國立雲林科技大學

機器學習專案作業一

神經網路預測類別數值

成員：

M11021009 吳宥霆

M11021028 劉軒瑋

M11021035 黃堉豪

M11021052 邱守燦

指導教授：

許中川 教授

日期：

2022 年 4 月 21 日

摘要

隨著時代的演進,使用傳統統計進行分類預測的方法,已逐漸被機器學習技術取代;使得在分類預測上更加準確且快速。本研究使用前饋式神經網路，分別針對 Adult 資料集的屬性資料預測 hours-per-week 數值；以及利用 Bike Sharing 資料集的總租借數資料來預測 season 類別。並經由不同的參數調整與模型績效指標找出最佳的模型績效，選擇較優之模型投入實際應用。而經研究發現，Adult 資料集與 Bike Sharing 資料集，當深度學習的優化器為 adam 時績效為最佳。

關鍵字：前饋式神經網路、深度學習優化器、模型績效指標。

一、 緒論

1.1 動機

1.1.1 Adult Dataset

根據勞動部國際勞動統計，台灣去年就業者平均每年工時為 2021 小時，在 40 個主要國家中排名第 4 名，第一名則是新加坡。工時長短一直是勞動階級所關心的問題，同時也是展現該國家工作文化的一種指標；如若能投入一個人相關資訊，例如職業、收入、國籍等，便能預測出該人之工時長短，從而了解該國整體之工作文化。

1.1.2 Bike Sharing Dataset

共享單車是一新的經濟模式，使用者從租賃到歸還的流程，通通可藉由系統操作來完成。隨著人口的增加，交通工具所造成的阻塞與汙染日益嚴重，共享單車租賃也是解決空氣汙染與交通阻塞的方法之一，此經濟模式也提升了人們對低碳環保的生活意識，許多政府也鼓勵人們使用單車與大眾運輸來做為日常通勤之工具，減少開車所造成的碳排放，達成美化生活環境與品質，以及減少碳排放的目標。

1.2 研究目的

1.2.1 Adult Dataset

為了預測出該人之工時長短，從而了解該國整體之工作文化，本組選取 Adult 資料集，並使用前饋神經網路進行預測，投入工作類別、婚姻狀況、教育程度、職業、所屬國籍等 15 種屬性資料，並透過測試資料集驗證上述幾種模型的績效指標(MAE、RMSE、MAPE)，並根據績效指標進行比較，從中選擇較優之模型投入實際使用情境。

1.2.2 Bike Sharing Dataset

隨著資訊系統的發達，能使原本傳統單車複雜的租賃流程，透過共享單車系統讓流程自動化，讓用戶可以輕鬆地進行使用，本組根據 Porto 大學所提供的 Bike Sharing 資料集，裡面包含季節、天氣資訊、周末…等屬性資料，採用前饋神經網路進行預測，並切割 80%的訓練資料集與 20%的測試資料集，透過模型的績效指標驗證，來預測總借車數輛。藉此希望可以提供該城市之各時段單車需求建議，使該公司能依據不同時段來配置相應的單車數量。

二、 資料集

2.1 資料集

2.1.1 Adult Dataset 說明

此資料集建立於 1996 年 5 月 1 日共有 48,842 筆，15 個欄位。

表 1 Adult Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
Age	年齡
Workclass	工作類別
Fnlwgt	連續數值
Education	教育程度
Education-num	教育人數(連續數值)
Marital-status	婚姻狀況
Occupation	職業
Relationship	關係
Race	種族
Sex	性別
Capital-gain	資本收益
Capital-loss	資本損失
Hours-per-week	小時/周
Native-country	所屬國家
Salary	年收入

2.1.2 Bike Sharing Dataset 說明

此資料集建立於 2013 年 12 月 20 日共有 17,389 筆，16 個欄位。

表 2 Bike Sharing Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
Instant	索引紀錄
Dteday	日期
Season	季節
Year	年
Month	月
Hr	小時
Holiday	假期
Weekday	周末
Workingday	工作日

Weathsit	氣象站天氣狀況
Temp	氣溫
Atemp	體感溫度
Hum	濕度
Windspeed	風速
Casual	臨時用戶
Registered	註冊用戶
cnt	總租借數量

三、 方法

3.1 程式架構

3.1.1 Adult Dataset

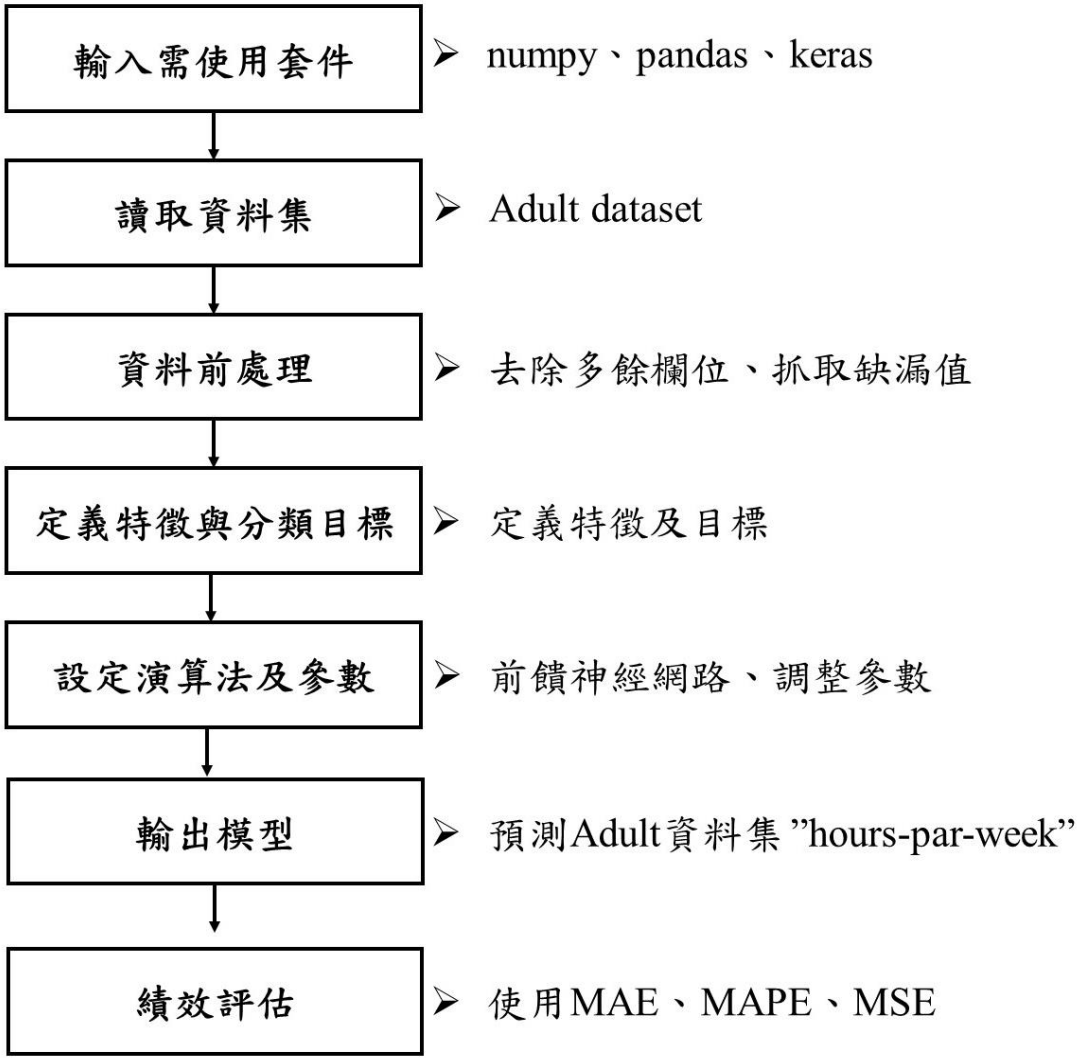

圖 1 Adult Dataset 之程式架構流程圖及說明

3.1.2 Bike Sharing Dataset

圖 2 Bike Sharing Dataset 之程式架構流程圖及說明

3.2 程式執行方法

3.2.1 前饋式神經網路（Feedforward Neural Network, FNN）

前饋式神經網路（Feedforward Neural Network, FNN）是一種前向傳遞類神經網路，包含了三層結構(輸入層、隱藏層及輸出層)，並利用「倒傳遞」的技術達到學習（model learning）的監督式學習，前饋視神經網路遵循人類神經系統原理，學習並進行數據預測，它首先學習，然後使用權重存儲數據，並使用算法來調整權重並減少訓練過程中的偏差，即實際值和預測值之間的誤差，主要優勢在於能夠快速解決複雜問題的能力。

四、 實驗

4.1 實驗數據

4.1.1 Adult Dataset 資料集

	age	workclass	education	…	…	Hours-per-week	Native-county	salary
0	39	5	9	…	…	40	38	0
1	50	4	9	…	…	13	38	0
2	38	2	11	…	…	40	38	0
3	53	2	1	…	…	40	38	0
…	…	…	…	…	…	…	…	…
32557	40	2	11	…	…	40	38	1
32558	58	2	11	…	…	40	38	0
32559	22	2	11	…	…	20	38	0
32560	52	3	11	…	…	40	38	1

圖 3 Adult Dataset 資料集

4.1.2 Bike Sharing 資料集

	instant	dteday	season	…	registered	cnt
0	1	2011-01-01	1	…	13	16
1	2	2011-01-01	1	…	32	40
2	3	2011-01-01	1	…	27	32
…	…	…	…	…	…	…
17377	17378	2012-12-31	1	…	48	61
17378	17379	2012-12-31	1	…	37	49

圖 4 Bike Sharing 資料集

4.2 前置處理

4.2.1 Adult Train Dataset

將 Adult Train 資料集，由原來的 txt 檔案格式轉換為 csv 檔案格式，發現在些許的資料欄位中具有不必要的符號資料，進行資料的清洗，利用 lambda 語法將具有'?'資料篩選與過濾，完成資料清洗，接著將 Adult Dataset 裡面文字屬性使用 Labelencoder 進行轉碼，最後讀取印出完成資料的前置處理。

4.2.2 Adult Test Dataset

將 Adult Train 資料集，由原來的 txt 檔案格式轉換為 csv 檔案格式，發現在些許的資料欄位中具有不必要的符號資料，進行資料的清洗，利用 lambda 語

法將具有'?'資料篩選與過濾，完成資料清洗，接著將 Adult Dataset 裡面文字屬性使用 Labelencoder 進行轉碼，最後讀取印出完成資料的前置處理。

4.2.3 Bike Sharing Dataet

將 Bike Sharing 資料集，利用 Pandas 套件的 DataFrame 物件讀取 csv 檔，刪除欄位，接者將 Bike Sharing Dataset 裡面文字屬性使用 Labelencoder 進行轉碼，最後讀取印出完成資料的前置處理。

4.3 實驗設計

4.3.1 Adult 資料集

(1)套件載入：載入協助資料分析的 Pandas 套件，並載入神經網路開發軟體 keras 的 Sequential 與 Dense 套件。

(2)正規化：讀取訓練（train_data）與測試（test_data）資料集，並個別將資料集取平均數與標準差，利用 Z-Score 標準化，將資料正規化。

(3)資料結構處理:將資料集中進行預測的欄位(hours-per-week)分離，並在模型訓練與模型績效評估中使用。

(4)建立神經網路模型：利用 keras 的 Sequential 與 Dense 套件撰寫前饋式神經網路，隱藏層（hidden layer）的激活函數以'Relu'、輸出層（output layer）的激活函數以'Sigmoid'為主，損失函數（loss function）利用 MSE 進行運算，最後使用迴歸問題的評估績效（MAE、MAPE、MSE）來檢測模型的成效。

(5)模型參數設定:在 adult 資料集隱藏層(hidden layer)設定 1~4 層、節點數(units)設定為 20，最佳化參數(optimizer)皆使用 adam，完成參數設定即可進行模型訓練。

(6)模型訓練：將訓練資料與經過結構處理的(hours-per-week)欄位資料匯入模型進行訓練，訓練次數（epochs）設定 20 次，批次更新的數量（batch-size）設定 20。

(7)模型績效評估：利用 kera 的模型評估（Evaluate）方法評估各模型的預測績效。

4.3.2 Bike sharing 資料集

(1)套件載入:載入協助資料分析的 Pandas 套件，並載入神經網路開發軟體 keras 的 Sequential 與 Dense 套件。

(2)正規化：以資料集取平均數與標準差，利用 Z-Score 標準化，將資料正規化。

(3)切割資料集：使用 Sklearn 套件將資料集切分成 80%訓練集與 20%測試集，並定義特徵與目標。

(4)建立神經網路模型：利用 keras 的 Sequential 與 Dense 套件撰寫前饋式神經網路，隱藏層（hidden layer）的激活函數以'Relu'、輸出層（output layer）的激活函數以'Sigmoid'為主，損失函數（loss function）利用評估

指標來檢測模型的成效。

(5)模型訓練：將訓練資料與經過結構處理的（Cnt、Season）欄位資料匯入模型進行訓練，訓練次數（epochs）設定 20 次，批次更新的數量（batch-size）設定 20。

(6)模型參數設定：將 Cnt 模型的隱藏層（hidden layer）分別設定為 1、2 和 3 層，而節點數（number of units）則是都設為 10，最後最佳化參數（optimizer）則使用 adam，完成參數設定即可進行模型訓練。

(7)模型績效評估：利用 kera 的模型評估（Evaluate）方法評估各模型的預測績效，Cnt 數值預測使用 MAE、MAPE、MSE 來評估目標值和預測值之偏差；Season 類別預測使用 Recall、Precision、F1-Score 來評估模型的準確度。

4.4 實驗結果

4.4.1 Adult dataset 結果預測績效評估

利用平均絕對誤差（MAE）、平均絕對誤差百分比（MAPE）、均方根誤差（RMSE）進行 hours-per-week 數值預測績效評估，評估目標值和預測值之偏差。

表 3 hours-per-week 績效評估總表

隱藏層	MAE	MAPE	RMSE
1	0.41	74.70	0.63
2	0.38	74.27	0.62
3	0.44	74.26	0.63
4	0.53	120.09	0.71

4.4.2 Bike Sharing 預測績效評估

利用平均絕對誤差（MAE）、平均絕對誤差百分比（MAPE）、均方根誤差（RMSE）進行 Cnt 數值預測績效評估，評估目標值和預測值之偏差。

表 4 Cnt 績效評估總表

隱藏層	MAE	MAPE	RMSE
1	0.55	70	0.53
2	0.64	68.59	0.53
3	0.67	71	0.53

利用 Recall、Precision、F1-Score 進行 Season 類別預測績效評估，評估模型的準確度。

表 5 Season 績效評估總表

優化器	隱藏層	Recall	Precision	F1-Score
Adam	1	1	0.93	1

Rmsprop	1	1	1	0.91
SGD	1	1	0.9	0.87

五、 結論

Adult Dataset 預測 hours-per-week 數值，當隱藏層層數為 2，輸入層與隱藏層使用 Relu 函數、輸出層使用 sigmoid 函數作為激活函數，並且優化器為 adam 時績效為最佳。

Bike Sharing Dataset 預測 cnt 數值，當隱藏層層數為 1，輸入層與隱藏層使用 Relu 函數、輸出層使用 sigmoid 函數作為激活函數，並且優化器為 adam 時績效為最佳。

Bike Sharing Dataset 預測 season 類別，當隱藏層層數為 2，輸入層與隱藏層使用 Relu 函數、輸出層使用 sigmoid 函數作為激活函數，並且優化器為 adam 時績效為最佳。

參考文獻

[1] 學習機器學習必知的程序-資料庫知識探索

https://medium.com/marketingdatascience/%E5%AD%B8%E7%BF%92%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E5%BF%85%E7%9F%A5%E7%9A%84%E7%A8%8B%E5%BA%8F-%E8%B3%87%E6%96%99%E5%BA%AB%E7%9F%A5%E8%AD%98%E6%8E%A2%E7%B4%A2-72bd2d73781c

[2] 過勞之島！台灣 2020 總工時全球第 4 比前年減 6 小時

https://udn.com/news/story/7238/5822033

[3] 用 Python 自學資料科學與機器學習入門實戰：Scikit Learn 基礎入門

https://blog.techbridge.cc/2017/11/24/python-data-science-and-machine-learning-scikit-learn-basic-tutorial/

[4] 深度學習：前饋神經網路 neural network

https://www.itread01.com/p/1422810.html

[5] Your First Deep Learning Project in Python with Keras Step-By-Step

https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

[6] 深度學習學習心得第四篇: FNN 實作

https://hackmd.io/@mumu-0/HJL_husUU?print-pdf#/

[7] Tensorflow 與 Keras 基本介紹

https://ithelp.ithome.com.tw/articles/10206261

[8] How to get accuracy, F1, precision and recall, for a keras model?

https://datascience.stackexchange.com/questions/45165/how-to-get-accuracy-f1-precision-and-recall-for-a-keras-model

[9] How to calculate precision and recall in Keras

https://stackoverflow.com/questions/43076609/how-to-calculate-precision-and-recall-in-keras

[10] 預測評價指標 RMSE、MSE、MAE、MAPE、SMAPE

https://blog.csdn.net/guolindonggld/article/details/87856780

[11] 10 Stochastic Gradient Descent Optimisation Algorithms + Cheatsheet

https://towardsdatascience.com/10-gradient-descent-optimisation-algorithms-86989510b5e9