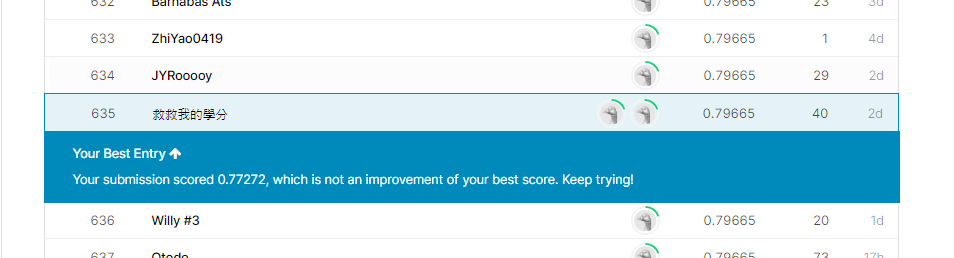
**深度學習介紹 project1**

**Team name**：救救我的學分

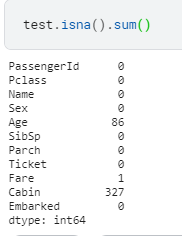
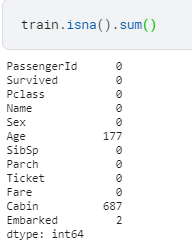
**Team member**：邱以中、歐亭昀

**Leader board** ：

**程式碼：**[**https://www.kaggle.com/fallantbell/titanic-mlp-with-age-prediction**](https://www.kaggle.com/fallantbell/titanic-mlp-with-age-prediction)

**資料前處理**

1. **填補缺失資料：**

經過觀察，我們可以發現train, test data當中都有很多缺失的資料，

因此，如何處理這些缺失的資料將會很大的影響我們預測的結果！

|  |  |
| --- | --- |
| 遺失資料 | 解決方法 |
| Cabin | 由於Cabin缺失的資料太多，因此在預測的過程中不考慮此項目。  ( train data 缺失687筆(77%)，test data 缺失327筆(78%) ) |
| Embarked | 因為只有缺失2個，所以選擇用數量最多的 ’S’ 來代替。 |
| Fare | 考慮擁有相同條件(Pclass, ticket, Embarked)的fare，選擇其中的中位數 |
| Age | train data 缺失177筆(20%)，test data 缺失86筆(20%)。參考許多top的文件後，看出age是會影響我們判斷結果的重要特徵。因此不依照常見方法以中位數填入，我們嘗試將age用模型預測的方式做填補，其中利用 Name、Fare、sibling … 等項目作為訓練特徵，放入隨機森林模型來預測缺失的年齡，最後並將Age做離散化，設為Age-bin存回training data中。 |

1. **資料內容處理**

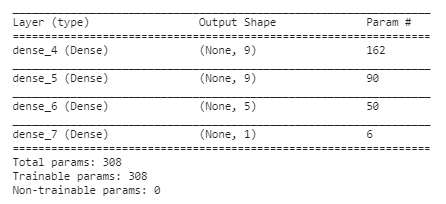
在參考別人的預測結果發現，他們都將male都預測為無法存活。所以我們為了能準確預測male是否能存活，便能以將資料做以下處理。

|  |  |
| --- | --- |
| 資料處理 | 方法 |
| Name | 我們依照乘客名字分為Miss, Mrs, Master, Mr, Others五項類別，並以 one-hot encoding的方式將這五類作為特徵存回data中。 |
| Fare | 因避免因數字大小差異影響訓練，所以fare離散化，依照大小五分位數分類分成[1-5]間，並設為Fare-bin存為data中。 |
| Ticket | 因ticket是數字及文字交雜的資料，因此我們取內容第一個字元當代表，if 字元為文字且不為A,W,F,L則填入4 else 則不便。因此ticket的內容為一個字元。 |
| SibSp & Parch | 從上面的圖可以看出，乘客如果是20-36歲的成人male、家族人數<=2、Pclass=1，竟有50%(18/32)的存活率。所以我們將SibSp +Parch+1的結果設為Family，並當作重要的特徵資訊放入data中。 |

最後，我們將以代替的項目 SibSp, Parch, Age, Fare, Title 清除後，我們的data 特徵有 PassengeID, Survived, Pclass, Sex, Master, Miss, Mr, Mrs, Others, Family, Fare-bin, Age-bin,Ticket1-4,C,P,S 。

**Model**

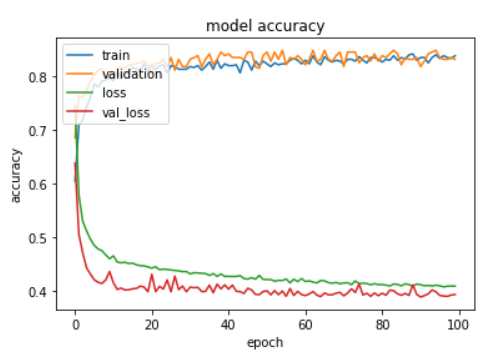
我們分別對以下四項做GridSearch 的比較，並取的表現最好的結果當作最終模型。

1. optimizer ( SGD, Adagrad, Adadelta, Adam, Nadam )
2. activation ( softmax, softplus, relu, tanh, selu, sigmoid )
3. layer number
4. kernel\_init(random\_normal, he\_normal)

*structure*

我們選擇 keras 的sequential model使用了3層 hidden layer + 1層output layer (如圖)

*Activation function*

**前3層活化函數使用 selu, 最後一層使用 sigmoid

*Loss function*

結果只有0 or 1, 因此使用 binary\_crossentropy

*Optimizers* - 使用 SGD

*Kernel\_init -* 使用 random\_normal

**結論 & 想法**

Prediction

Prediction

Model1

No age

Model2

Have age

1. 因為age為重要資訊同時又有部分資料缺失，所

以我們有想到將data以是否有age分開做模型訓練

(右圖)，但訓練結果不如預期。

1. 我們也有用Drop的方式訓練模型，因準確率一

直落在約77%左右，所以我們改用前述所用的模型

**工作分配:**

邱以中: 資料前處理、model 訓練+參數優化、報告撰寫

歐亭昀: 資料前處理、model 訓練+參數優化、報告撰寫