國立雲林科技大學資訊管理系

機器學習-作業一

Department of Information Management

National Yunlin University of Science & Technology Assignment

MNIST和波士頓房價和成人資料集分析

MNIST and Boston Housing Price and Adult Data Set analysis

楊欣蓓、黃裕鳴、游棨翔

指導老師：許中川 博士

Advisor: Chung-Chian Hsu, Ph.D.

 中華民國113年4月

April 2024

**摘要**

本研究使用神經網路對MNIST數字辨識資料集、波士頓房價資料集和成人收入資料集進行實驗分析。在MNIST實驗中,發現使用單隱藏層、128個神經元的相對簡單模型就能達到98.24%的準確率,無需過度複雜化網路結構。在波士頓房價實驗中,單隱藏層且神經元數量較少的模型收斂最快且預測性能最優,過於複雜的模型會導致過擬合。在成人收入實驗的分類問題中,relu激活函數的準確率最高達85.28%;回歸問題中,適度增加模型複雜度可提升性能,但需權衡模型穩定性。總結來說,對於小規模數據集,簡單模型就能取得出色性能;而對大規模數據集,適度增加模型複雜度有助於提升性能,但也要考慮模型穩定性。實際應用中需根據任務和數據特點審慎選擇適當的神經網路模型和超參數配置,以達到最佳機器學習效果。

關鍵字：relu、神經網路、隱藏層、神經元

**一、緒論**

1. **研究動機**

本實驗欲使用神經網路搭配適當的激活函數研究MNIST資料集、波士頓房價資料集、成人資料集。

關於MNIST資料集研究，現在數字辨識系統被廣泛運用在各個產業中，透過這些系統可以有效率地提升工作效率，其中在設計辨識系統時，提高辨識正確率是很重要的環節，所以本實驗欲透過MNIST資料集，訓練神經網路模型，並從中觀察出最適當的參數搭配，以提高模型辨識正確率。

關於波士頓房價資料集研究，現今隨著城市的發展，人們居住的房屋不僅更舒適，社區的交通環境也整治地更安全便利，但隨之而來的是房價的起伏，房價不僅考慮其本身房屋的品質，更看重視外在因素，因此本實驗欲透過分析波士頓房價資料集，觀察房價與外在因素之間的關係，以幫助城市規劃以及政策的考量。

關於成人資料集研究，通過對收入的預測分析，更準確地評估個人的價值和競爭力，對於求職者和用人單位都具有重要意義。對求職者而言，可以根據自身的預測收入水平選擇合適的職業方向和教育規劃；對企業而言，可以根據候選人的預測收入水平制定更合理的薪資策略和人才招聘計劃，提高用工效率和競爭力。工作時間的管理對個人和企業的生產效率和生活品質都有著重要的影響。通過對工作時間的預測分析，可以幫助個人更有效地安排工作與生活，提高生活品質和工作效率；對企業而言，也可以通過優化人力資源配置和工作時間管理提高生產力和競爭力。

1. **研究目的**

本研究目的想以神經網路對MNIST資料集、波士頓房價資料集、成人資料集進行實驗。本研究於上述不同的資料集實驗中，探討要使用何種激活函數下去做實驗，並來回調整模型的超參數，最後用對應的績效指標進行模型績效測試，最終從實驗結果中探討出各個實驗有價值的訊息。

**二、實驗方法**

1. **實作說明**

本實驗使用神經網路對MNIST資料集、波士頓房價資料集、成人資料集進行實驗。進行各個實驗前，需各自做適當的資料前處理，以獲得乾淨的資料，再來開始建立神經網路，通過調整神經網路的隱藏層數、神經元顆數、優化器、batch size、epoch，其中激活函數的部分，於隱藏層時大部分使用relu、分類預測使用softmax、迴歸預測因目標為預測數值，而無需要激活函數，直接輸出數值即可，透過上述調整使模型更加優化，最後將調整好的模型對測試集進行績效衡量，對類別預測績效時，使用Precision, Recall及F1績效指標；數值預測績效則使用MAE、MAPE與RMSE績效指標。

1. **操作說明**

本研究執行環境皆採用Python3.10.10，以Visual Studio Code作為開發工具，神經網路進行研究分析，並使用Pandas、Numpy、Mlxtend等函式庫來讀取資料。於各個資料前處理，利用Pandas套件功能，刪除空值、重複資料、數值標準化、對名目資料做one-hot-encoding變成模型所需的型態…等。

**三、實驗設計**

1. **資料集**

名稱：MNIST資料集

原始資料筆數：6000筆

資料前處理後資料筆數：6000筆

**表1**

*MNIST資料集欄位介紹*

|  |  |  |
| --- | --- | --- |
| 欄位 | 屬性 | 內容 |
| 0 | pixel | Continuous |
| 1 | label | Nominal |

名稱：波士頓房價資料集

原始資料筆數：404筆

資料前處理後資料筆數：404筆

**表2**

*波士頓房價資料集欄位介紹*

|  |  |  |
| --- | --- | --- |
| 欄位 | 屬性 | 內容 |
| 0 | CRIM | Continuous |
| 1 | ZN | Continuous |
| 2 | INDUS | Continuous |
| 3 | CHAS | Continuous |
| 4 | NOX | Continuous |
| 5 | RM | Continuous |
| 6 | AGE | Continuous |
| 7 | DIS | Continuous |
| 8 | RAD | Continuous |
| 9 | TAX | Continuous |
| 10 | PTRATIO | Continuous |
| 11 | B | Continuous |
| 12 | LSTAT | Continuous |
| 13 | MEDV | Continuous |

名稱：成人資料集

原始資料筆數：32561筆

資料前處理後資料筆數：30139筆

**表3**

*成人資料集欄位介紹*

|  |  |  |
| --- | --- | --- |
| 欄位 | 屬性 | 內容 |
| 0 | age | Continuous |
| 1 | workclass | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. |
| 2 | fnlwgt | Continuous |
| 3 | education | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. |
| 4 | education-num | continuous |
| 5 | marital-status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| 6 | occupation | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces |
| 7 | relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| 8 | race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| 9 | sex | Female, Male |
| 10 | capital-gain | Continuous |
| 11 | capital-loss | Continuous |
| 12 | hours-per-week | Continuous |
| 13 | native-country | United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands |
| 14 | income | <=50,>50 |

1. **資料前處理**
2. **MNIST資料集**

* 資料前處理
* 將從原本二維向量圖片攤平成一維向量，再來將原始數據類型轉換成float32，最後將所有數值除以255，使數值映射到0到1之間，使不同特徵以同標準表示出。
* 將訓練集和測試集做OneHotEncoder，使原本的數字標標籤轉為長度為10的向量，且對應到的位置為1，其餘則為0。
* 將為了讓模型避免overfitting，進行分割訓練集，分成訓練集80%和驗證集20%，可以檢視模型是否過度訓練。

**表4**

*MNIST資料集的圖片數字2資料處理前向量值*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1 | **2** | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

**表5**

*MNIST資料集的圖片數字2資料處理後向量值*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

1. **波士頓房價資料集**

* 資料前處理
* 確認資料集使否有重複值、空值，若有則將其移除。
* 對資料集做正規化，其中不可將全部的資料集一起作z-score，因測試集為未知項目，若一開始將測試集一起作z-score，則會使模型預測不準確，所以先行算出訓練集的平均值跟標準差，對訓練集做出z-score，再使用訓練集的平均值跟標準差，對測試集做出z-score。最終將數值都壓縮到0到1的範圍內。
* 將為了讓模型避免overfitting，進行分割訓練集，分成訓練集80%和驗證集20%，可以檢視模型是否過度訓練。

**表6**

*部分資料處理後的波士頓房價資料集*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 特徵  資料 | CRIM | ZN | INDUS | … | B | LSTAT | MEDV |
| No.0 | -0.271909 | -0.483017 | -0.435222 | 0.447522 | 0.824198 | -0.781184 |
| No.1 | -0.402927 | 2.988079 | -1.33226 | 0.431371 | -1.327556 | 2.161129 |
| No.2 | 0.124785 | -0.483017 | 1.027052 | 0.220344 | -1.30688 | 2.997136 |
| No. 3 | -0.400996 | -0.483017 | -0.868325 | 0.447522 | -0.652118 | -0.140607 |
| No. 4 | -0.005627 | -0.483017 | 1.027052 | 0.389399 | 0.263171 | -0.509753 |

1. **成人資料集**

* 資料前處理
* 確認資料集使否有重複值、空值，若有則將其移除。
* 用LabelEncoder把名目特徵欄位轉換成數字型態。
* 用OneHotEncoder把salary欄位做型態轉換，「>=50」轉換為0、「<50」轉換為1。
* 對資料做z-score正規化，將數值都壓縮到0到1的範圍內。

**表7**

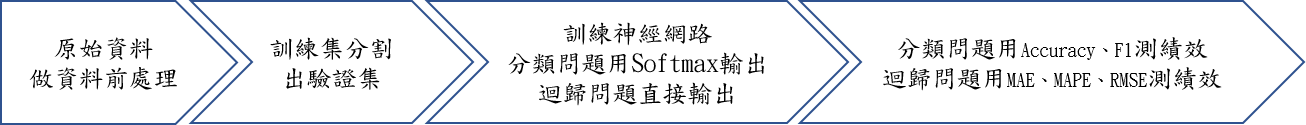
*部分資料處理後的成人資料集*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 特徵  資料 | age | workclass | fnlwgt | … | capital-loss | hours-per-week | salary |
| No.0 | 0.042516 | 5 | -1.062676 | 0.145925 | 40 | 0 |
| No.1 | 0.880215 | 4 | -1.007829 | -0.147502 | 13 | 0 |
| No.2 | -0.033639 | 2 | 0.244669 | -0.147502 | 40 | 0 |
| No. 3 | 1.108678 | 2 | 0.425206 | -0.147502 | 40 | 0 |
| No. 4 | -0.795183 | 2 | 1.406572 | -0.147502 | 40 | 0 |

1. **實驗設計**

***圖 1***

*各個實驗設計流程圖*



1. **MNIST資料集**

本研究將MINST資料集做資料前處理，首先將訓練集和測試集做flattened，把二維資料轉換成一維向量，為後面輸入神經網路模型時做好準備，接著因神經網路訓練時是使用浮點做運算，所以先將原始數據轉換成float32，再來因圖片像素的數值都在0到255的範圍內，所以要對圖像數據進行正規化處理的話，透過除以255可以將數值都壓縮在0到1的範圍內，使得模型在訓練時可以更加穩定，最後對訓練、測試集的數字標籤進行one-hot-encoding轉換成長度為10的向量，對應到的位置顯示1，其餘則顯示0。在進行訓練前，為了避免overfitting，所以先對訓練集進行切割，各別切割成80%訓練和20%驗證，然後開始進行神經網路訓練，透過調整超參數找出訓練最佳模型，再進行Precision、Recall、Accuracy、F1績效測試。

1. **波士頓房價資料集**

本研究將波士頓房價資料集做資料前處理，確認資料集使否有重複值、空值，若有則將其移除，對資料集做z-score正規化，將數值都壓縮到0到1的範圍內，此外為了讓模型避免overfitting，進行分割訓練集，分成訓練集80%和驗證集20%，可以檢視模型是否過度訓練，然後開始進行神經網路訓練，透過調整超參數找出訓練最佳模型，再進行MAE、MAPE、RMSE績效測試。

1. **成人資料集**

本研究將成人資料集做資料前處理，確認資料集使否有重複值、空值，若有則將其移除，接著用LabelEncoder把名目特徵欄位的資料對應到數字上、用OneHotEncoder把salary欄位的「>=50」轉換為0、「<50」轉換為1，對資料集做z-score正規化，將數值都壓縮到0到1的範圍內，此外為了讓模型避免overfitting，進行分割訓練集，分成訓練集95%和驗證集5%，檢視模型是否過度訓練，然後開始進行神經網路訓練，透過調整超參數找出訓練最佳模型，其中分類問題使用Softmax輸出結果用Precision、Recall、Accuracy、F1績效測試、迴歸問題直接輸出結果，用MAE、MAPE、RMSE績效測試。

1. **實驗結果**

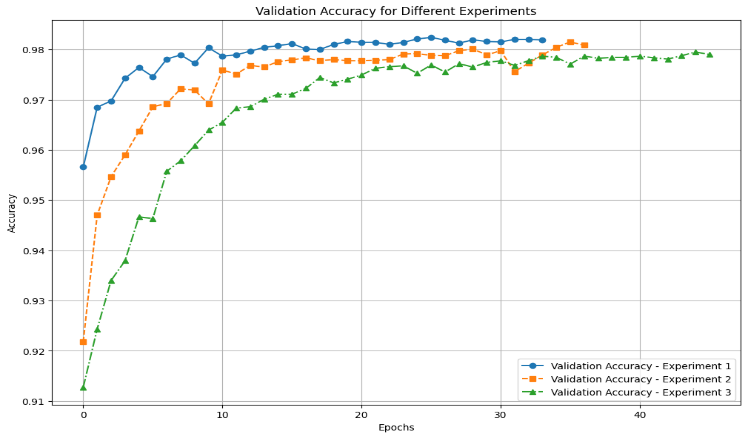
本實驗主要研究對各個資料集調整神經網路架構和超參數設定，分類研究和迴歸研究的績效分別以驗證集的Accuracy和MAE下去觀察，探討其影響的原因。

1. **MNIST資料集分析**

本實驗為了有效調整架構跟參數，所以先固定神經網路的架構，比較激活函數的績效，從下表8和圖2，可以觀察出relu的績效最佳，所以本實驗將激活函數固定為relu。

**圖2**

*MNIST資料集不同activation function的Accuracy績效*

**

**表8**

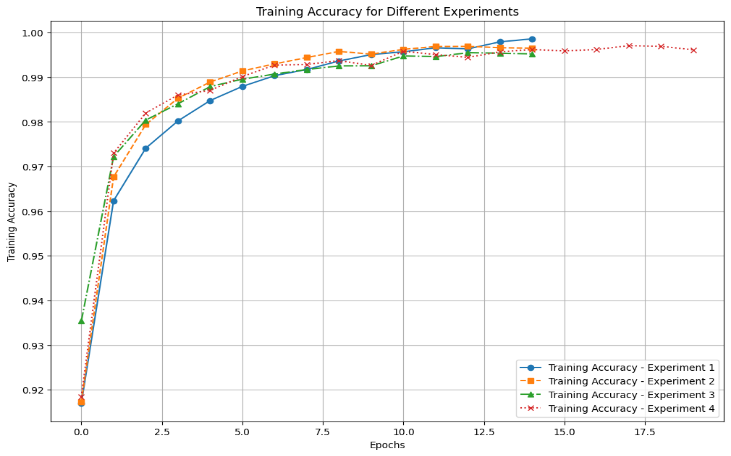
*MNIST資料集不同activation function的績效*

|  |  |  |  |
| --- | --- | --- | --- |
|  | relu(Experiment1) | tanh(Experiment2) | sigmoid(Experiment3) |
| Accuracy | **0.9824** | 0.9815 | 0.9794 |
| Loss | **0.0800** | 0.0819 | 0.0837 |

接著調整神經網路架構和超參數，本研究設計了四種實驗，設定的參數值和績效於下方表9、表10，從實驗結果下方圖3判斷，各組的模型隨著複雜度上升，驗證集的準確度皆有逐漸變好，但最佳與最差績效只差了0.0025，彼此的神經網路架構和超參數卻相差甚大，推測造成該結果之原因可能為資料集筆數較少，只需簡單架構就可以達到高績效的預測。本研究認為對此實驗用最簡單的架構和超參數即可，無需進行複雜的訓練，也可避免模型overfitting。

**圖3**

*MNIST資料集不同架構跟超參數Accuracy績效*



**表9**

*MNIST資料集不同架構跟超參數和Accuracy、Loss績效*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Batch size | patient | Number of  hidden layers | Number of units | Activation function | optimizer | Accuracy | Loss |
| Experiment1 | 32 | 10 | 1 | 128 | RELU | Adam | 0.9778 | 0.0993 |
| Experiment2 | 128 | 10 | 2 | 256 | RELU | Adam | 0.9795 | 0.1033 |
| Experiment3 | 64 | 10 | 3 | 512 | RELU | Adam | 0.9790 | 0.1220 |
| Experiment4 | 256 | 10 | 4 | 1024 | RELU | Adam | **0.9803** | **0.0961** |

**表10**

*MNIST資料集不同架構跟超參數和precision、recall、f1\_score績效*

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1\_score |
| Experiment1 | 0.9786 | 0.9768 | 0.9776 |
| Experiment2 | 0.9800 | 0.9791 | 0.9793 |
| Experiment3 | 0.9801 | 0.9778 | 0.9789 |
| Experiment4 | **0.9812** | **0.9797** | **0.9802** |

1. **波士頓房價資料集分析**

本研究固定神經網路的架構，找出最佳激活函數為tanh，於表11，接著進行不同架構和不同超參數探討，從圖4觀察出實驗組一的收斂速度最快，並且績效也為最佳，本研究臆測為神經網路層數較低，會使模型較為簡單，相對於其他對照組，模型較快收斂，另，此實驗集筆數較少，使用高層數的模型績效反而不佳，固推測實驗筆數也會影響模型訓練的績效。

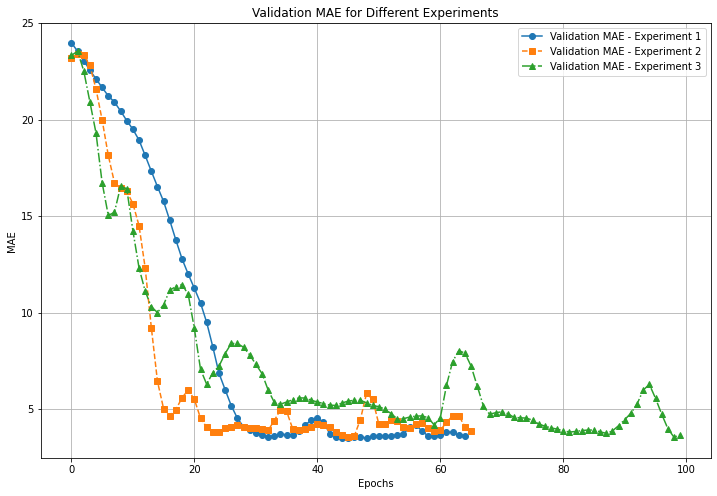
**表11**

*波士頓房價資料集不同activation function績效*

|  |  |  |  |
| --- | --- | --- | --- |
|  | relu(Experiment1) | tanh(Experiment2) | sigmoid(Experiment3) |
| MAE | 3.6021 | **3.0734** | 4.2276 |
| MAPE | 16.1507 | **13.3289** | 16.6902 |
| RMSE | 4.8237 | **4.4602** | 6.2986 |

**圖4**

*波士頓房價資料集不同架構跟超參數MAE績效*



**表12**

*波士頓房價資料集不同架構跟超參數、MAE績效*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Batch size | patient | Number of  hidden layers | Number of units | Activation function | optimizer | MAE |
| Experiment1 | 64 | 10 | 1 | 1000 | tanh | Adam | **3.5053** |
| Experiment2 | 128 | 10 | 2 | 1000 | tanh | Adam | 3.5720 |
| Experiment3 | 256 | 10 | 3 | 1000 | tanh | Adam | 3.5493 |

**表13**

*波士頓房價資料集不同架構跟超參數MAPE、RMSE績效*

|  |  |  |
| --- | --- | --- |
|  | MAPE | RMSE |
| Experiment1 | 16.2207 | **2.4286** |
| Experiment2 | 16.0683 | 5.1480 |
| Experiment3 | **14.7474** | 4.9304 |

1. **成人資料集分析**

本研究分為對‘income’欄位進行分類研究和對‘hours-per-week’進行迴歸研究。首先個別固定神經網路的架構，從下方表14和表15得知分類研究的最佳激活函數分類研究為relu、而迴歸研究為tanh，接續探討不同架構和不同超參數影響績效為何。

關於分類研究的實驗結論，從圖5觀察出增加神經網路層數時，其驗證集的績效穩定度起伏也會變大，本研究猜測更深的神經網路結構會增加模型的複雜度,也同時容易出現overfitting的問題，因此本研究認為在找出最佳績效的同時,也需要衡量模型的穩定性和泛化能力。

關於迴歸研究的實驗結論，本研究從下方圖6的MAE曲線中觀察出最佳績效為實驗組三，但其起伏劇烈不穩定，相對於實驗組一，雖然並非有最佳績效，但其穩定度則為三者最佳，因此本研究認為雖然其資料筆數大，普遍增加神經層數可以使模型績效變好，但於此實驗結果得知，使用簡單模型在高績效同時，也可保持其穩定度。

**表14**

*成人資料集分類研究的不同activation function績效*

|  |  |  |  |
| --- | --- | --- | --- |
|  | relu(Experiment1) | tanh(Experiment2) | sigmoid(Experiment3) |
| Accuracy | **0.8528** | 0.8522 | 0.7487 |
| Loss | **0.3257** | 0.3291 | 0.7268 |

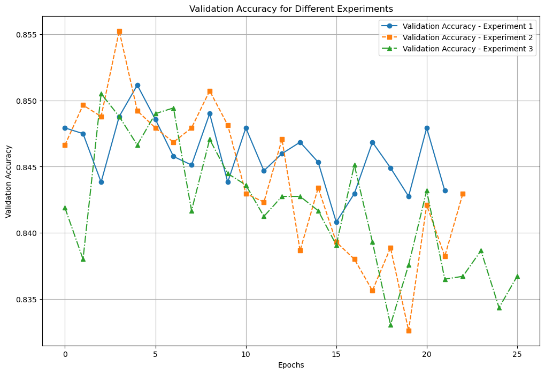
**表15**

*成人資料集迴歸研究的不同activation function績效*

|  |  |  |  |
| --- | --- | --- | --- |
|  | relu(Experiment1) | tanh(Experiment2) | sigmoid(Experiment3) |
| MAE | 7.730341 | **7.184071** | 7.289593 |
| MAPE | 0.250105 | **0.238316** | 0.254253 |
| RMSE | 11.312093 | **10.663094** | 10.756505 |

**圖5**

*成人資料集分類研究的不同架構跟超參數Accuracy績效*



**表16**

*成人資料集分類研究的不同架構跟超參數和Accuracy、Loss績效*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Batch size | patient | Number of  hidden layers | Number of units | Activation function | optimizer | Accuracy | Loss |
| Experiment1 | 64 | 10 | 1 | 1000 | relu | Adam | 0.8511 | 0.3188 |
| Experiment2 | 128 | 10 | 2 | 1000 | relu | Adam | **0.8552** | **0.3179** |
| Experiment3 | 256 | 10 | 3 | 1000 | relu | Adam | 0.8505 | 0.3210 |

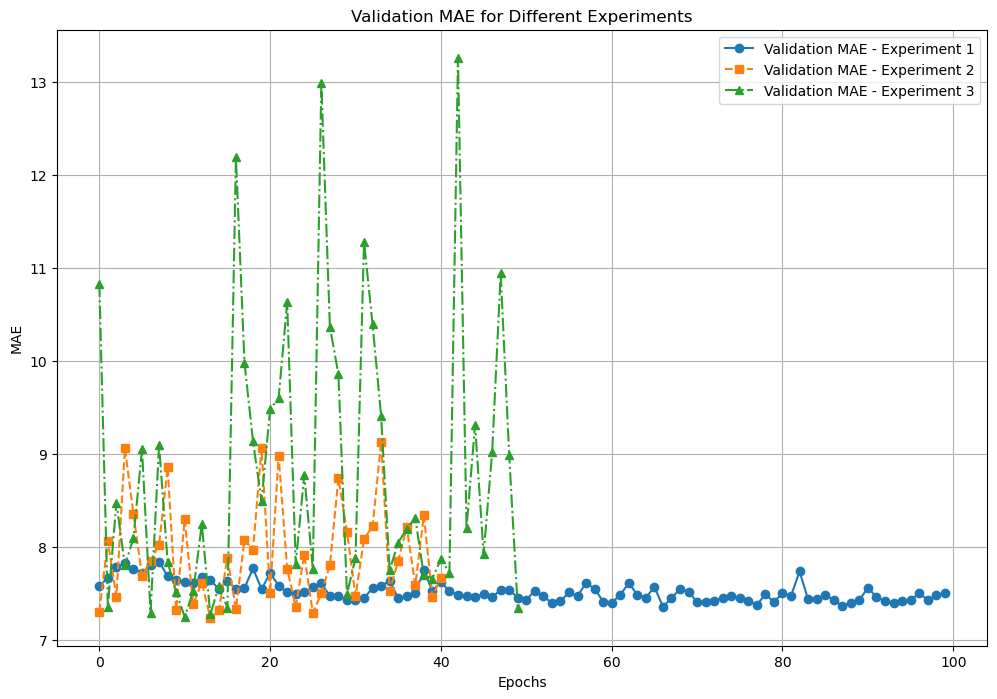
**表17**

*成人資料集分類研究的不同架構跟超參數precision、recall、f1\_score績效*

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1\_score |
| Experiment1 | **0.7994** | 0.4886 | **0.3953** |
| Experiment2 | 0.7395 | 0.6016 | **0.3953** |
| Experiment3 | 0.7341 | **0.6165** | **0.3953** |

**圖6**

*成人資料集迴歸研究的不同架構跟超參數MAE績效*



**表18**

*成人資料集迴歸研究的不同架構跟超參數MAE績效*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Batch size | patient | Number of  hidden layers | Number of units | Activation function | optimizer | MAE |
| Experiment1 | 64 | 10 | 1 | 1000 | tanh | Adam | 7.3580 |
| Experiment2 | 128 | 10 | 2 | 1000 | tanh | Adam | 7.2380 |
| Experiment3 | 256 | 10 | 3 | 1000 | tanh | Adam | **7.2483** |

**表19**

*成人資料集迴歸研究的不同架構跟超參數MAPE、RMSE績效*

|  |  |  |
| --- | --- | --- |
|  | MAPE | RMSE |
| Experiment1 | **27.9536** | **10.6053** |
| Experiment2 | 31.4397 | 12.0248 |
| Experiment3 | 31.4219 | 12.0248 |

**四、結論**

機器學習在各個領域的應用越來越廣泛,從數字辨識到房價預測再到收入預測,都需要建立合適的機器學習模型。本次實驗通過對MNIST、波士頓房價和成人三個不同數據集的分析,探討了神經網路模型在這些任務中的表現。

首先是MNIST數字辨識資料集,實驗發現使用單隱藏層、128個神經元的相對簡單模型就能達到98.24%的準確率,無需過度複雜化網路結構。這是由於MNIST資料集規模相對較小,簡單的模型就能捕捉到足夠的特徵。因此對於此類數據集,不必過度追求模型複雜度,反而可能導致過擬合,適度的簡單模型就能取得優秀的性能。

其次是波士頓房價預測資料集,實驗結果顯示單隱藏層且神經元數量較少的模型收斂最快,且預測性能也相對最優。這可能是因為波士頓房價資料集規模較小,使用過於複雜的模型會導致過擬合。因此對於此類中等規模的回歸問題,適度簡單的模型結構就能取得較佳的效果。

最後是成人收入預測資料集,實驗從分類和回歸兩個角度進行探討。分類問題中,ReLU激活函數的準確率最高,達到85.28%。而回歸問題中,儘管增加模型複雜度可以進一步提升性能,但需要權衡模型穩定性。簡單模型雖然性能略有遜色,但變化更加平穩,是更安全的選擇。這說明對於大規模數據集,適度增加模型複雜度能帶來收益,但也要顧及模型的可靠性。

總的來說,不同特徵和規模的數據集,其最佳的模型配置也存在差異。對於小規模數據集,相對簡單的模型結構就能取得出色的性能,無需追求過于複雜的網路架構。而對於大規模數據集,適度增加模型複雜度可以進一步提升性能,但也要平衡模型的穩定性。因此在實際應用中,需要根據具體任務和數據特點,審慎選擇合適的神經網路模型和超參數配置,以達到最佳的機器學習效果。

**參考文獻**

Dario Radečić (2021)。如何使用 TensorFlow 優化學習率——比你想像的要容易。

[How to Optimize Learning Rate with TensorFlow — It’s Easier Than You Think | by Dario Radečić | Towards Data Science](https://towardsdatascience.com/how-to-optimize-learning-rate-with-tensorflow-its-easier-than-you-think-164f980a7c7b)

Ryan Lu (2018)。Preprocessing Data : 類別型特徵\_OneHotEncoder &LabelEncoder 介紹與實作。<https://medium.com/ai%E5%8F%8D%E6%96%97%E5%9F%8E/preprocessing-data-onehotencoder-labelencoder-%E5%AF%A6%E4%BD%9C-968936124d59>