國立雲林科技大學資訊管理系

機器學習

Department of Information Management

National Yunlin University of Science and Technology

Machine Learning

專案作業一

Project assignment one

|  |  |
| --- | --- |
| Student Name：  Student ID：  E-mail： | 吳承峻  B10523036  B10523036@yuntech.edu.tw |
| Student Name：  Student ID：  E-mail： | 吳沛錡  B10523006  B10523006@yuntech.edu.tw |
| Student Name：  Student ID：  E-mail： | 紀政廷  B10523012  B10523012@yuntech.edu.tw |
| Student Name：  Student ID：  E-mail： | 林郁凱  B10523032  B10523032@yuntech.edu.tw |

指導教授：許中川博士

Advisor：Chung-Chian Hsu, Ph.D.

中華民國109年4月

April 2020

**摘要**

本研究使用Python語言及Keras開發的ANN模型，對根據人口普查數據進行類別預測與數值預測，分別是:預測收入是否>$50K/年或<=$50K/年，以及預測每週的工作時數。使用UCI machine learning repository中的Adult資料集進行模型建置，此資料集包含的特徵有年齡、工作類別、教育程度、職業、資本收益、資本損失、工作時數、祖國…等等，從這些特徵中，預測收入狀況與可能的工作時數。 經過模型訓練結果，類別預測之結果Accuracy約為”0.84”的準確率，數值預測之結果RMSE約為”0.01”。

1. **緒論**

**1.1動機**

UCI machine learning repository中的Adult資料集內收集了48842位成年人的個人資料，選擇此資料集的原因是此資料集給予了許多相關的基本資料，使本研究能夠透過機器學習來進行更多面相的判別工作時數以及收入是否與一些個人資料有相對應的關聯。

**1.2目的**

透過ANN找出UCI machine learning repository的Adult資料集之最佳模型訓練成果，並能夠準確預測收入是否>$50K/年或<=$50K/年，以及預測每週的工作時數。

1. **方法**
   1. **實作說明**

本研究的實作過程，首先將原始資料做前置處理。透過使用Pandas套件將資料移除缺失值和sklearn.preprocessing的LabelEncoder套件加以編碼，再以sklearn.preprocessing的MinMaxScaler套件將資料進行正規化即完成資料的前置處理。

1. 類別預測方面，收入>$50K/年或<=$50K/年；預測模型的績效衡量指標為Accuracy：

將正規化後的資料給予Keras的models.sequential套件當中，利用Keras的layer套件增加隱藏層，再以RMSprop進行優化。最後以matplotlib.pyplot套件進行性能指標Accuracy評估及輸出圖表，比較training error及generalization error的差異。

1. 數值預測方面，每週工作時數；預測模型的績效衡量指標為RMSE：

將正規化後的資料給予Keras的models.sequential套件當中，利用Keras的layer套件增加隱藏層，再以Adam進行優化。最後以matplotlib.pyplot套件進行性能指標RMSE評估及輸出圖表，比較training error及generalization error的差異。

* 1. **操作說明**

本研究執行環境為 Python 3.6，使用 Jupyter Notebook 開啟 hw1B10523036.ipynb 檔案，然後依序執行程式碼，程式執行過程中會載入 data 資料夾下的所有檔案，並且需要引入 Pandas、numpy、sklearn、Keras matplotlib等套件。

1. **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

表一：Adult資料集欄位介紹

| 欄位 | 欄位名稱 | 內容 |
| --- | --- | --- |
| 0 | age | continuous |
| 1 | workplace | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked |
| 2 | fnlwt | 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 | salary | <=50K,>50K |

表二：顯示部分Adult資料集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 欄位 | 0 | 1 | 2 | 3 | 4 |
| 0 | 39 | 50 | 38 | 53 | 28 |
| 1 | State-gov | Self-emp-not-inc | Private | Private | Private |
| 2 | 77516 | 83311 | 215646 | 234721 | 338409 |
| 3 | Bachelors | Bachelors | HS-grad | 11th | Bachelors |
| 4 | 13 | 13 | 9 | 7 | 13 |
| 5 | Never-married | Married-civ-spouse | Divorced | Married-civ-spouse | Married-civ-spouse |
| 6 | Adm-clerical | Exec-managerial | Handlers-cleaners | Handlers-cleaners | Prof-specialty |
| 7 | Not-in-family | Husband | Not-in-family | Husband | Wife |
| 8 | White | White | White | Black | Black |
| 9 | Male | Male | Male | Male | Female |
| 10 | 2174 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 |
| 12 | 40 | 13 | 40 | 40 | 40 |
| 13 | United-States | United-States | United-States | United-States | Cuba |
| 14 | <=50K | <=50K | <=50K | <=50K | <=50K |

**3.2前置處理**

對Adult資料集進行資料前置處理。先將原始資料Adult.test的第一行去掉，使用Numpy套件的nan將缺失值以NaN代替，再以dropna將擁有NaN的整筆資料移除，即完成處理移除擁有缺失值的資料。接下來將資料集當中的”,”及空白移除，再以sklearn.preprocessing的LabelEncoder套件將種類型屬性值的欄位轉換成label標籤。最後將資料以sklearn.preprocessing的MinMaxScaler套件進行正規化即完成資料的前置處理，分別進行Adult資料集的類別預測及數值預測。

**3.3實驗設計**

|  |  |  |
| --- | --- | --- |
|  | 優化器 | 績效衡量指標 |
| 類別預測方面  >$50K/年或<=$50K/年 | RMSprop | Accuracy |
| 數值預測方面  每週工作時數 | Adam | RMSE |

1. 類別預測方面，收入>$50K/年或<=$50K/年

實驗設計步驟：

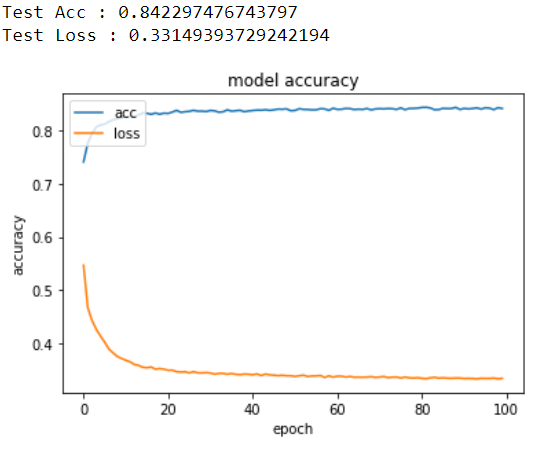
1. 匯入前置處理完的資料
2. 經過6層隱藏層
3. 使用RMSprop優化
4. 將1-3的步驟重複100次且每128筆資料就要修正權重一次。
5. 以Accuracy當作績效衡量指標，比較training error及generalization error的差異。
6. 數值預測方面，每週工作時數

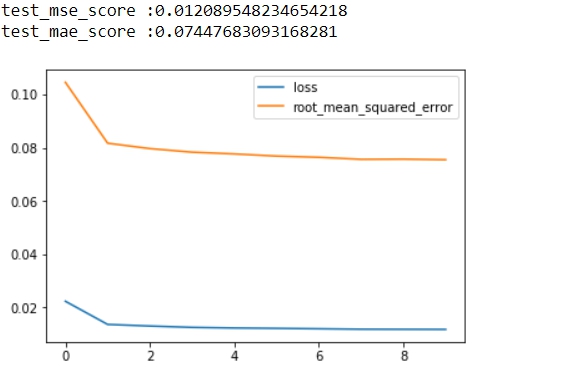
實驗設計步驟：

* 1. 匯入前置處理完的資料
  2. 經過6層隱藏層
  3. 使用Adam優化
  4. 將1-3的步驟重複100次且每128筆資料就要修正權重一次。
  5. 以RMSE當作績效衡量指標，比較training error及generalization error的差異。

**3.4實驗結果** (呈現結果)

* 類別預測方面，收入>$50K/年或<=$50K/年



* ****數值預測方面，每週工作時數

1. **結論**

我們使用類神經網路對UCI ML repository Adult dataset分別進行數值預測與類別預測。我們能夠確認我們的模型根據人口普查數據分別能夠預測收入是否>$50K/年或<=$50K/年，以及預測每週的工作時數。經過本研究的測試發現，若將遺失值運用找尋該欄位出現最多次的資料補齊，將有可能會降低學習效果，因此我們將有缺失值的資料移除，在機器學習的方面我們嘗試了各種不同的組合，最終我們選用六層的學習架構，並且其中包含了一層Dropout，而在每一層的寬度我們選用了64、32以及16，如此一來得出本研究最佳的結果。