概要

使用 SIRD 模型分析 Covid-19 在不同國家的疫情，並藉由現有資料預測往後幾天的結果。

資料分析

接下來我們將對手頭上現有的不同的地區每天的人口資料進行分析，了解該地區因疫情而產生的人口變化。我們分成下面幾個步驟來說明資料分析的內容。

1 資料來源

資料來源為 John Hopkins University (JHU) CSSE COVID-19 Dataset，裡面收集大量官方組織的與 COVID-19 有關的資料。

2 資料內容

資料集中包含各個國家及地區每天的資料。一筆資料中有某地區某天的關於 covid_19 的感染人口數，回復人口數及死亡人口數以及該地區的經緯度。藉由分析資料來比較不同國家在疫情下的防疫情形。

Province_State	Country_Region	Last_Update	Lat	Long_	Confirmed	Deaths	Recovered	Active	Combined_Key	Incidence_Rate	Case-Fatality_Ratio
Tyumen Oblast	Russia	2020-10-17 04:24:12	58.820649	70.365884	11275	51	8025	3199.0	Tyumen Oblast, Russia	305.356949	0.452328

3 資料彙整

第一步，先將所有的資料讀入，整理成一張有著當下所有天數，地區和確診、回復、死亡人數的資料。由於原資料中有包含一些地理資訊，這不會用在我們實際訓練的模型中，我們會將其過濾掉，只保留與人口相關的資料。
接著我們從中取一部分國家做我們的訓練集，先從部分國家開始分析。這裡實際例子，先從部分歐洲國家開始。

4 帶入模型

第二步，套用 SIRD 模型，找出可以吻合資料分布的參數。SIRD 是一個基本的描述傳染病在封閉環境下造成疑似(susceptible)、感染(infected)、回復(recovered)和死亡(deceased)人口變化的模型。這個模型由下方的四個常微分方程描述，各項人數隨時間的變化。另外參數也是隨時間變化，意思是說造成每日的人口變化會因應該地區或國家對於疫情的應對有所不同。不同國家因應 COVID-19 的方式應該不會完全相同，藉由分析不同國家模型參數的相異或相同之處，得知各國抗疫作為是否有效和 COVID-19 在不同地區傳播是否有共同性。

$$\dot{S} = -\beta \frac{I}{N} S$$

$$\dot{I} = \beta \frac{I}{N} S - (\mu + \gamma) I$$

$$\dot{R} = \gamma I$$

$$\dot{D} = \mu I$$

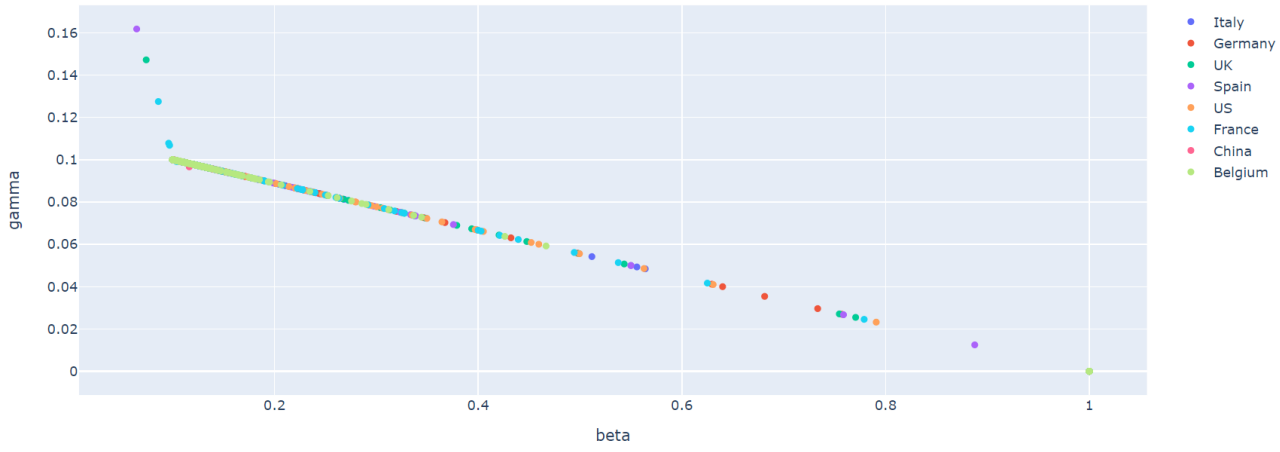
S 是疑似確診人數，I 是感染人數，R 是回復人數，D 是死亡人數。N 是總人數 β 是感染率、 γ 是回復率、 μ 是死亡率。

5 實際計算

離散化SIRD模型，使用有限差分法，可以得到下一天的人口數，在方程式中已有今日和明日的人口數，未知的只有參數(感染率、回復率、死亡率)，於是現在我們要做的是一個 data fitting的問題。意思是說，藉由適當的loss function，找出最佳的參數，使得在離散化SIRD模型結果跟實際資料吻合。這裡的loss function 使用文章First-principles machine learning modelling of COVID-19中設計的loss function。藉由最佳化程式套件，求出每日的參數。

6 分析呈現

將訓練實用到的國家的結果繪製在同一張圖上，找出資料的相關程度，觀察不同地區參數隨日期的變化。



結果發現感染率和回復率呈負相關，感染越嚴重，回復的情形就不好。高感染率主要發生在初期，部分零星分布在後續天數，後期感染率變化不大，因為總感染人數遠比初期多，些微的感染率就會帶來眾多的感染人口。當高感染率的時期回復率不會高，表示該疾病本身在圖表的國家中，醫療能力和個人自癒力無法應對病毒。這符合 COVID-19 是現今人類難以治癒的疾病的事實。基於難以治癒的事實，使得感染者會長期，存在於一個群體之中，隨時間增加接觸他人使新感染者增加的風險提高，造成一個地區會陷入長期有大量染疫者的局面。這表示至現今的局面而言，提高回復率，意指增強醫療能力尚未有明確且有效

的做法，要能控制疫情的做法在於降低感染率，竟可能的減少與他人接觸，減緩染疫的人口，用現有的醫療資源，盡可能的治癒已感染者。避免更大規模的感染發生。

預測

藉由資料分析得知 COVID-19 的傳播在不同國家間也著共同的特性，我們試想用手上現有部分國家資料預測後面幾天的感染和死亡人數。這裡我們試著訓練一個模型可以預測不同地區。

1 訓練方法

使用前幾天的資訊預測幾天後的結果。例如: 用含今天在內的前 30 天的死亡和感染人數做輸入，預測明天的死亡和感染人數。藉由不同的方法，使的預測結果吻合手上的資料。在實作中，採用 linear regression 和 XGboost 兩個方法，並比較這兩個方法的差別。將這兩個方法應用在後述兩項實驗中。這裡做兩種類型的預測。第一種，用含今天在內的前 30 天的死亡和感染人數做輸入，預測明天的死亡和感染人數。第二種，用含今天在內的前 30 天的死亡和感染人數做輸入，預測後 10 天的死亡和感染人數。

2.linear regression

線性回歸方法指藉由調整模型的參數使得，預測值跟手頭上現有資料的差距最小化。例如: 給 N 筆資料，資料點用x_i表示，標籤(label)用 y_i表示。一筆訓練資料為(x_i, y_i)。模型$\mathrm{Y}(x_i) = Wx_i + b$，找到參數(W, b)使得 loss function

$$\sum_{i=1}^{N} (Y(x_i) - y_i)^2$$

得到最小值。

線性回歸方法重點在於，藉由一條直線或超平面使資料點到此線或平面的距離最短。使預測的整體誤差可以最小。

3.XGboost

XGboost 方法是建立在 Tree ensemble 的模型。將手上資料分配到 tree 的節點上，每個節點上都有分數，找到適當的參數使得手頭資料可以正確分類。上述是一個 tree 的情形，將多個分辨不同特徵的 tree 組合起來，使得訓練資料可以正確分對。這個模型要最佳化較為困難，這裡採用 addictive training 的方法，藉由增加一個 tree 的結構修正原模型的不足，使模型的到最好的結果。

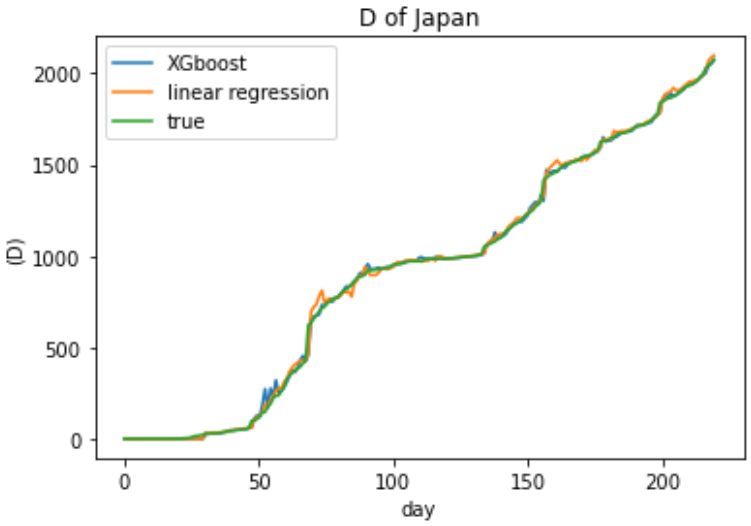
XGboost 方法目前被廣泛應用在機器學習領域和資料探勘的競賽中，需多競賽的最佳解方中都有採用 XGboost 方法。說明這個方法在廣泛的問題上應該可以

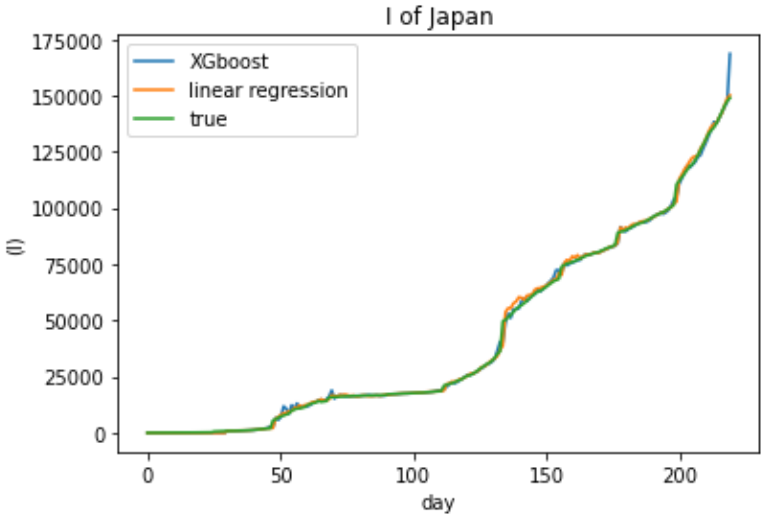
給予優秀的結果。

4 實驗結果

(a) 用含今天在內的前 30 天的死亡和感染人數做輸入，預測明天的死亡和感染人數。

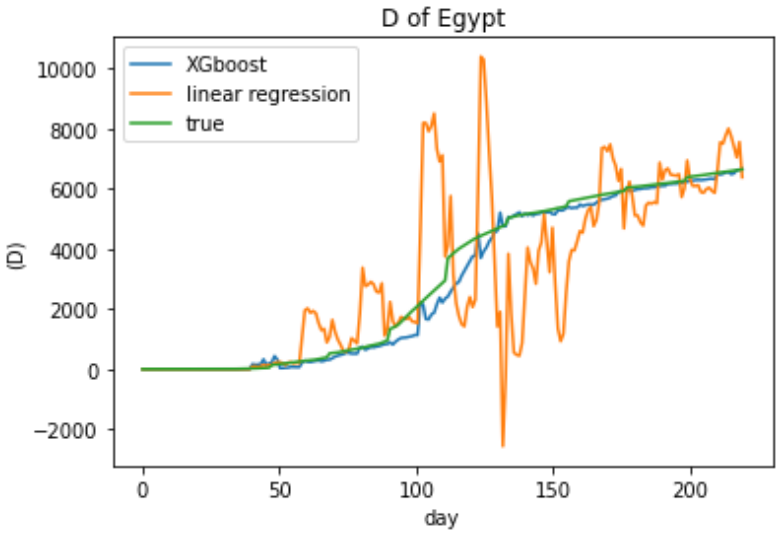
下方第一張圖為日本的死亡人數預測，而第二張圖為日本的感染人數預測。
約第 150 天開始為預測結果的圖形比較。

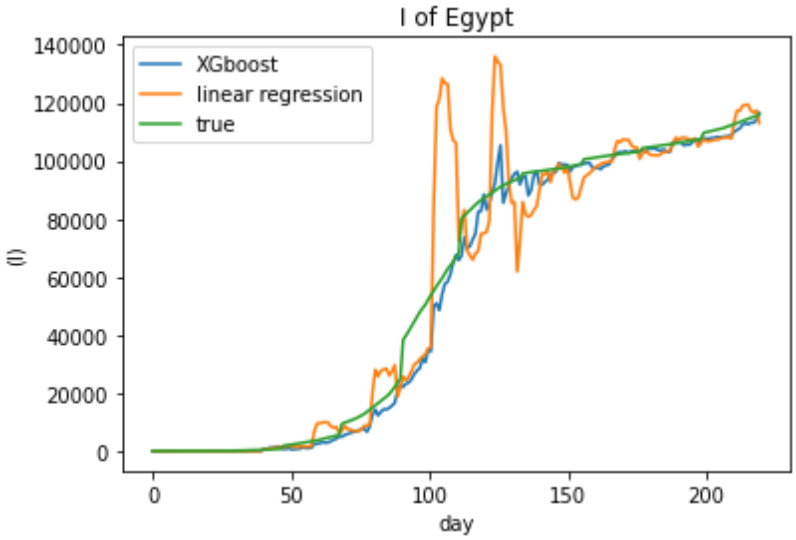




(b) 用含今天在內的前 30 天的死亡和感染人數做輸入，預測 10 天後的死亡和感染人數。

下方第一張圖為埃及的死亡人數預測，而第二張圖為埃及的感染人數預測。
約第 150 天開始為預測結果的圖形比較。





在預測明日死亡和感染人數的實驗，linear regression 的表現較 XGboost 好，兩者就圖形上而言，預測結果跟實際情形非常接近。
然而，在第二個實驗中，XGboost 效果較好，比較吻合實際結果。XGboost 的方法較適合用來預測據現在較遠的日子的死亡和感染人數。

討論

用現有的 30 天資料預測後面某天感染人數和死亡人數，會因為據預測時間點越久，誤差情形增加。藉由觀察不同國家的預測人數圖，從疫情爆發至現在時間點，感染人數和死亡人數有上升趨勢，意味著這個疫情並未有明顯趨緩或可能又再度爆發，目前大部分國家都仍深深受到 Covid-19 的影響。