社會科學/公共衛生領域之機器學習應用

鄭守開(Shou-Kai Cheng) [1]

[1] 國立臺灣大學; tw841117@gmail.com

摘要: 近年來，在機器學習、人工智慧等技術的逐漸普及，相關技術被嘗試應用在各領域上，當中作者最為感興趣的便是在資訊科學工具在社會科學、公共衛生領域的應用。作者過去曾於臺灣人口學會、中央研究院歐美研究所從事跨國人口社會經濟地位、健康狀況之研究，因此本文使用人口學、醫療相關的資料庫進行分析，資料來源包括：世界銀行(World Bank)、聯合國人類發展報告(Human Develoment Reports)、中華民國內政部統計處、2021 李宏毅教授機器學習課程內容等。本文包括(0)資料介紹。(1)基本資料統計分析。(2)資料視覺化呈現。(3)機器學習預測。(4)機器學習應用討論。

關鍵字: 生育率、人均收入、教育年限、性別差異、社會經濟地位、傳染病預測、公共衛生、新冠病毒(COVID-19)、機器學習。

0. 資料介紹

本文使用的資料來源主要使用 2013 年世界銀行(World Bank)、2013 年聯合國人類發展報告(Human Development Reports)中的各國資料，從中擷取出平均每位婦女總生育率(fertility rate, total (births per woman))、男性及女性的平均教育年限(Mean years of schooling (years)、人均國內生產毛額(GDP per capita (current US$))，此外因為這兩粉報告當中皆沒有包括臺灣的相關資料，因此再從中華民國內政部統計處另外獲得資料整合進資料原始檔案當中。共有兩份資料來源檔案(Fertility Rate Per Woman and Male Education Relationship Data 2013.xlsx、Fertility Rate Per Woman and Female Education Relationship Data 2013.xlsx)，各自檔案當中每位婦女總生育率、人均國內生產毛額皆相同，但男性及女性的平均教育年限則不同，下文中將稱為男性資料檔案及女性資料檔案，檔案讀取後的摘要如下(表一、表二)。

原始檔案讀取摘要(上：表一、下：表二)。

	Country Name	Fertility rate, total (births per woman) 2013	Mean years of schooling, female (years) 2013	GDP per capita (current US$) 2013
0	Afghanistan	5.359	1.4	637.165523
1	Albania	1.690	9.8	4413.060861
2	Algeria	2.990	6.0	5499.581487
3	Argentina	2.322	10.5	13080.254732
4	Armenia	1.728	11.5	3838.185801
...	...	...	...	...
163	Vietnam	1.978	7.5	1886.671896
164	Yemen, Rep.	4.326	1.7	1607.152365
165	Zambia	5.132	6.2	1878.907001
166	Zimbabwe	4.030	7.3	1430.000818
167	Taiwan	1.070	11.2	21973.000000

	Country Name	Fertility rate, total (births per woman) 2013	Mean years of schooling, male (years) 2013	GDP per capita (current US$) 2013
0	Afghanistan	5.359	5.1	637.165523
1	Albania	1.690	9.5	4413.060861
2	Algeria	2.990	7.8	5499.581487
3	Argentina	2.322	9.6	13080.254732
4	Armenia	1.728	10.8	3838.185801
...	...	...	...	...
159	Vietnam	1.978	5.7	1886.671896
160	Yemen, Rep.	4.326	3.8	1607.152365
161	Zambia	5.132	7.2	1878.907001
162	Zimbabwe	4.030	7.8	1430.000818
163	Taiwan	1.070	12.1	21973.000000

1. 基本資料統計分析

男性資料檔案及女性資料檔案中各有 168 及 164 國家的資料，基本的描述統計資料如下(表三、表四)，可以觀察到男性及女性之間平均教育年限有較明顯的差異(男性 8.422561 對比女性 7.998810)。

基礎描述統計資料(上：男性資料檔案、下：女性資料檔案)

```
mean_df = df.mean()
print(mean_df)
```

```
Fertility rate, total (births per woman) 2013        2.798765
Mean years of schooling, male (years) 2013            8.422561
GDP per capita (current US$) 2013                 15059.599438
dtype: float64
```

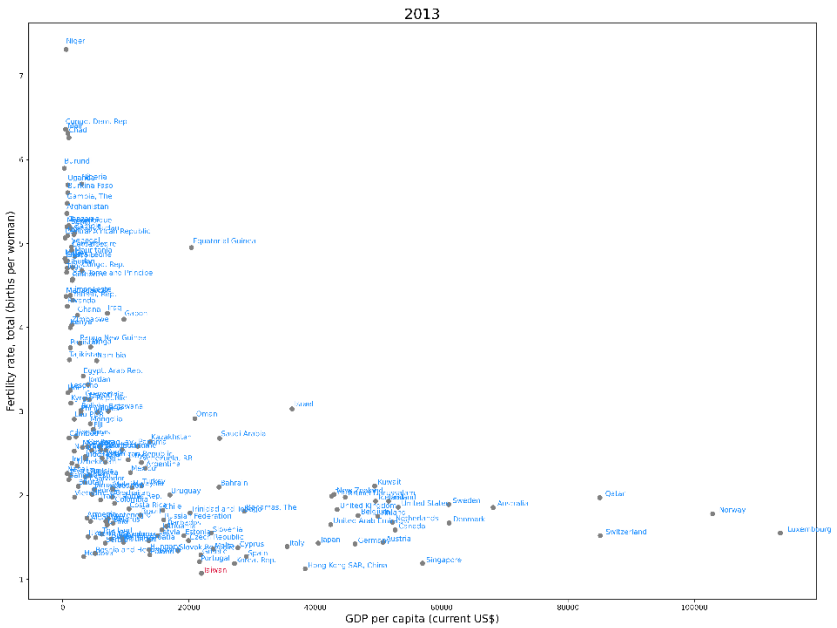
```
mean_df = df.mean()
print(mean_df)
```

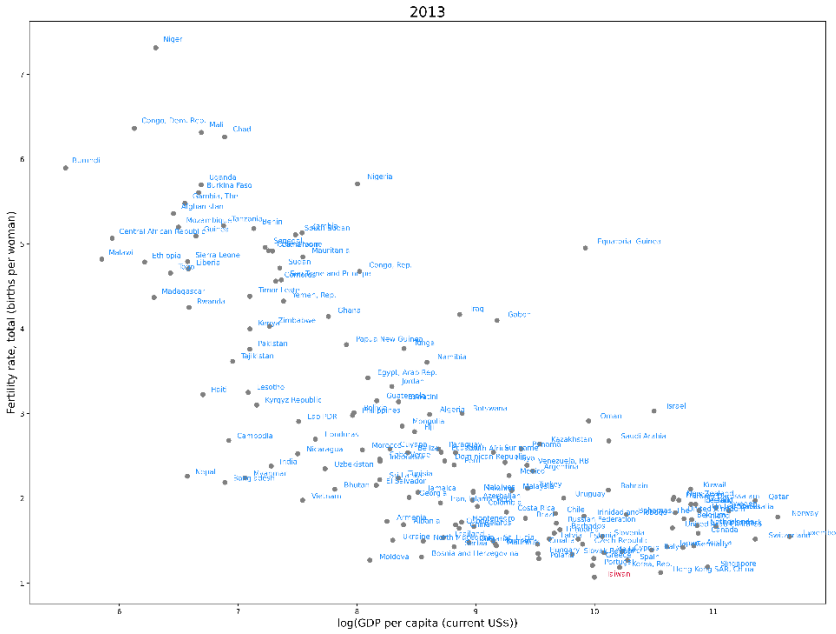
```
Fertility rate, total (births per woman) 2013        2.789872
Mean years of schooling, female (years) 2013          7.998810
GDP per capita (current US$) 2013                 14962.667247
dtype: float64
```

2. 資料視覺化呈現

為了能夠清楚的呈現資料的分佈情況，採用散佈圖來呈現兩兩一組資料之間的分佈情況，首先使用人均國內生產毛額(GDP per capita (current US$))、平均每位婦女總生育率(fertility rate, total (births per woman))來繪製散佈圖(圖一)，但因為相比於生育率的單位(介於 1~7 區間內)，生產毛額的單位過大導致散佈圖的分佈偏向於圖片左側，因此對人均國內生產毛額取自然對數，得到人 log(均國內生產毛額)(log(GDP per capita (current US$))，再次繪製新圖(圖二)，可以觀察到兩者之間呈現負相關的趨勢，人均國內生產毛額越高，平均每位婦女總生育率越低，而臺灣也屬於高人均國內生產毛額、低平均每位婦女總生育率的右下角部分。(註：臺灣資料使用紅色字體顯示。)

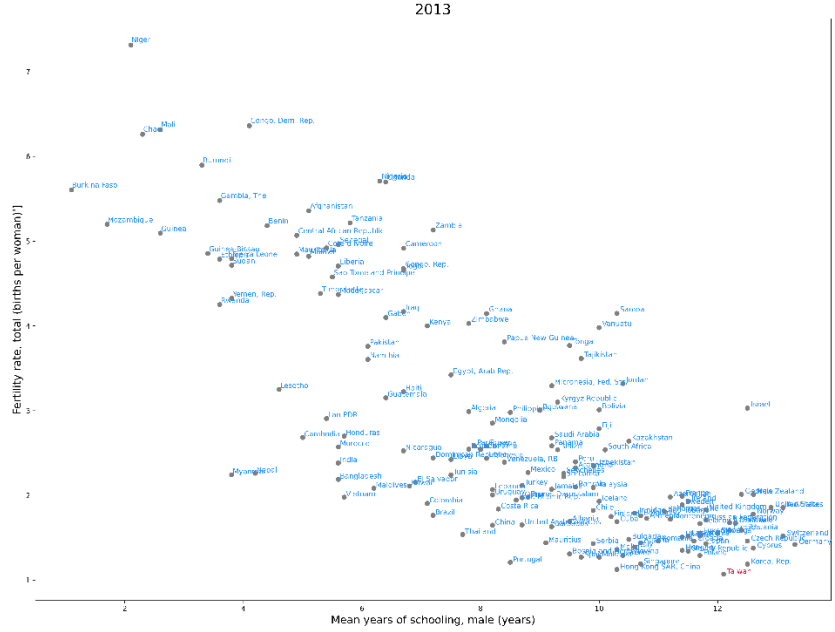
人均國內生產毛額、平均每位婦女總生育率散佈圖(上：圖一、下：圖二)

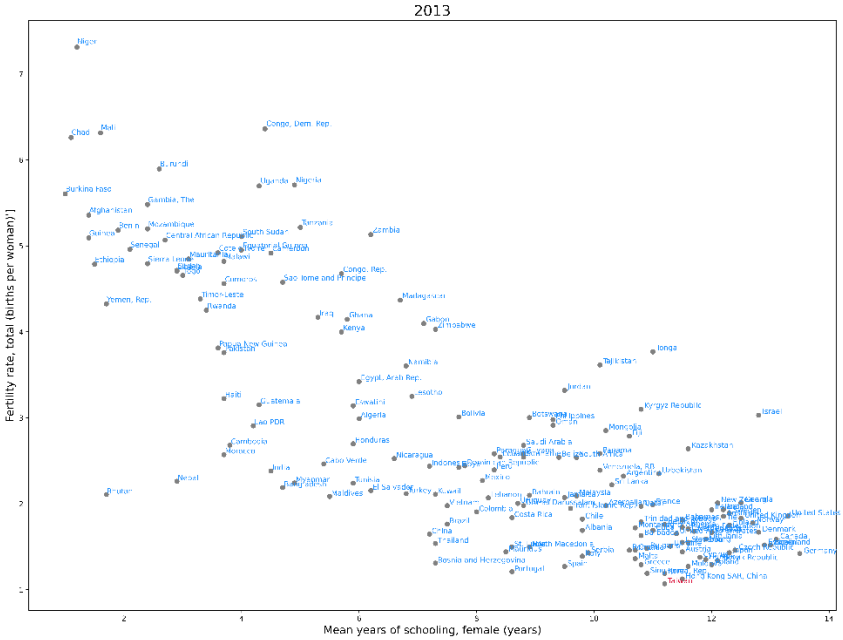




接下來使用男性及女性的平均教育年限(Mean years of schooling (years)、平均每位婦女總生育率(fertility rate, total (births per woman))來繪製散佈圖(圖三：男性資料檔、圖四：女性資料檔)，因為兩者的資料單位相近(皆為 1~15 區間內)，因此散佈圖相當清楚，不須再使用自然對數處理資料。整體來說，可以觀察到兩者之間呈現負相關的趨勢，平均教育年限越高，平均每位婦女總生育率越低，且女性的教育年限相較於男性整體分佈更偏左側，顯示有些國家的女性教育年限相比男性來說較低。而臺灣在此圖當中也屬於高平均教育年限、低平均每位婦女總生育率的右下角部分，值得注意的是臺灣的男性教育年限(約 12 左右)也同樣比女性(約 11 左右)較長。(註：臺灣資料使用紅色字體顯示。)

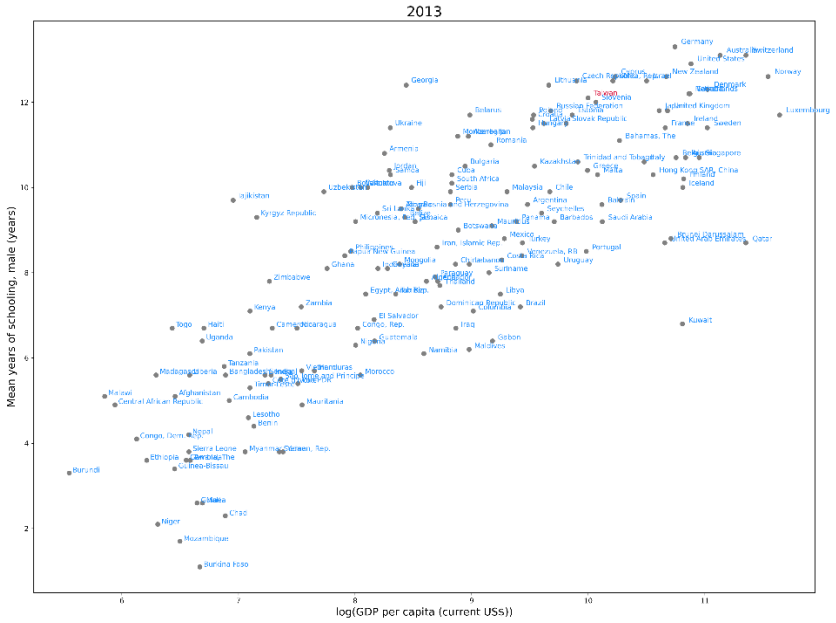
平均教育年限、平均每位婦女總生育率散佈圖
(上：圖三 男性資料檔、下：圖四 女性資料檔)

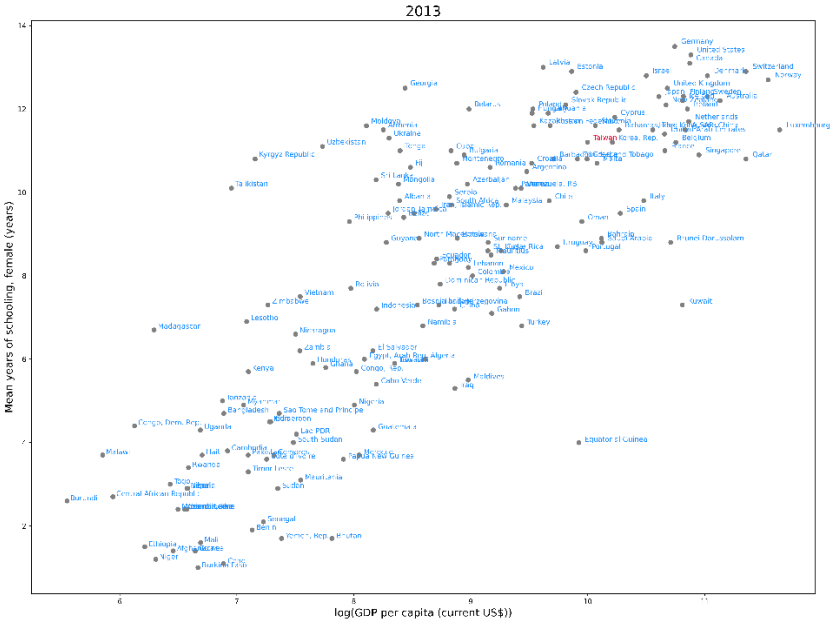




接下來使用男性及女性的平均教育年限(Mean years of schooling (years)、人均國內生產毛額取自然對數 log(均國內生產毛額)(log(GDP per capita (current US$))來繪製散佈圖(圖五:男性資料檔、圖六:女性資料檔)，平均教育年限及人均國內生產毛額取自然對數的資料單位相近(皆為 1~15 區間內)，因此散佈圖相當清楚，可以觀察到兩者之間呈現正相關的趨勢，平均教育年限越高，人均國內生產毛額取自然對數也越高。而臺灣在此圖當中也屬於高平均教育年限、高人均國內生產毛額取自然對數的右上角部分。(註：臺灣資料使用紅色字體顯示。)

平均教育年限、log(均國內生產毛額)散佈圖(上：圖五、下：圖六)





3. 機器學習預測

首先使用使用男性及女性的平均教育年限(Mean years of schooling (years)作為自變數，使用機器學習預測應變數人均國內生產毛額(均國內生產毛額)(log(GDP per capita (current US$))，使用的優化器為 Stochastic Gradient Descent(SGD)，進行 500 次的訓練，最後得到如下結果(表三：男性資料檔、表四：女性資料檔)。可以看到統過使用男性的平均教育年限來預測的人均國內生產毛額為$17507.783(US)，使用女性的平均教育年限來預測的人均國內生產毛額為$23760.86(US)。

機器學習預測人均國內生產毛額結果(左：圖三 男性資料檔、右：圖四 女性資料檔)

```
1 print(model.predict([10.0]))
```

```
[[17507.783]]
```

```
1 print(model.predict([10.0]))
```

```
[[23760.86]]
```

4. 機器學習應用討論

使用機器學習的技術的確可以在有足夠資料量的情況下進行有效率的預測，隨著計算機軟、硬體的進步，機器學習的實作越來越普及，但在進行的同時也有些問題讓我開始思考，此文當中使用機器學習來進行預測，的確的到具體的數字，但每次的預測因為包含有非決定性運匱，導致每次的結果都有所不同，可能也無法得知如何設置參數才能得到最佳結果，而在訓練次數上的設定也難有一定的標準，雖然大量的訓練次數往往能得到較好的結果，但若是過多的訓練次數可能也導致過度貼合(overfitting)，導致訓練出來的模型效果不佳。而除了機器學習技術上的問題，另外讓我感到難以解釋的便是如何找到學術上的佐證來支持得到的結果，雖然能夠使用調整參數來得到更佳的模型，但為何採用這樣的參數依然有學理上的疑慮。

另外一個案例便是在新冠疫情病毒傳染預測時遇到的問題(參考 COVID-19 Cases Prediction.ipynb)，而來源資料是 2020 年 4 月 Facebook 上的 3 天調查資料(包括：是否前往餐廳、使用大眾運輸工具)，應用神經網路(Neural Network)模型輸入前兩天的調查資料來進行訓練，並用前兩天的新冠病毒感染結果測試資料來驗證，最後輸入第三天的資料來預測特定受試者否會得到新冠病毒，再使用第三天的測試資料去驗證模型的準確率。在參數調整的部分同樣有與上面同樣的疑慮，雖然能夠得到較好的預測結果，但為何要使用這樣的參數設定也還是難得到明確答案。另外有些參數相當直觀與新冠病毒有相關性(如：搭乘大眾運輸工具、前往餐廳用餐等)，直接放入模型納入參考相當易懂，但有些較不直觀的資料被

加入模型時卻有更好的測試結果，這時便難尋求解釋，也是我在進行機器學習實作時思考的問題，未來累積更多經驗後希望能獲得答案。

參考資料

1. World Bank. https://data.worldbank.org/indicator/SP.DYN.CBRT.IN?end=2013&start =2002
2. UN Development Programme. http://hdr.undp.org/en/en/content/human-development-report-2013
3. 中華民國內政部統計處. https://www.moi.gov.tw/cp.aspx?n=5590
4. 李弘毅教授機器學習課程(2021)