雲林科技大學資訊管理學系

機器學習

Department of Information Management National Yunlin University of Science and Machine Learning

專案作業一

基於兩個資料集使用前饋式神經網路進行回歸預測與分類預測之績效比較

|  |  |
| --- | --- |
| Student Name: | 彭冠穎 |
| Student ID: | M10923025 |
| E-mail: | M10923025@gmail.yumtech.edu.tw |
| Student Name: | 趙國翔 |
| Student ID: | M10923013 |
| E-mail: | M10923013@gmail.yumtech.edu.tw |
| Student Name: | 曾信嘉 |
| Student ID: | M10923032 |
| E-mail: | M10923032@gmail.yumtech.edu.tw |
| Student Name: | 陳楚柔 |
| Student ID: | M10923011 |
| E-mail: | M10923011@gmail.yumtech.edu.tw |

指導教授:許中川

Advisor：Chung-Chian Hsu, Ph.D.

中華民國110年4月

April 2021

**摘要**

本實驗使用前饋式神經網路針對UCI Datasets中的Adult Data及Abalone Data進行分類與數值預測，並透過Talos進行參數優化調整，且使用不同層數以及不同架構的結果查看績效預測之影響，最後實驗結果表明，hidden layer 以及 hidden neuron並不是越多越好。

關鍵字：前饋式神經網路、分類預測、數值預測、Talos

**第一章、緒論**

1.1動機

企業能夠透過機器學習演算法來幫助管理找尋找大量資料中所包含的意義，而深度學習式機器學習中的分支，深度學習是指在機器學習的基礎架構上，強調使用連續、多層的學習方式，而這種多層次的表示法幾乎都是來自於神經網路模型。本研究將透過兩個不同的資料集進行實驗，分別為UCI Datasets中的Adult Data及Abalone Data，並透過目前所學之前饋式神經網路模型進行分類、數值預測。

1.2目的

本研究主要目的為透過前饋式神經網路模型進行分類、數值預測分析，將使用UCI Datasets中兩個資料集中進行實驗，資料及分別為Adult Data及Abalone Data，透過talos進行參數優化調整，以加速調整參數所需花費時間，並嘗試使用不同層數變換神經網路模型實驗，比較不同架構對於結果造成什麼影響，並將績效結果紀錄於本實驗。

**第二章、方法**

本研究透過兩個資料集(Adult Data及Abalone Data)的歷史資料，結合目前學習的機器學習的方法(ex.前饋式神經網路)，資料集Adult透過歷史資料中找到各個工作者的教育程度(分類預測)，而資料集Abalone透過歷史資料中的鮑魚表面特徵來去預測出鮑魚的實際年齡(數值預測)、性別(分類預測)，並比較不同層數神經網路模型所產生的績效，從中找到績效最好的層數。

**2.1程式操作方法**

Step1：請至UCI下載本研究所使用的adult資料集與Abalone資料集。

Step2：請至本小組GitHub專案資料夾下載程式碼

Step3：請將下載的程式碼放在資料夾底下，使用Jupyter notebook執行程式碼檔案即可操作程式。

**第三章、實驗**

**3.1資料集**

* 名稱:Adult資料集
* 原始資料筆數:48842
* 正規化後之訓練資料筆數:30162
* 正規化後之測試資料筆數:15059

表3-1-1：Adult資料集欄位介紹表

|  |  |  |
| --- | --- | --- |
| 欄位 | 欄位名稱 | 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 |
| 續下表 | | |
| 呈上表 | | |
| 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 | salary | <=50K,>50K |

* 名稱: Abalone資料集
* 原始資料筆數: 4177
* 缺失值筆數: 0

表3-1-2 Abalone欄位介紹表

|  |  |  |
| --- | --- | --- |
| 名稱 | 資料類型 | 資料描述 |
| Sex | nominal | M, F, and I (infant) |
| Length | continuous | Longest shell measurement |
| Diameter | continuous | perpendicular to length |
| Height | continuous | with meat in shell |
| Whole Weight | continuous | whole abalone |
| Shucked Weight | continuous | weight of meat |
| Viscera Weight | continuous | gut weight (after bleeding) |
| Shell Weight | continuous | after being dried |
| Ring | integer | +1.5 gives the age in years |

**3.2前置處理**

* **Label encoding轉換**

因機器在進行學習時無法辨識非數值資料，因此須將非數值資料轉換為數值資料，透過Labelencoding方法將每個非數值映射到某個整數，並且不會增加欄位。

* **資料正規化**

因資料中不同欄位的分布狀況不同，而在進行訓練時可能會使模型複雜度提升，進而造成Overfitting ，因此為預防此狀況可透過正規化的方式將資料按照比例縮放，使資料落在某一特定區間。本研究將透過資料正規化將輸入參數進行特徵縮放至0~1。

* **資料欄位選擇**

由於欄位education-num為預測目標的正確答案，因此將本欄位刪除，不納入輸入資料。

**3.3實驗設計**

表3-3-1類神經網路參數定義表

|  |  |
| --- | --- |
| 變數 | 定義 |
| activation | 1. ReLU在神經網絡中，線性整流作為神經元的激活函數，定義了該神經元在線性變換之後的非線性輸出結果。 2. Linear對輸入數據應用線性轉換 |
| kernel\_initializer | 設置 Keras各層權重隨機初始值的方法 |
| loss | loss這個函數是用來最小化甚麼目標函數的計算方法的。sklearn提供了多種選擇： 1）輸入均方誤差「mean\_squared\_error」，使用最小平方法的目標函數預測值與實際值的差距之平均值。  2）輸入Hinge Error 「hinge」，是一種單邊誤差，不考慮負值  3) 輸入Cross Entropy「categorical\_crossentropy」，預測值與實際值愈相近，損失函數就愈小，反之差距很大，就會更影響損失函數的值。 |
| Optimizer | Adam 是一種可以替代傳統隨機梯度下降過程的一階優化算法，它能基於訓練數據疊代地更新神經網絡權重。 |
| validation\_split | 用於在沒有提供驗證及的時候，按一定比例從訓練集中取出一部份作為驗證集 |
| epochs | epochs被定義為向前和向後傳播中所有批次的單次訓練疊代。 |
| batch\_size | 每次修正的樣本數 |

**3.4挑選參數**

* Talos

因為在前饋式神經網路中可以使用的參數與可調參數有無數種組合，而人力有限無法每次都實際測試，因此我們使用了Talos套件來挑選參數，Talos可以讓使用者事先設定好參數的範圍，讓Talos按照使用者輸入的參數範圍對模型進行測試，根據使用者所選擇的績效來反映出最優參數。

**3.5實驗結果**

本研究針對Adult資料集進行欄位education的分類預測，並透過Talos調整參數，調整結果與範圍如表3-5-1。

表3-5-1 Adult分類預測使用Talos所挑選的範圍與最優參數表

|  |  |  |  |
| --- | --- | --- | --- |
| 參數名稱 | 參數說明 | 參數範圍 | 最優參數 |
| first\_neuron | 第一層輸入層所設置的neuron | [32, 64, 128, 256] | 128 |
| hidden\_neuron | 第n層隱藏層所設置的neuron | [32, 64, 128] | 64 |
| batch\_size | 每次修正的樣本數 | [32, 64] | 32 |
| epochs | 迭代次數 | [ 200, 300] | 200 |
| optimizer | 優化器 | [SGD, Adam] | Adam |

本研究進行Adult分類預測時是使用表3-5-1最優參數後再進行調整Hidden layer之動作。

表3-5-2 Adult分類預測績效表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 4 Hidden layers | | 5 Hidden layers | |
| 績效指標 | Train | Test | Train | Test |
| precision | 0.42 | 0.41 | 0.26 | 0.23 |
| recall | 0.12 | 0.11 | 0.001 | 0.001 |
| F1 | 0.19 | 0.17 | 0.003 | 0.002 |

本研究針對Abalone資料集進行欄位sex的分類預測，以及欄位Rings的數值預測，並透過Talos調整參數，調整結果與範圍如表3-5-3。

表3-5-3 Abalone數值預測Talos所挑選的範圍與最優參數表

|  |  |  |  |
| --- | --- | --- | --- |
| 參數名稱 | 參數說明 | 參數範圍 | 最優參數 |
| first\_neuron | 第一層輸入層所設置的neuron | [32, 64, 128, 256] | 128 |
| hidden\_neuron | 第n層隱藏層所設置的neuron | [32, 64, 128] | 64 |
| batch\_size | 每次修正的樣本數 | [32, 64] | 32 |
| epochs | 迭代次數 | [ 200, 300] | 200 |
| optimizer | 優化器 | [SGD, Adam] | Adam |

本研究進行Adult預測數值時是使用表3-5-4最優參數後再進行調整Hidden layer之動作。

表3-5-4 Abalone預測數值績效表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 2 Hidden layers | | 3 Hidden layers | |
| 績效指標 | Train | Test | Train | Test |
| mae | 1.49 | 1.48 | 1.55 | 1.55 |
| mape | 14.35 | 14.52 | 14.09 | 14.34 |
| rmse | 2.16 | 2.15 | 2.36 | 2.34 |

本研究針對Abalone資料集的分類預測使用Talos調整參數，調整的參數與參數範圍如表3-5-5。

表3-5-5 針對Abalone分類預測使用Talos所挑選的範圍與最優參數表

|  |  |  |  |
| --- | --- | --- | --- |
| 參數名稱 | 參數說明 | 參數範圍 | 最優參數 |
| first\_neuron | 第一層輸入層所設置的neuron | [32, 64, 128, 256] | 128 |
| hidden\_neuron | 第n層隱藏層所設置的neuron | [32, 64, 128] | 64 |
| batch\_size | 每次修正的樣本數 | [32, 64] | 32 |
| epochs | 迭代次數 | [ 200, 300] | 200 |
| optimizer | 優化器 | [SGD, Adam] | Adam |

本研究進行Abalone分類預測時是使用表3-5-6最優參數後再進行調整Hidden layer之動作。

表3-5-6 Abalone分類預測績效表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 3 Hidden layers | | 4 Hidden layers | |
| 績效指標 | Train | Test | Train | Test |
| precision | 0.65 | 0.65 | 0.72 | 0.71 |
| recall | 0.36 | 0.36 | 0.27 | 0.26 |
| F1 | 0.46 | 0.47 | 0.38 | 0.39 |

**第四章、結論**

本研究在Abalone資料集中的分類預測，經過4層hidden layer以及3層hidden layer比較後，發現使用4層hidden layer出來的數據較為優秀。在數值預測中，使用了3層hidden layer以及2層hidden layer做比較，經過實驗後3層hidden layer的績效反而比2層的還要糟糕。而在Adult資料集中的分類預測，以5層和4層hidden layer比較，結果為4層hidden layer較為優秀。經由Talos調整也發現在兩個資料集中，優化器使用Adam與SGD進行比較後，得知Adam的績效都較為優秀。本研究經過上述實驗發現，使用前饋式神經網路預測Abalone與Adult資料集時hidden layer並不是越多越好。

**參考資料:**

Autonomio Talos [Computer software]. (2019). Retrieved from: http://github.com/autonomio/talos.

sklearn.preprocessing.LabelEncoder. Retrieved from: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>

sklearn.metrics. precision\_score. Retrieved from: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html>

sklearn.metrics.recall\_score. Retrieved from: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html#sklearn-metrics-recall-score>

TensorFlow.keras. Retrieved from: <https://keras.io/getting_started/>