機器學習專案作業二

摘要

近年來，平均學歷已達新高峰。為了研究學歷對個人的影響。本專案使用個人生活情況預測學歷，進而找出社會對學歷的期待。實驗流程先將各屬性轉為數值、將正確解答轉為獨熱編碼。並製作前饋式神經網路作為分類器，以個人生活情況為輸入，輸出預測學歷。最終模型預測的正確率為 0.4083。

壹、 緒論

1.1. 動機

近年來，平均學歷已達新高峰。在教育普及的同時，也因為高學歷人才增加，造成學歷價值變動。為了研究學歷對個人的影響，本專案使用個人之生活地區、婚姻、種族、性別、家庭、職業、工作機關預測學歷，目的是找出各屬性與學歷是否存在關聯，藉此能夠推斷社會對於學歷的期待。

1.2. 目的

探討是否因為所在地、種族、性別不同影響個人對學歷的要求，以及學歷是否影響個人之婚姻、家庭關係、職業、工作機關。

貳、 方法

首先執行 adult-preprocess 將資料欄位之文字標記為數值類別並存成 csv 檔案，再執行 adult 使用 tensorflow 提供之方法，將數值類別作為分類標籤輸入模型，並將學歷標籤轉為獨熱編碼。檔案皆以 jupyter 執行。

參、 實驗

3.1 資料集

本專案使用 adult 資料集，由 Barry Becker 從 1994 年的人口普查數據庫中提取，此研究使用 native-country, marital-status, race, sex, relationship, occupation, workclass 欄位預測 education-num，欄位皆為數值類別。使用之資料筆數分別為：訓練集 26049 筆、驗證集 6512 筆、測試集 16281 筆。

部分資料內容：

Native-country	Marital-status	race	Sex	Relationship	occupation	workclass	Education-num
1	1	1	1	1	1	1	13
1	2	1	1	2	2	2	13
1	3	1	1	1	3	3	9
1	2	2	1	2	3	3	7
2	2	2	2	3	4	3	13

3.2 前置處理

由於模型必須輸入數值資料，我們將欄位字串處理成數值類別，並將學歷之數值類別處理成獨熱編碼。

3.3 實驗設計

第一次實驗使用一層隱藏層，寬度為 32、激活函數為 relu、輸出層激活函數使用 softmax、優化器使用 adam、batch_size 為 32，epoch 為 200，並使用 callback 機制防止過擬合。測試集的準確度有 0.4029。第二次實驗增加隱藏層的深度到兩層，寬度分別是 16、32，激活函數皆使用 relu，輸出層激活函數使用 softmax。測試集的準確度有 0.4056。

3.4 實驗結果

使用訓練集評估模型，多類別分類器的結果是 loss: 1.6788 - accuracy: 0.4170 - f1_m: 0.1507 - precision_m: 0.5526 - recall_m: 0.0891。

使用測試集評估模型，多類別分類器的結果是 loss: 1.7547 - accuracy: 0.4056 - f1_m: 0.1425 - precision_m: 0.5070 - recall_m: 0.0849。

由結果可以判斷，個人之學歷與屬性存在關聯性。

肆、 結論

由這次實驗的結果得知，個人學歷與生活地區、婚姻、種族、性別、家庭、職業、工作機關存在關聯性。但可以藉由個人的生活情況預測學歷，合理推論學歷會影響個人生活。

伍、 參考文獻

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

https://keras.io/examples/structured_data/structured_data_classification_from_scratch/