【智慧企業整合】

Project 3

困惑學生的腦電圖預測

110034563 許宇韶

指導教授：邱銘傳 教授

中華民國 111 年 1 月 7 日

目錄

圖目錄

表目錄

一、 背景說明

(一)、現況描述

每個大學生都難免會面臨課業壓力、人際壓力、感情壓力或家庭的壓力，通常課業壓力被認為是大學生的主要壓力來源，不同科系面臨的課業壓力可能有異。其中包含了因為無法適應教師的上課方式與不敢詢問教師有關課業上的問題而造成學習障礙；而學習方法欠佳，不知如何計劃，還有學習能力的不足，則容易造成挫折感。

(二)、問題描述 5W1H

我們的目標是建立一個深度學習模型，該模型可以透過腦波儀收集學生腦波資料後，分析學生對於這些視頻可能會讓普通大學生感到困惑。
其中 10 位大學生觀看視頻剪輯時的 EEG 信號數據。假定提取了不會讓大學生感到困惑的在線教育視頻，例如介紹基本代數或幾何的視頻。還準備了一些視頻，如果學生不熟悉量子力學和幹細胞研究等視頻主題，這些視頻可能會讓普通大學生感到困惑。

表 1. 5W1H

What	困惑學生的腦電圖預測
When	學習時間
Who	學生
Where	教室
Why	協助對於學習有困難的學生從而落實快樂學習
How	LSTM、LogisticRegression

二、 資料介紹、處理

(一)、資料內容

由 Kaggle 公開數據集中取得 Confused Student EEG prediction 的資料集，資料集共分成 2 種類別：

總共包含 10 個視頻，每個視頻時長約 2 分鐘。
學生們佩戴 MindSet 來測量額葉的活動。MindSet 測量放置在額頭上的

電極和分別與耳朵接觸的兩個電極（一個接地和一個參考）之間的電壓。

每次課程結束後，學生按照 1-7 的等級評定他/她的困惑程度，其中 1 對應於最不困惑，7 對應於最易混淆。這些標籤如果進一步規範化為學生是否感到困惑的標籤。除了我們預定義的混淆標籤之外，此標籤還作為自我標記的混淆提供。

這些數據來自十名學生，每人觀看十個視頻。因此，對於這 12000 行，可以看作只有 100 個數據點。如果你這樣看，那麼每個數據點由 120 多行組成，每 0.5 秒採樣一次（所以每個數據點都是一分鐘的視頻）。頻率較高的信號報告為每 0.5 秒的平均值。

EEG_data.csv : 包含 10 名學生記錄的 EEG 數據
人口統計.csv：包含每個學生的人口統計信息
視頻數據：每個視頻大約持續兩分鐘，我們去掉前 30 秒和後 30 秒，只收集中間 1 分鐘的腦電數據。

(二)、資料處理

- get_dummies 函數對 ethnicity & gender 進行 One hot encoding 編碼
- 一些 EEG 特徵的範圍很廣，因此我使用標準化

(三)、資料分析

Pandas Profiling 是一個非常強大的開放原始碼套件，可以使用最少的程式碼快速實現探索式資料分析(EDA)，並且透過報表提供的統計數據和視覺化圖表，能夠幫助資料分析人員對於陌生資料集的有效分析和探索，非常值得列為資料分析的工具之一。因此本次使 pandas_profiling 探索資料集特徵包含統計資訊、特徵分布的情況以及特徵之間的關係。提供資料分析人員快速瀏覽資料集的變數個數、遺失值比率、重複值比率與變數型態等

顯示資料集的各個欄位變數統計資訊，除此之外，還可以點擊右下角的「Toggle details」按鈕查看更詳細的欄位資訊。

在 warnings 頁籤則會指出哪些欄位是 High cardinality(多唯一值的)、High correlation(高相關性的)及 Missing(遺失值比率較高)

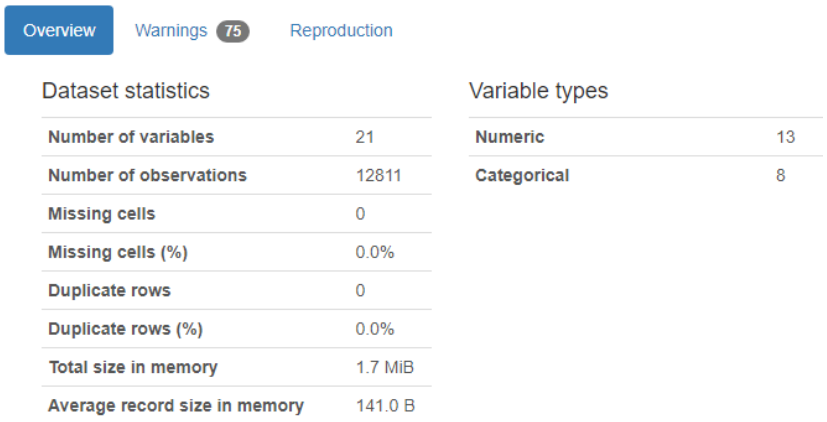

圖 1. 特徵 Overview

顯示資料集的各個欄位變數簡單的統計資訊

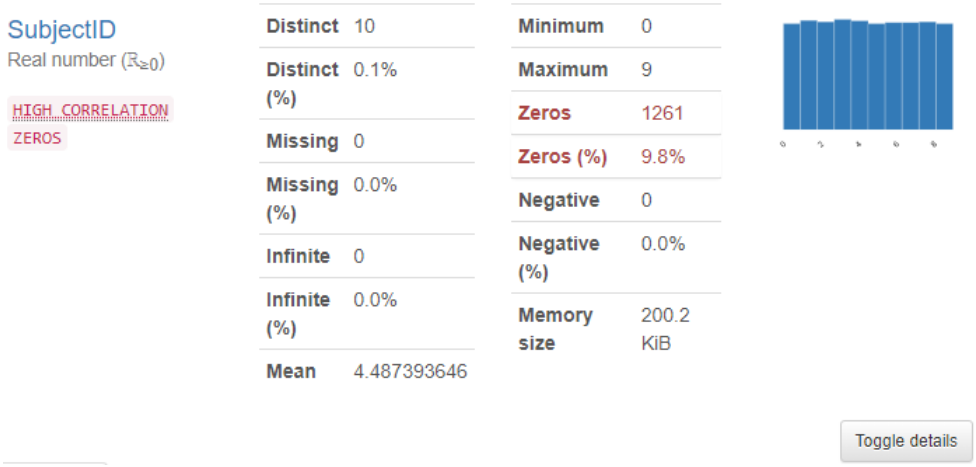

圖 2. 特徵統計資訊-1

除此之外，呈上圖點擊右下角的「Toggle details」按鈕查看更詳細的欄位資訊我們可以得到特徵的更多統計資訊包括

- 最大最小值
- 全距
- IQR
- 平均數
- 標準差
- 變異數等

Quantile statistics

Minimum	0
5-th percentile	0
Q1	2
median	4
Q3	7
95-th percentile	9
Maximum	9
Range	9
Interquartile range (IQR)	5

Descriptive statistics

Standard deviation	2.86537261
Coefficient of variation (CV)	0.6385382776
Kurtosis	-1.222148798
Mean	4.487393646
Median Absolute Deviation (MAD)	2
Skewness	0.009631892513
Sum	57488
Variance	8.210360193
Monotonicity	Increasing

圖 3. 特徵統計資訊-2

透過切換頁籤的方式，來瞭解高相關性的不同欄位之間交互關係

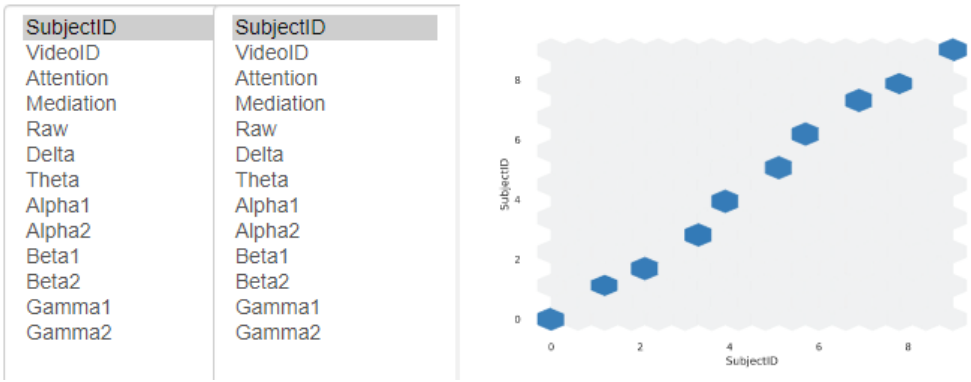

圖 4. 特徵交互關係

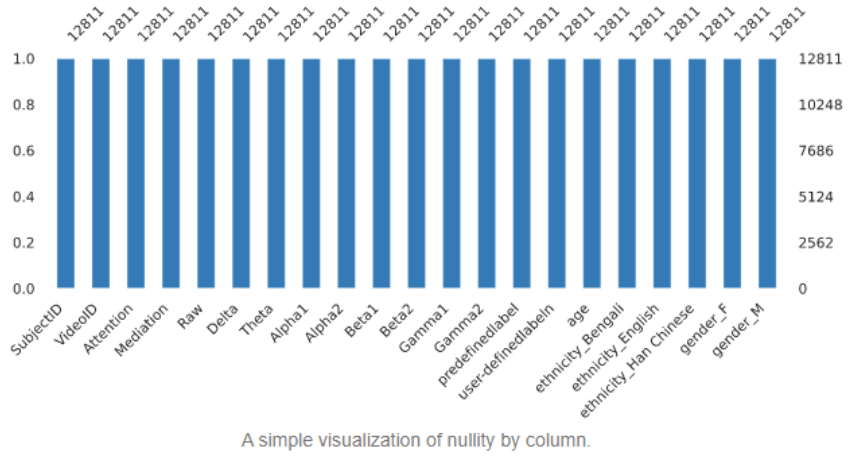

圖 5. 顯示遺失值的個數

透過熱力圖發現以下特徵有高度正相關性

- VideoID & Predefinedlabel
- Beta2 & Gamma1
- Gamma1 & Gamma2

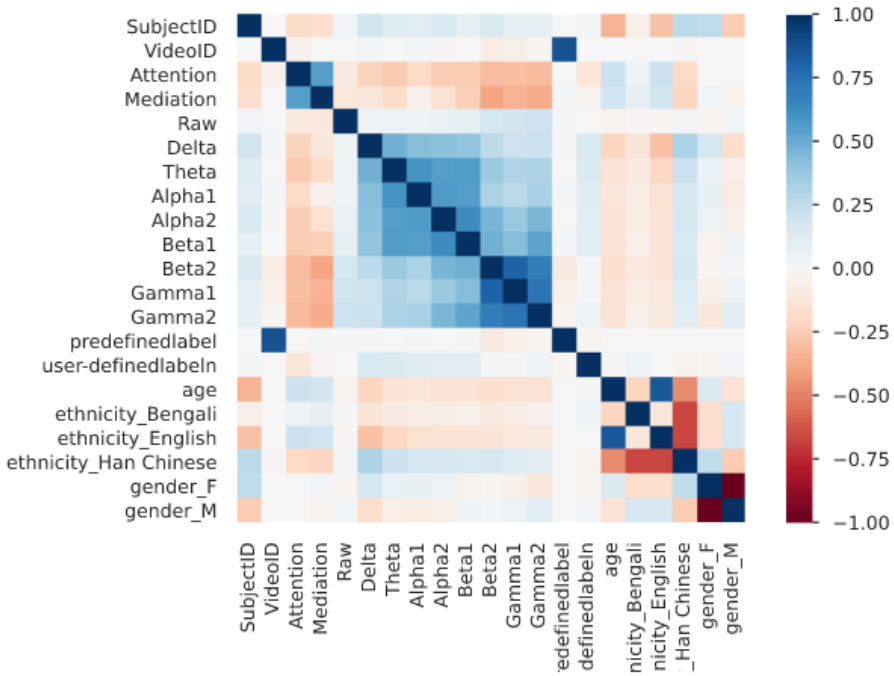

圖 6. 熱力圖

（四）、特徵篩選

- 在數據集的描述中提到，具有 VideoID 和 SubjectID 等功能。SubjectID 和 VideoID 怕可 能將阻礙模型訓練，因為 10 個學生有 10 個剪輯。所以模型很可能會根據 ID 進行學習，但我希望它根據 EEG 記錄、種族以及性別和年齡參數進行學習。

- 利用 mutual_info_classif 進行分離特徵集和目標列給出每個特徵的分數，描述目標與特徵間的關係，可以看到 gender 與 ethnicity 對目標是沒有任何影響力，因此嘗試把兩項特徵拿掉。

Index	0
VideoID	6.77143
Alpha2	2.54267
Delta	2.44771
Gamma1	2.29428
Theta	2.14002
Beta1	2.09874
Alpha1	2.03632
Gamma2	1.88297
Raw	1.57544
Beta2	1.43891
Attention	1.06214
Mediation	0.543061
predefinedlabel	0.509191
SubjectID	0.381945
age	0.195931
ethnicity_Bengali	0
ethnicity_English	0
ethnicity_Han Chinese	0
gender_F	0
gender_M	0

圖 7. 特徵篩選

三、 模型介紹

（一）、Logistic Regression

logistic regression 是迴歸分析的一種，但與一般線性迴歸的目標(Y)須為連續型變數不同，Logistic Regression 的目標(Y)型態是類別，若是類別只有兩個，則為二元的邏輯式迴歸(Binary logistic regression)，若是類別超過三 個以上則為 Polytomous logistic

regression，model 相對複雜許多，本次使用 Binary logistic regression，二元分類為有困惑和沒困惑，困惑設為 1，沒有設為 0。為了方便結果的解釋與理解，一般來說我們會將依變項為 0 設為參照組。自變項(X)可為類別變數或連 續變數，用來討論對依變項(Y)的關係。

(二)、LSTM

長短期記憶（LSTM）是一種特殊的 RNN 模型。與 RNN 相比， LSTM 可以更好地處理長期順序數據，並通過存儲功能解決長期依賴的 問題。與常規 RNN 相比，LSTM 更為複雜。它還具有三個控制門，即 輸入門，輸出門和忘記門。當將值寫入存儲單元時，它具有一個存儲單 元，必須通過輸入門並且只有在門打開時才能寫入該值。至於輸入門是 否打開，LSTM 會自己學習，輸出門則控制是否可以讀取存儲單元的值， 而遺忘門決定何時應清除存儲單元的值。

```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 17, 1)]           0
_________________________________________________________________
dense (Dense)                (None, 17, 64)            128
_________________________________________________________________
bidirectional (Bidirectional (None, 17, 512)           657408
_________________________________________________________________
dropout (Dropout)            (None, 17, 512)           0
_________________________________________________________________
bidirectional_1 (Bidirection (None, 17, 256)           656384
_________________________________________________________________
dropout_1 (Dropout)          (None, 17, 256)           0
_________________________________________________________________
flatten (Flatten)            (None, 4352)              0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               557184
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 129
=================================================================
Total params: 1,871,233
Trainable params: 1,871,233
Non-trainable params: 0
_________________________________________________________________
```

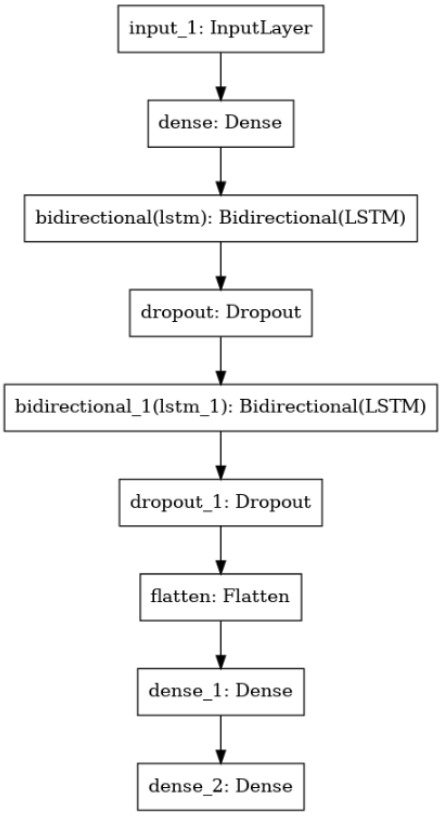

圖 8. 模型架構

四、 訓練過程

(一)、Logistic Regression

表 2. Logistic Regression accuracy

Drop SubjectID & VideoID	0.5298478345688646
Drop age & gender & ethnicity	0.5306281701131487
Drop SubjectID & VideoID & age & gender & ethnicity	0.5645727662895045

(二)、LSTM

使用的損失函數 Binary Cross Entropy，並使用 Early_Stopping 的回調函數來避免過度擬合，並使用 lr_scheduler 來改變模型訓練時的學習率。從 learning_rate = 0.001 和 batch_size = 20 開始訓練 100 個 epochs。

Drop SubjectID & VideoID

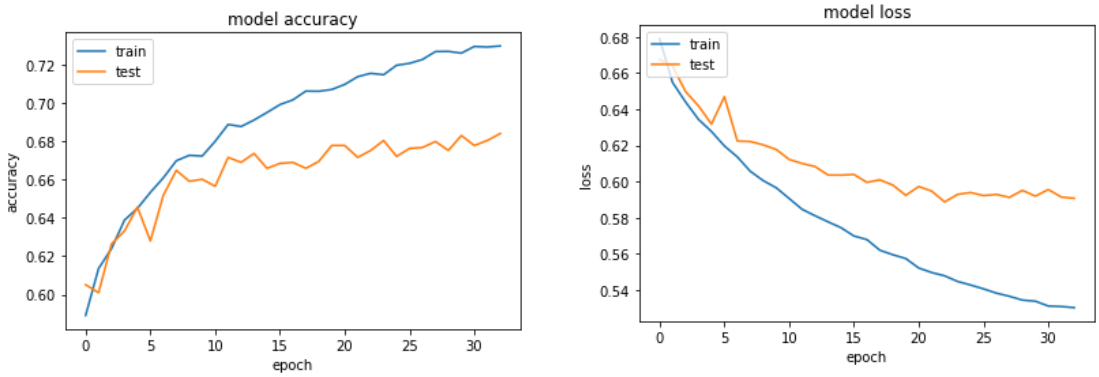

圖 9. Drop SubjectID & VideoID Accuracy(左) & Loss(右)

Drop SubjectID & VideoID age & gender & ethnicity

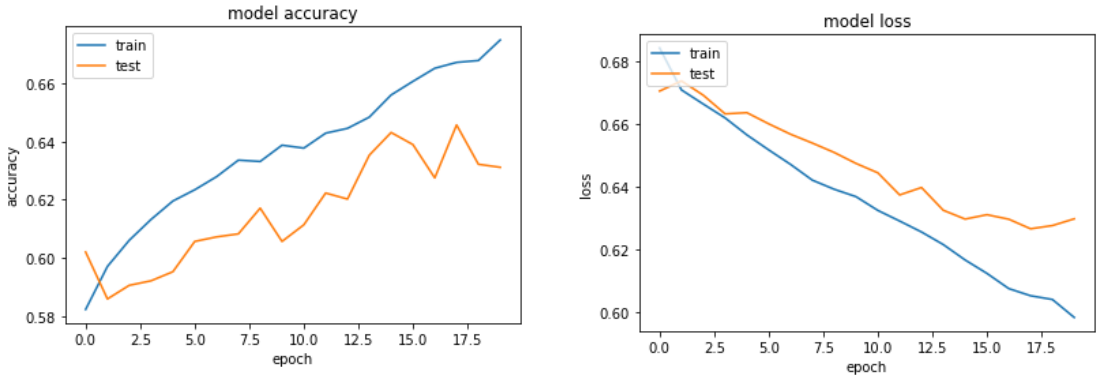

圖 10. Drop SubjectID & VideoID age & gender & ethnicity Accuracy（左）& Loss(右)

五、 結論與展望

(一)、結論

Drop age & gender & ethnicity

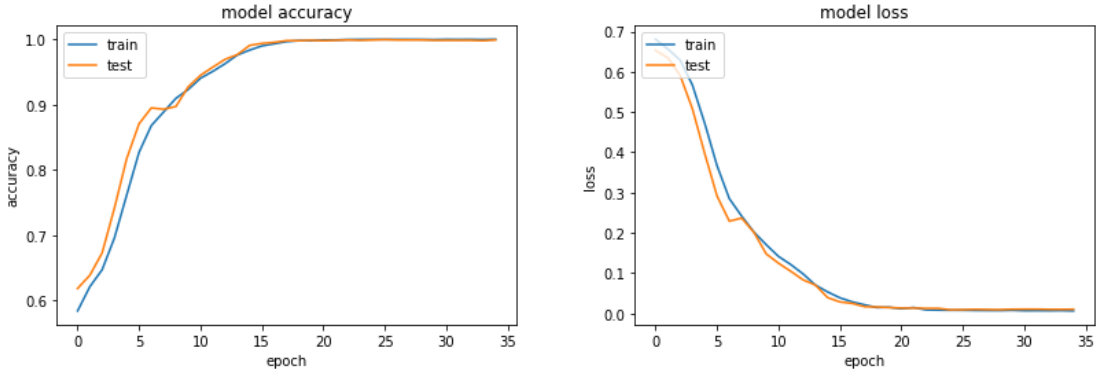

圖 11. Drop age & gender & ethnicity Accuracy（左）& Loss(右)

	precision	recall	f1-score	support
0.0	0.99	1.00	1.00	974
1.0	1.00	0.99	1.00	948
accuracy			1.00	1922
macro avg	1.00	1.00	1.00	1922
weighted avg	1.00	1.00	1.00	1922

圖 12. Drop age & gender & ethnicity test result

總結以上，經由資料前處理(包含資料相關分析與資料視覺化)等手法找出特徵值作為訓練模型變數，並配合資料轉換與標準化，最終挑選出 14 項變數作為訓練模型之因子，而透過 keras 深度學習模型建立，並透過 Early_Stopping 的回調函數來避免過度擬合，並使用 lr_scheduler 來改變模型訓練時的學習率大幅減少超參數調整所耗費的時間，最終得出之訓練模型與其他機器學習相比有著最好的表現。

(二)、展望

資源班是指學校針對有學習障礙的學生所提供的支援服務，需要特別申請，同時經過專業鑑定後才能進資源班。家長可能會擔心師資或外界眼光，但不能因此就無視於孩子需求把他留在普通班。

資源班（亦有稱作資源教室），是一種普通中小學設置的教學環境。學生具有各類輕度身心障礙的程度，但是有意願在普通校學校就讀，就依其意願安置在普通學校，在校大部分時間在普通班級上課，抽出部分時間到資源班接受彈性化、個別化及功能性教學，根據兒童的能力及特殊需要來施以個別化教育方案，使學生在普通學校亦能獲得適當資源的輔導及幫助。

透過本次實驗我們可以利用腦波檢測學生於學習中是否遇到困難，再根據腦波儀檢測協判斷學生是否需額外的輔導功課，讓小朋友能夠知道同儕之間是因為學習上有困難，所以需要資源班的服務，同時協助專家鑑定是否真的需要進入資源班，減少資源班在使學生受阻能力能夠排除障礙，順利發展。避免不關注成長和發展以及老師們即使辛勤付出，教學效果卻不盡理想的結果。

六、 參考資料

- https://www.kaggle.com/shreyaspj/confused-student-eeg-prediction/notebook
- https://www.kaggle.com/deepak915/are-they-confused-99-5-accuracy#Scaling-our-Feature-set
- https://class.kh.edu.tw/1188/page/view/15