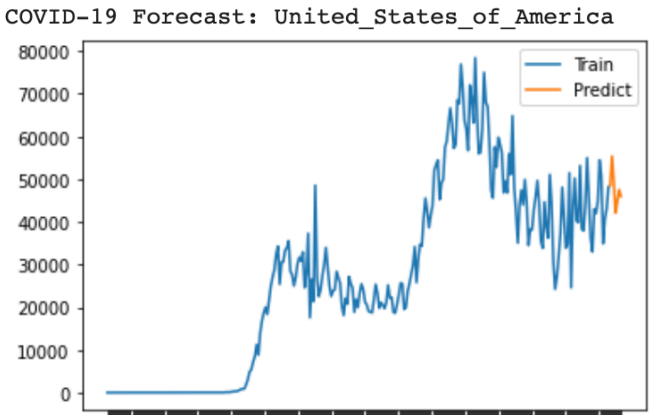
	countriesAndTerritories	Parameter1	Parameter2	Parameter3
0	Afghanistan	0.9990115309952480	0.7584223111835490	0.7584223111835490
1	Albania	0.16991104792219200	0.16991104792219200	0.16991104792219200
2	Algeria	0.06080750672633300	0.06080750672633300	0.06080750672633300
3	Andorra	0.23292263292263300	0.25328005328005300	0.23292263292263300
4	Angola	0.20540349012312900	0.20540349012312900	0.20540349012312900
5	Anguilla	0.0	0.0	0.0
6	Antigua_and_Barbuda	0.1142857142857140	0.1142857142857140	0.1142857142857140
7	Argentina	0.21919211499952500	0.21919211499952500	0.21919211499952500
8	Armenia	0.29301081075627100	0.29301081075627100	0.29301081075627100
9	Aruba	1.157936507936510	1.6391156462585000	1.6391156462585000
10	Australia	0.4758981214863570	0.3959015494309610	0.3959015494309610
11	Austria	0.2969245674816790	0.2969245674816790	0.2969245674816790
12	Azerbaijan	0.2980953379517440	0.2600099131763340	0.2600099131763340
13	Bahamas	0.2349901001741940	0.25506354660993800	0.2349901001741940
14	Bahrain	0.5330527390096780	0.5330527390096780	0.5330527390096780
15	Bangladesh	0.09577728326496710	0.09577728326496710	0.09577728326496710
16	Barbados	0.6666666666666670	0.38095238095238100	0.38095238095238100
17	Belarus	0.07383246045176310	0.09455441613285480	0.07383246045176310
18	Belgium	1.2074928768804600	1.2074928768804600	1.062782449270180
19	Belize	0.8203683961146650	0.7806653724125190	0.7806653724125190
20	Benin	0.12433862433862400	0.12433862433862400	0.12433862433862400

- 最後 prediction 的實作方式是用一個 for loop 跑 210 次，每次迴圈跑一個國家，將所有資料作為訓練資料，最後再輸出成 csv 檔。

  下圖為以美國為例，auto-arima 跑出來的模型參數

```
Best model:  ARIMA(5,1,5)(0,0,0)[0] intercept
Total fit time: 13.456 seconds
                               SARIMAX Results
==========================================================================
Dep. Variable:                      y   No. Observations:                  283
Model:               SARIMAX(5, 1, 5)   Log Likelihood               -2764.120
Date:                Mon, 12 Oct 2020   AIC                           5552.239
Time:                        09:39:18   BIC                           5595.942
Sample:                             0   HQIC                          5569.765
                                  - 283
Covariance Type:                  opg
==========================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------
intercept    360.7395    320.068      1.127      0.260    -266.583     988.062
ar.L1          0.4693      0.130      3.601      0.000       0.214       0.725
ar.L2         -1.1366      0.113    -10.042      0.000      -1.358      -0.915
ar.L3          0.3256      0.185      1.758      0.079      -0.037       0.689
ar.L4         -0.7123      0.109     -6.533      0.000      -0.926      -0.499
ar.L5         -0.3162      0.116     -2.732      0.006      -0.543      -0.089
ma.L1         -1.0848      0.126     -8.610      0.000      -1.332      -0.838
ma.L2          1.4304      0.153      9.373      0.000       1.131       1.730
ma.L3         -1.0976      0.201     -5.447      0.000      -1.493      -0.703
ma.L4          0.9890      0.149      6.657      0.000       0.698       1.280
ma.L5         -0.2187      0.113     -1.943      0.052      -0.439       0.002
sigma2     2.044e+07      0.007   2.78e+09      0.000    2.04e+07    2.04e+07
==========================================================================
Ljung-Box (Q):                       58.32   Jarque-Bera (JB):               370.24
```

下圖為以美國為例，藍線為歷史資料、橘線為模型預測的確診人數



下圖為模型跑出來的結果

Model Prediction:

	Greece	India	Russia	Turkey	United_States_of_America
10/9	366	75759	11576	1581	48507
10/10	359	70819	11716	1581	55395
10/11	361	68283	11943	1581	49237
10/12	362	68014	12128	1581	42093
10/13	363	68110	12313	1581	44803
10/14	364	67518	12486	1581	47542
10/15	366	66381	12650	1581	45989

二、How to use the model file

讀入從老師提供網站下載的 csv 檔（將檔名改為 input），執行所有程式碼，得到 output.csv，內含五個國家 10/9-10/15 確診人數預測。

output

	Greece	India	Russia	Turkey	United_States_of_America
10/9	366	75759	11576	1581	48507
10/10	359	70819	11716	1581	55395
10/11	361	68283	11943	1581	49237
10/12	362	68014	12128	1581	42093
10/13	363	68110	12313	1581	44803
10/14	364	67518	12486	1581	47542
10/15	366	66381	12650	1581	45989

若將程式碼中 if 條件式刪除，則可以預測所有國家的確診人數。

三、Summary

（一）使用的套件：panda、matplotlib、pmdarima.arima

（二）嘗試 LSTM、Linear Regression 和 ARIMA 三種方法，最終使用 ARIMA 模型，用 MAPE 來調整 p、d、q 參數的範圍和 AIC/BIC，過程中有切 Train 和 Validation，最後的結果用所有的資料當訓練集。