國立雲林科技大學

機器學習專案作業一

前饋式神經網路預測類別與數值

指導教授：許中川　教授

學生：M10921002宋沂芸

　　　M10921016林恩杰

　　　M10921036童湘庭

　　　M10921038張珮柔

摘要

葡萄酒從過去被視為奢侈品到現今越來越受到消費者的青睞。各種產地、工法的葡萄酒品質也不盡相同，而葡萄酒的品質評估成為防止葡萄酒的非法參雜及確保葡萄酒市場的品質的關鍵要素。本研究採用UCI資料庫的Wine Quality資料作為分析，運用類神經網路進行分類及預測紅酒品質。分析結果顯示，對品質進行預測與分類均有不錯的結果分別為，在學習率0.001下MAE達0.268而Accuracy達0.829。

據研究發現低開發中國家之教育報酬率較高，也代表著教育能夠使人獲得獲取更高的報酬和獲得穩定的工作機會，教育程度與職業和收入的關聯業也影響著社經地位的高低，因此本研究欲了解造成教育程度高低之影響，故選取UCI資料庫的Adult資料，使用前饋神經網路進行預測分類透過 Precision, Recall及F1做績效評估，研究結果顯示在學習率為0.0005下準確率高達0.8997。

關鍵字：機器學習、前饋式神經網路

1. 緒論

1.1動機

1.1.1 wine quality white dataset

葡萄牙是世界十大葡萄酒出口國，從1997年到2007年葡萄酒出口增長了36%。質量評估通常是認證過程的一部分，可用於改善葡萄酒釀造（通過確定最有影響力的因素）以及對諸如高級品牌之類的葡萄酒進行分層（可用於確定價格）。葡萄酒的品質評估通常是透過其物理性質、化學性質及感官測試進行評估的。前二者通常測試密度、酒精、PH值等，而感官測試主要依靠品酒師進行。人類的感官測試較為主觀，且理化特質與人類的感官之間關係複雜，因此葡萄酒品質分類是一大挑戰。

1.1.2 adult dataset

根據 Psacharopoulos(1973)研究指出在低開發國家的教育報酬率一般較高，而人們常說教育能改變命運，是獲得進步的途徑，學習使人們更充實專業知識及技能，在工作時比起別人更具競爭力，且接受良好的教育能獲取更高的報酬和獲得穩定的工作機會，因此教育的重要性愈來愈受到重視(周新富，2008)。在人力資本論裡也明白指出教育與職業的關聯，學習與生產活動有關的高專業知識技能愈多，也進而提高工作效率以及收入。在快速的社會經濟發展過程中，經濟結構變動推動了社會階層化，並產生了社會流動，而教育程度也跟職業與收入相關，並影響了社經地位的變化，教育也因此得以充分發揮 (蔡瑞明，林大森，2000) 。故本研究欲了解職業、收入、年齡及資本等14項背景對教育程度的影響並進行分類。

1.2目的

1.2.1 wine quality white dataset

　　本研究採用UCI資料庫的Wine Quality資料作為分析，運用類神經網路進行分類及預測葡萄酒品質。根據葡萄酒的10個物化特質作為特徵進行預測，後續可以做為品質評估依據或可以給葡萄酒產地進行品質改善。

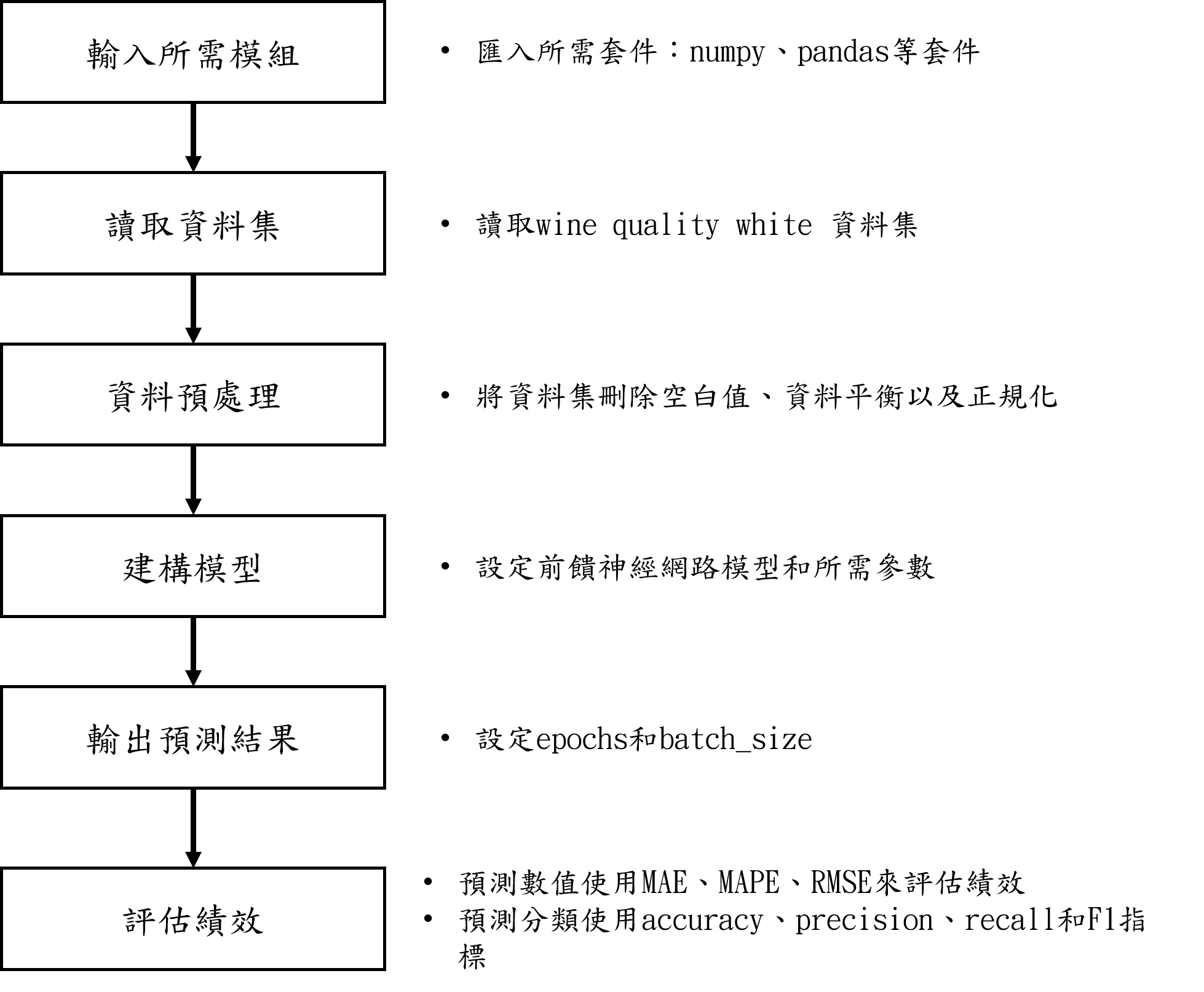
1.2.2 adult dataset

本研究欲探討影響教育程度之因素，故選取Adult資料集，並使用前饋神經網路對教育進行分析並預測類別，分析年齡、工作類別、教育程度、職業、種族等14種屬性資料，預測出一個人的教育程度為何，藉此更加了解什麼樣背景的人會接受到多高的教育程度。

1. 方法

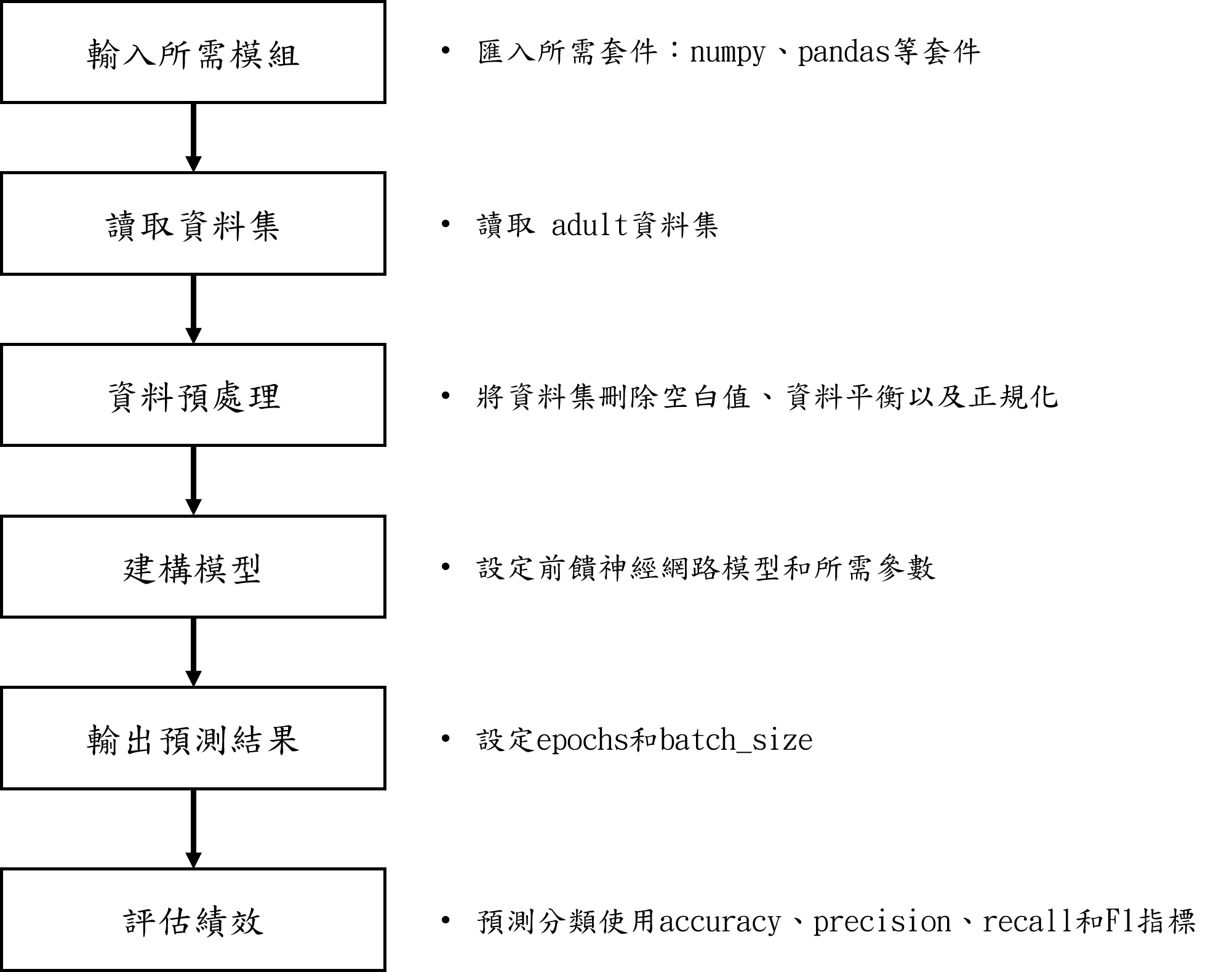
2.1程式架構

2.1.1 wine quality white dataset



圖一wine quality white dataset程式架構圖

2.1.2 adult dataset



圖二adult dataset程式架構圖

2.2 執行方法

本組專案使用前饋式神經網路（Feedforward Neural Network），參數從輸入層通過隱藏層向輸出層單向傳播，是屬於早期發明的簡單人工神經網路類型。

1. 實驗

3.1資料集

3.1.1 wine quality white dataset

此資料建立於2009年10月07日共有4896筆，12個欄位。

表一wine quality white dataset欄位資料說明彙總表

|  |  |
| --- | --- |
| 欄位名稱 | 欄位說明 |
| fixed acidity | 固定酸度 |
| volatile acidity | 揮發性酸度 |
| citric acid | 檸檬酸 |
| residual sugar | 殘留糖 |
| chlorides | 氯化物 |
| free sulfur dioxide | 游離二氧化硫 |
| total sulfur dioxide | 二氧化碳總量 |
| density | 密度 |
| PH | 衡量成熟度 |
| sulphates | 硫酸鹽 |
| quality | 品質 |

3.1.2 adult dataset

此資料集建立於1996年5月1日共有48,842筆，14個欄位。

表二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 | 教育年級 |
| 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 |
|  | 續下頁 |
| 承上頁 | |
| hours-per-week | 每周多少小時 |
| native-country | 祖國 |
| annual salary | 年收入 |

3.2前置處理

3.2.1 wine quality white dataset

　　本組將資料集用分號分隔欄位，而後檢查資料集是否有空白值和異常值，此資料集無任何缺失值或異常值。資料集quality欄位資料分布不均，quality值為6有2198筆，但品質為9僅有5筆，因此將資料進行平衡。接著將資料進行正規化，最後再分割資料集。

3.2.2 adult dataset

　　本組將訓練及測試資料的「連續數值」欄位刪除，同時也刪除空值的部分，接著將訓練及測試資料進行正規化，最後再將「education」欄位設為分類的目標，其他欄位為資料的特徵。

3.3實驗設計

3.3.1 wine quality white dataset

　　本組利用keras套件建立模型。預測數值將模型設置兩層，第一層節點為64個，第二層為128個。activation function使用relu，optimizer選擇rmsprop，loss function使用MSE，評估績效方法為MAE。Epochs為80個，batch\_size為16個。

預測分類時本組模型設置為4層，第ㄧ、二層皆為500個，而後棄置比例為0.3的節點，第四層設置50個節點，最後有7個分類結果。activation function使用relu和sigmoid，optimizer選擇rmsprop，loss function使用categorical\_crossentropy，評估績效方法為準確率。Epochs為8個，batch\_size為16個。

本組設置不同的學習率，藉由調整學習率觀察對於模型的影響。

3.3.2 adult dataset

　　本組利用keras套件建立模型。模型設置三層，分別為輸入層、一層隱藏層以及輸出層。表三為模型設計說明。模型編譯的優化器使用Adam，損失函數為Binary cross-entropy，評估績效方法為準確率，並且設置Epochs為8，batch\_size為16。

　　本組設置不同的學習率，藉由調整學習率觀察對於模型的影響。

表三 模型設計說明

|  |  |  |  |
| --- | --- | --- | --- |
| 層級 | 節點數 | 激活函數 | 參數數量 |
| 輸入層 | 8 | Relu | 112 |
| 隱藏層 | 16 | Relu | 144 |
| 輸出層 | 16 | Sigmoid | 272 |

3.4實驗結果

3.4.1 wine quality white dataset

　　本組設置4種不同的學習率分別為0.002、0.0015、0.001、0.005。實驗結果發現學習率設置為0.001時，預測數值績效最佳。

表四 wine quality white dataset預測數值績效評估表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 學習率 | MSE | MAE | MAPE | RMSE |
| 0.0005 | 0.2066 | 0.3128 | 44.2894 | 0.4545 |
| 0.001 | 0.1688 | 0.2682 | 44.7771 | 0.4109 |
| 0.0015 | 0.1863 | 0.2972 | 45.5380 | 0.4316 |
| 0.002 | 0.3268 | 0.4186 | 46.5557 | 0.5717 |

　　由表四可發現學習率為0.001時，模型準確率最高。表五至表八為各學習率不同的Precision、Recall和F1指標彙總表。

表五 wine quality white dataset分類準確率彙總表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 學習率  準確率 | 0.002 | 0.0015 | 0.001 | 0.0005 |
| accuracy | 0.8110 | 0.8142 | 0.8226 | 0.7979 |

表六 學習率=0.0005績效評估表

|  |  |  |  |
| --- | --- | --- | --- |
| 類別 | Precision | Recall | F1 |
| 0 | 0.98 | 1.00 | 0.99 |
| 1 | 0.91 | 0.87 | 0.89 |
| 2 | 0.63 | 0.68 | 0.65 |
| 3 | 0.55 | 0.30 | 0.39 |
| 4 | 0.61 | 0.77 | 0.68 |
| 5 | 0.85 | 0.95 | 0.90 |
| 6 | 1.00 | 1.00 | 1.00 |

表七 學習率=0.001績效評估表

|  |  |  |  |
| --- | --- | --- | --- |
| 類別 | Precision | Recall | F1 |
| 0 | 0.97 | 1.00 | 0.98 |
| 1 | 0.94 | 0.94 | 0.94 |
| 2 | 0.71 | 0.71 | 0.71 |
| 3 | 0.56 | 0.39 | 0.46 |
| 4 | 0.63 | 0.79 | 0.70 |
| 5 | 0.91 | 0.92 | 0.92 |
| 6 | 0.99 | 1.00 | 1.00 |

表八 學習率=0.0015績效評估表

|  |  |  |  |
| --- | --- | --- | --- |
| 類別 | Precision | Recall | F1 |
| 0 | 0.98 | 1.00 | 0.99 |
| 1 | 0.93 | 0.93 | 0.93 |
| 2 | 0.66 | 0.74 | 0.70 |
| 3 | 0.54 | 0.35 | 0.43 |
| 4 | 0.64 | 0.76 | 0.70 |
| 5 | 0.191 | 0.90 | 0.91 |
| 6 | 0.98 | 1.00 | 0.99 |

表九 學習率=0.002績效評估表

|  |  |  |  |
| --- | --- | --- | --- |
| 類別 | Precision | Recall | F1 |
| 0 | 0.98 | 1.00 | 0.99 |
| 1 | 0.85 | 0.96 | 0.90 |
| 2 | 0.65 | 0.69 | 0.67 |
| 3 | 0.54 | 0.30 | 0.39 |
| 4 | 0.68 | 0.73 | 0.70 |
| 5 | 0.86 | 0.98 | 0.91 |
| 6 | 1.00 | 1.00 | 1.00 |

3.4.2 adult dataset

　　本組設置4種不同的學習率分別為0.01、0.005、0.001、0.0005。實驗結果發現學習率設置為0.0005時，預測數值績效最佳。表十一至表十四為各學習率的績效評估表。

表十 adult dataset分類準確率彙總表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 學習率  準確率 | 0.01 | 0.005 | 0.001 | 0.0005 |
| accuracy | 0.8655 | 0.8885 | 0.8987 | 0.8997 |

表十一 學習率=0.01績效評估表

|  |  |  |  |
| --- | --- | --- | --- |
| 類別 | Precision | Recall | F1 |
| 0 | 0.90 | 0.86 | 0.88 |
| 1 | 0.96 | 0.92 | 0.94 |
| 2 | 1.00 | 0.12 | 0.22 |
| 3 | 0.00 | 0.00 | 0.00 |
| 4 | 0.57 | 0.75 | 0.65 |
| 續下頁 | | | |
| 承上頁 | | | |
| 5 | 0.77 | 0.66 | 0.71 |
| 6 | 0.74 | 0.64 | 0.68 |
| 7 | 1.00 | 0.88 | 0.94 |
| 8 | 1.00 | 0.86 | 0.93 |
| 9 | 1.00 | 0.81 | 0.90 |
| 10 | 1.00 | 0.68 | 0.81 |
| 11 | 0.75 | 1.00 | 0.86 |
| 12 | 0.83 | 0.78 | 0.80 |
| 13 | 0.51 | 0.78 | 0.62 |
| 14 | 1.00 | 0.01 | 0.02 |
| 15 | 1.00 | 0.89 | 0.94 |

表十二 學習率=0.005績效評估表

|  |  |  |  |
| --- | --- | --- | --- |
| 類別 | Precision | Recall | F1 |
| 0 | 0.95 | 0.79 | 0.86 |
| 1 | 1.00 | 0.93 | 0.96 |
| 2 | 1.00 | 0.96 | 0.98 |
| 3 | 0.85 | 0.65 | 0.74 |
| 4 | 0.67 | 0.78 | 0.72 |
| 5 | 0.76 | 0.72 | 0.74 |
| 6 | 0.69 | 0.67 | 0.68 |
| 7 | 1.00 | 0.88 | 0.94 |
| 8 | 1.00 | 0.86 | 0.93 |
| 9 | 1.00 | 0.81 | 0.90 |
| 10 | 1.00 | 0.68 | 0.81 |
| 11 | 0.77 | 1.00 | 0.87 |
| 12 | 1.00 | 0.78 | 0.88 |
| 13 | 0.90 | 0.67 | 0.77 |
| 14 | 1.00 | 0.58 | 0.74 |
| 15 | 1.00 | 0.89 | 0.94 |

表十三 學習率=0.001績效評估表

|  |  |  |  |
| --- | --- | --- | --- |
| 類別 | Precision | Recall | F1 |
| 0 | 0.98 | 0.92 | 0.95 |
| 1 | 1.00 | 0.93 | 0.96 |
| 續下頁 | | | |
| 承上頁 | | | |
| 2 | 1.00 | 0.96 | 0.98 |
| 3 | 0.92 | 0.85 | 0.88 |
| 4 | 0.90 | 0.85 | 0.87 |
| 5 | 0.92 | 0.85 | 0.89 |
| 6 | 0.94 | 0.84 | 0.89 |
| 7 | 1.00 | 0.88 | 0.94 |
| 8 | 1.00 | 0.86 | 0.93 |
| 9 | 1.00 | 0.81 | 0.90 |
| 10 | 1.00 | 0.68 | 0.81 |
| 11 | 0.77 | 1.00 | 0.87 |
| 12 | 1.00 | 0.78 | 0.88 |
| 13 | 1.00 | 0.74 | 0.85 |
| 14 | 1.00 | 0.58 | 0.74 |
| 15 | 1.00 | 0.89 | 0.94 |

表十四 學習率=0.0005績效評估表

|  |  |  |  |
| --- | --- | --- | --- |
| 類別 | Precision | Recall | F1 |
| 0 | 0.97 | 0.94 | 0.95 |
| 1 | 1.00 | 0.93 | 0.96 |
| 2 | 1.00 | 0.96 | 0.98 |
| 3 | 0.86 | 0.90 | 0.88 |
| 4 | 0.98 | 0.84 | 0.90 |
| 5 | 0.94 | 0.89 | 0.91 |
| 6 | 0.97 | 0.85 | 0.91 |
| 7 | 1.00 | 0.88 | 0.94 |
| 8 | 1.00 | 0.86 | 0.93 |
| 9 | 1.00 | 0.81 | 0.90 |
| 10 | 1.00 | 0.68 | 0.81 |
| 11 | 0.77 | 1.00 | 0.87 |
| 12 | 1.00 | 0.78 | 0.88 |
| 13 | 1.00 | 0.59 | 0.74 |
| 14 | 1.00 | 0.58 | 0.74 |
| 15 | 1.00 | 0.89 | 0.94 |

1. 結論

4.1 wine quality white dataset

　　本組透過這次專案發現到學習率雖變化浮動非常小，但對於模型還是會產生影響。由這次結果可得知學習率為0.001時，在預設數值和分類的績效表現皆為最佳。在實驗過程中發現Epochs參數設定過大時，訓練結果很好而測試結果不佳，測試模型產生過度擬合狀況，合適的Epochs數值對模型的學習能力也很重要。

4.2 adult dataset

　　本組經由專案發現，在模型設計方面，設置不同層數會造成資料在訓練及測試時，準確率會有很大的差異性，同時節點數量的多寡也會對準確率造成影響。因此本組將層數、節點數、epoch及batch size的參數設置相同，查看優化器在不同的學習率是否會有影響績效，實驗結果可以發現，設置越小的學習率會有較好的績效，不過當數值到一定的值，其績效也會趨於相同，不會有明顯的差異。

1. 參考文獻
2. G. Psacharopoulos, Returns to Education. Amsterdam, Elsevier Scientific Publishing Company, 1973, XIII p. 216 p., DFL 27.50. (1975). *Louvain Economic Review,* *41*(1), 82-82. doi:10.1017/S0770451800000361
3. 蔡瑞明、林大森 (2000),。教育與勞力市場的連結：以台灣的教育分流為例,。載於劉兆佳等（主編），市場、階級與政治：變遷中的華人社會(頁143-190)。香港: 香港中文大學。
4. 周新富(2008)。社會階級對子女學業成就的影響⎯⎯ 以家庭資源為分析架構。臺灣教育社會學研究，8(1)，1-43。
5. 前饋式神經網路https://zh.wikipedia.org/wiki/%E5%89%8D%E9%A6%88%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C