應用深度學習預測 YouBike 每月借用次數
-以台北市為例

吳玉文 Yu-Wen Wu

摘要

環保意識抬頭、大眾運輸逐漸完善等原因，使得公共自行車需求逐漸增加，為預測未來的需求以進行因應，本實驗使用深度學習進行迴歸分析，並透過 F 檢定找出適合預測自行車租借次數所使用的自變數，探討各相關變數(氣候、私人運具等)對於自行車租借次數的影響。

關鍵詞：深度學習、迴歸分析、公共自行車

一、研究動機

近年環保意識逐漸抬頭，越來越多人開始重視燃油車的替代方案，除了近年廣受討論的電動車議題外，因為 COVID-19 疫情的緣故，引發了一波騎乘自行車的熱潮，加上各國大眾運輸系統逐漸完善，如何去完善「第一哩路」及「最後一哩路」，公共自行車扮演了相當重要的角色。

而面對逐漸攀升的需求，國內最主要的公共自行車租賃系統營運商「微笑單車股份有限公司」積極的與各縣市政府進行合作，更是於 2020 年推出了 YouBike 2.0 系統，隨著使用者人數以及自行車站點數逐漸上升，能夠準確的預測需求對於「微笑單車股份有限公司」會是相當重要的課題，可以更有彈性的對自行車進行調度以及衡量未來的營運情形。

二、文獻回顧

臺北市政府交通局統計室在<<大數據分析臺北市公共自行車使用特性>>中歸納出了使用者的十大使用特性，其中包含了租借行為易受天氣晴雨影響、假日租借時間較平日長等。

而在<<公共自行車租借量之影響因素分析–地理加權迴歸和函數資料分析方法之應用>>中則提到了租借次數與傳統市場面積、商業區面積、租借站車位數、每千人擁有汽機車數具有一定的關係。

三、研究方法

1. 相關變數考量

較易收集的資料皆以月別進行記錄，因此本研究之資料皆以月做為單位，而在參考文獻所述後，篩選出較適合進行分析的因子，資料中將以下相關變數(氣候、自行車、私人運具、平假日)列入以進行迴歸分析。

氣候相關變數：

1. 每月總降雨量(毫米)
2. 每月均溫(攝氏)
3. 每月總日照時數(小時)
4. 每月總降雨天數(天)

自行車相關變數：

5. 公共自行車租借站點數(站)
6. 公共自行車車輛數(輛)
7. 已建設之自行車道長度(公里)

私人運具相關變數：

8. 每千人擁有汽車數(輛)
9. 每千人擁有機車數(輛)

平假日相關變數(同月份放假天數較為接近)：

10. 前一年同月份之租借次數(百萬次)

2. F 檢定

將以上 10 項相關變數分別與當月租借次數(百萬次)進行 F 檢定，查看與預測目標之間的相關性，以進行特徵選擇，結果如下：

Variables	Score
rainfall(mm)	2.534780
num_of_rainy_day	1.311336
num_of_bikes	0.476641
insolation_duration(hour)	0.263690
last_12_months	0.221171
total_bike_lane_length(km)	0.175166
monthly_average_temperature(Celsius)	0.144311
num_of_bike_stations	0.112669
num_of_scooters_owned_per_thousand_people	0.070508
num_of_cars_owned_per_thousand_people	0.006026

從上圖可以得知，10 項變數與預測目標之間的相關性由強到弱依序為：降雨量、降雨天數、自行車車輛數、日照時數、前一年同月份之租借次數、已建設之自行車道長度、月均溫、租借站點數、每千人擁有機車數、每千人擁有汽車數。

3. 深度學習應用及相關變數設定

目標是透過選擇不同數量的相關變數進行特徵提取，來找出最適合的模型及其中考慮的相關變數，進行 10000 個 Epoch(若經過 1000 個 Epoch 訓練結果沒有變好則提早停止)、Learning Rate 為 0.00001，使用 MSE(Mean-Square Error)定義 Loss，以 AdamW 做為 optimizer。

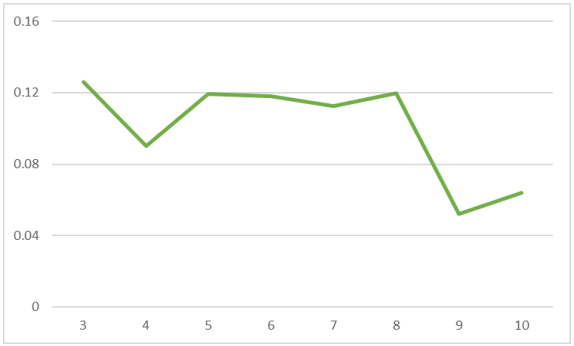
訓練資料從 2019 年 1 月至 2021 年 12 月，共計 36 組，從中取出 7 組做為驗證集，其餘 29 組做為訓練集，測試資料則為 2019 年 1 月至 2022 年 5 月，共計 41 組。每個提取不同特徵數量的模型會在測試資料進行預測，將預測結果與實際結果計算均方誤差，均方誤差最小者為最適模型，再以最適模型為基礎，觀察其中各相關變數的變動對於預測目標(自行車租借次數)的影響。

四、執行結果

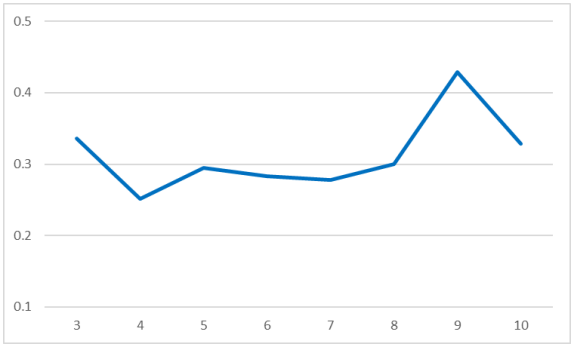
1. 以不同相關變數總數進行預測所得之誤差

將測試資料(2019 年 1 月至 2022 年 5 月)拆分為兩個部份：(1)在已知資料上的預測(2019 年 1 月至 2021 年 12 月)；(2)在完全未知資料下的預測(2022 年 1 月至 2022 年 5 月)

以下為不同相關變數總數在(1)部份表現之情形，其中 x 軸為變數選取數目(依相關性強弱，如選取 4 個則為降雨量、降雨天數、自行車車輛數、日照時數)，y 軸則為均方誤差：



而下圖則為相關變數總數在(2)部份表現之情形，x 軸為變數選取數目，y 軸為均方誤差：

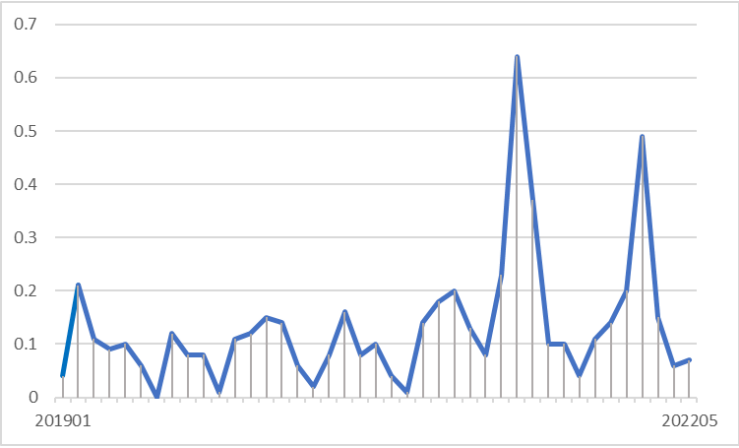


從上面兩張圖中可以看出，如果能參考進更多的相關變數，大致上是能夠使得模型在已知資料的解釋能力更好，但卻會在未知資料上的預測變差，由於實驗目的是為了找出適合預測未來的模型，因此對(1)、(2)兩部份進行加權，並將重點著重於第(2)部份，以選出最適合的模型，採用「第(1)部份的誤差*0.3+第(2)部份的誤差*0.7」的形式，結果如下：

相關變數 3 項：0.1259*0.3 + 0.3359*0.7 = 0.27290
相關變數 4 項：0.0902*0.3 + 0.2511*0.7 = 0.20283(最佳)
相關變數 5 項：0.1192*0.3 + 0.2942*0.7 = 0.24170
相關變數 6 項：0.1180*0.3 + 0.2832*0.7 = 0.23364
相關變數 7 項：0.1126*0.3 + 0.2774*0.7 = 0.22796
相關變數 8 項：0.1197*0.3 + 0.2996*0.7 = 0.24563
相關變數 9 項：0.0520*0.3 + 0.4288*0.7 = 0.31576
相關變數 10 項：0.0641*0.3 + 0.3284*0.7 = 0.24911

就以上結果呈現，選擇相關變數總數 4 項(包含降雨量、降雨天數、自行車車輛數、日照時數)的模型，在舊有資料上有著足夠的解釋能力，在未知資料上表現也是最好的，因此選定此模型以進行後續實驗。

2. 模型的誤差



上圖為相關變數 4 項的模型(以下簡稱最適模型)的誤差情形，以下為其數據：

第(1)部份：誤差平均為 12%；中位數為 10%；標準差為 11%

第(2)部份：誤差平均為 19%；中位數為 15%；標準差為 16%

全部資料：誤差平均為 13%；中位數為 10%；標準差為 12%

其中較為異常的值為 2021 年的 5、6、7 月(誤差 23%、64%、37%)，推測為疫情緣故(2021 年 5 月 15 日發布全國疫情第三級警戒)以及 2022 年的 1、2 月(誤差 20%、49%)，同樣應為疫情而導致(2022 年 1 月開始，新冠病毒 Omicron 變異株導致大規模感染，而後疫情指揮中心宣布 2022 年 3 月起開始經濟防疫新模式)。

排除掉以上提及之異常值後，以下為最適模型的統計數據：

第(1)部份：誤差平均為 10%；中位數為 10%；標準差為 5%

第(2)部份：誤差平均為 9%；中位數為 7%；標準差為 4%

全部資料：誤差平均為 10%；中位數為 10%；標準差為 5%

另外，最適模型存在著每年 2 月都會出現較大誤差率的情形(2019 年：21%、2020 年：15%、2021 年：20%、2022 年：49%)，推測是因為年節的關係。

3. 相關變數影響：

在固定其餘變數下，利用相同模型計算以得出每個變數對於預測結果的影響，結果如下：

(1) 每月總降雨量(毫米)：約略為每增加 25mm，減少 8.48 萬次租借次數

(2) 每月總降雨天數(天)：約略為每增加 1 天，減少 4.45 千次租借次數

(3) 公共自行車車輛數(輛)：約略為每增加 500 輛，增加 4.74 萬次租借次數

(4) 每月總日照時數(小時)：約略為每增加 10 小時，增加 8.35 千次租借次數

可以整理出簡化過後的公式如下：

$$f(x_1,x_2,x_3,x_4) = -0.34x_1 - 0.445x_2 + 0.00948x_3 + 0.0835x_4 + b(\text{常數項})$$

其中 $f(x_1,x_2,x_3,x_4)$為租借次數(萬次)；x_1 為每月總降雨量(毫米)；x_2 為每月總降雨天數(天)；x_3 為公共自行車車輛數(輛)；x_4 為每月總日照時數(小時)。將各項變數在資料中的平均值代入(總降雨量=174mm；降雨天數=12 天；公共自行車=15325 輛；日照時數=115 小時)，得到 b=133.5869，最終式子如下：

$$f(x_1,x_2,x_3,x_4) = -0.34x_1 - 0.445x_2 + 0.00948x_3 + 0.0835x_4 + 133.5869$$

將 2019 年 1 月至 2022 年 5 月之資料代入此式，所得誤差平均為 20%；中位數為 15%；標準差為 16%。將前述因疫情影響之異常值剔除後，所得誤差平均為 17%；中位數為 13%；標準差為 14%。

五、研究不足之處

1. 天氣的難以預測性

本次實驗主要採用之相關變數包含了天氣因子，在實際情況下不會是先得到準確的月總雨量才預測總租借次數，而是需要先對天氣本身進行預測，但天氣本身較為難以預測，因此會放大預測總租借次數的誤差，此為本實驗之不足之處，如果能降低預測的時間範圍(月→日或小時)，就能縮小這樣的缺點。

2. 不同時間單位下的預測

本研究因為以日為單位及以小時為單位的資料難以蒐集的緣故，因此選擇了以月為單位進行預測，但如果可以從較短時間範圍下進行預測，在資料量增多、能考慮更多的相關變數(如：尖離峰)的情況下，結果勢必會更為準確，也能讓相關單位較快速的進行反應及規劃。

3. 不同區域大小下的預測

本研究同樣因為資料蒐集難易程度的緣故，因此是以整個台北市為單位，如果能以各區(大安區、文山區等)或是以站點做為單位，可能可以考慮進更多的相關變數(如社會經濟相關變數-與學校的距離、傳統市場的面積等等)，或許能讓誤差降到更低，同樣能夠增加實用性。

4. 不同縣市的比較

考慮到台灣北部與南部的氣候分布具有差異，縣市之間也有著不同的特性，因此有可能本實驗結果無法完全適用於其他縣市的營運情形，但其他縣市目前資料仍不夠完善，所以暫時無法進行比較。

六、結論

本實驗顯示出公共自行車租借次數以月為單位而言，主要會受到天氣因素及車輛數影響，透過以上的結果，也顯示出了深度學習確實是能幫助進行租借次數的預測的，所得數據及公式也能提供除台北市外的縣市參考，但就誤差值看來，仍有相當大的進步空間，如果有足夠的資料能改善研究不足之處，誤差值就能被縮小，供公共自行車營運商進行更好的利用(如安排人力、車輛等等)。

七、參考資料

政府開放資料平台-臺北市氣候按月別
政府開放資料平台-臺北市機動車輛登記數
政府開放資料平台-臺北市公共自行車概況按月別
臺北市政府交通局統計室<<大數據分析臺北市公共自行車使用特性>>
交大運管所-倪如霖<<公共自行車租借量之影響因素分析-地理加權迴歸和函數資料分析方法之應用>>
程式代碼參考台灣大學李弘毅教授 2022 機器學習課程-作業一