團隊測驗報告

報名序號：110087

團隊名稱：教授哩底隊

註1：請用本PowerPoint 文件撰寫團隊程式說明，**請轉成PDF檔案繳交**。
註2：依據競賽須知第七條，第4項規定：
測試報告之簡報資料不得出現企業、學校系所標誌、提及企業名稱、學校系所、教授姓名及任何可供辨識參賽團隊組織或個人身分的資料或資訊，違者取消參賽資格或由評審會議決議處理方式。

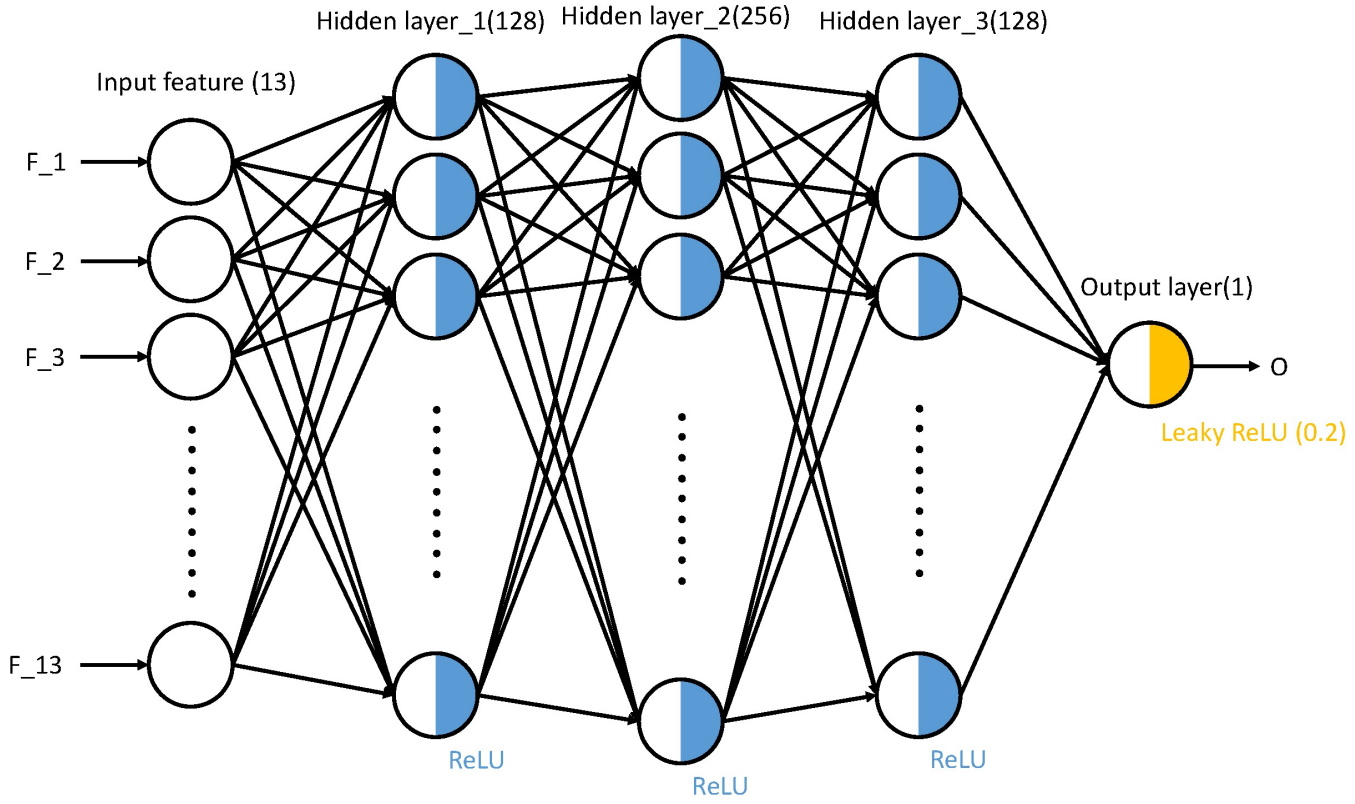
一、資料前處理

- 資料預處理
  - ➢目的：整理重複數據資料
  - ➢檔名：dataset.py
  - ➢自定義函式toIndependent_csv( )，將數據集中F_1到F_13欄位相同但O欄位輸出資料不同的訓練資料做處理，使得每筆訓練資料皆不重複
  - ➢O欄位輸出處理可為平均數(mean)、中位數(median)或眾數(mode)

- 資料標準化(Standardization/ Z-score normalization)
  - ➢目的：轉換數據資料結構以加速收斂
  - ➢檔名：csv_utils.py
  - ➢將清理過後的資料利用sklearn內建函式StanderScaler( )與fit_transform( )，使得所有數據的平均為0、標準差為1，並將資料轉換成符合運算結構的張量(tensor)

二、演算法和模型介紹(P.1)

- 演算法介紹
  - ➢使用Multi-Layer Perceptron(MLP)演算法
  - ➢並利用Model ensemble來增加訓練的預測表現



MLP模型架構圖 →

二、演算法和模型介紹(P.2)

- 模型介紹與優化
  1. 使用 Multi-Layer Perceptron(MLP) 模型架構，並建立三層隱藏層
     ① 檔名：model.py
     ② 輸入層節點數：13 (資料特徵值F_1~F_13)
     ③ 第一隱藏層節點數：128
     ④ 第二隱藏層節點數：256
     ⑤ 第三隱藏層節點數：128 or 256
     ⑥ 輸出層節點數：1 (預測O值)
  2. 使用 Adam(Adaptive Moment Estimation) 學習優化器
     ① 目的：用於優化類神經網路中的權重與偏差值
     ② 優化器建立在train.py檔之中

二、演算法和模型介紹(P.3)

- 模型評估與收斂

1. 建立損失函數(Loss Function)
   ① 目的：評估預測結果與真實值間之差異
   ② 使用均方誤差(Mean Squared Error, MSE)來建立函數
   ③ 損失函數建立在train.py檔之中
2. 使用Gradual Warm-up策略 (訓練初期)
   ① 目的：起初訓練的Loss值通常都很大，而我們為了避免資料的過擬合(Overfitting)，所以建立Warm-up策略，而Gradual Warm-up是用漸進式的方式由低到高增加Learning rate來訓練每個Epoch。此機制不僅能避免Learning rate突然增加，也可以幫助訓練初期能有較好的收斂
   ② Gradual Warm-up策略建立在train.py檔之中

二、演算法和模型介紹(P.4)

- 模型評估與收斂

3. 使用Cosine Annealing策略 (訓練中後期)

① 目的：訓練後期容易因為過大的Learning rate使得模型陷入局部最小值(Local minimum)而影響效能，利用Cosine Annealing漸進式的方式由高到低減少Learning rate來訓練後期的模型，幫助找到全局最小值(Global minimum)

② Cosine Annealing策略建立在train.py檔之中

二、演算法和模型介紹(P.5)

- Dataloaders -- PyTorch中用於讀取數據的接口
    1. 透過資料前處理後，所篩選出的資料筆數共 21,766 筆
    2. 為了避免篩選過後的資料過擬合(Overfitting)，我們利用sklearn內建函式 train_test_split ( ) ，透過設定超參數訓練檔(training.yaml)中的 VAL_RATE 參數來劃分訓練資料與驗證資料的比重
        ① VAL_RATE：0.8(or 0.9)
        ② 訓練資料(train_loader)：17,412(or 19,589) 筆資料
        ③ 驗證資料(val_loader)：4,354(or 2,177) 筆資料

三、預測結果(P.1)

- 在進行預測後，以下有幾種較佳之模型設定：

模型編號	輸入層節點數	隱藏層層數	各隱藏層節點數	VAL_RATE參數設定	資料前處理(O欄位)	輸出層節點數
1	13	3	128-256-128	0.8	O欄位輸出處理使用**平均數**	1
2	13	3	128-256-128	0.8	O欄位輸出處理使用**中位數**	1
3	13	3	128-256-128	0.9	O欄位輸出處理使用**平均數**	1
4	13	5	128-256-256-256-128	0.8	O欄位輸出處理使用**平均數**	1

三、預測結果(P.2)

- 在演算法的介紹當中，我們提及利用Model ensemble來增加訓練的預測表現，所以我們將四種模型所預測的分數加以平均作為我們的預測值結果。

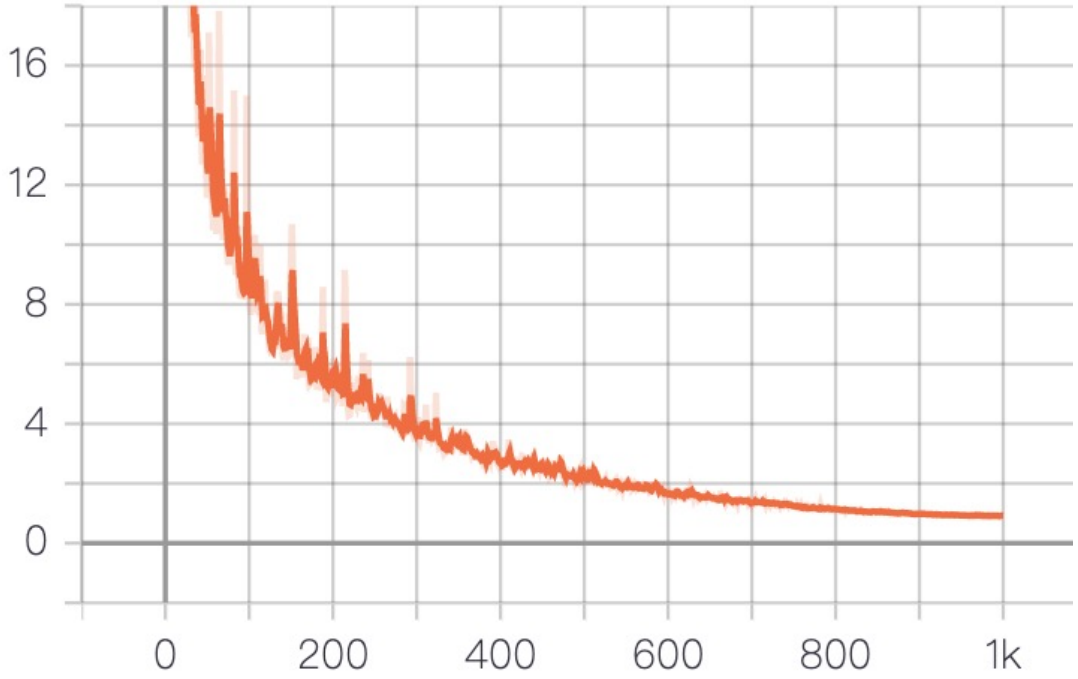
Seq	MLP_13-128-256-128	MLP_13-128-256-128	MLP_13-128-256-128	MLP_13-128-256-256-256-128	Average Output
1	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
2	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
3	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
4	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
5	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
6	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
7	2.819832325	-0.026237428	0.512207687	0.970664978	1.069116891
8	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
9	2.819832325	-0.026237428	0.512207687	0.970664978	1.069116891
10	2.819832325	-0.026237428	0.512207687	0.970664978	1.069116891
11	2.819832325	-0.026237428	0.512207687	0.970664978	1.069116891
12	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
13	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
14	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
15	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
16	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
17	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
18	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
19	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375
20	2.389365196	-0.010343552	-0.004091883	0.829967737	0.801224375

註：資料筆數眾多僅擷取部分示意

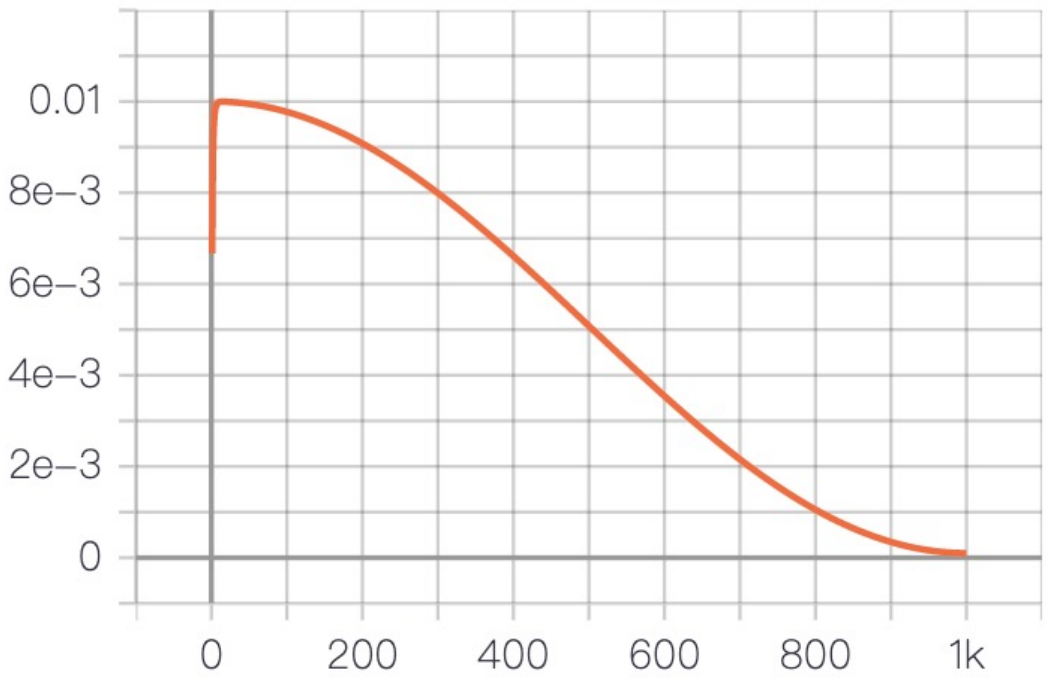
四、Tensorboard訓練圖示(P.1)

A.訓練資料(Train)

註：Epoch數為1,000



(a) Loss function

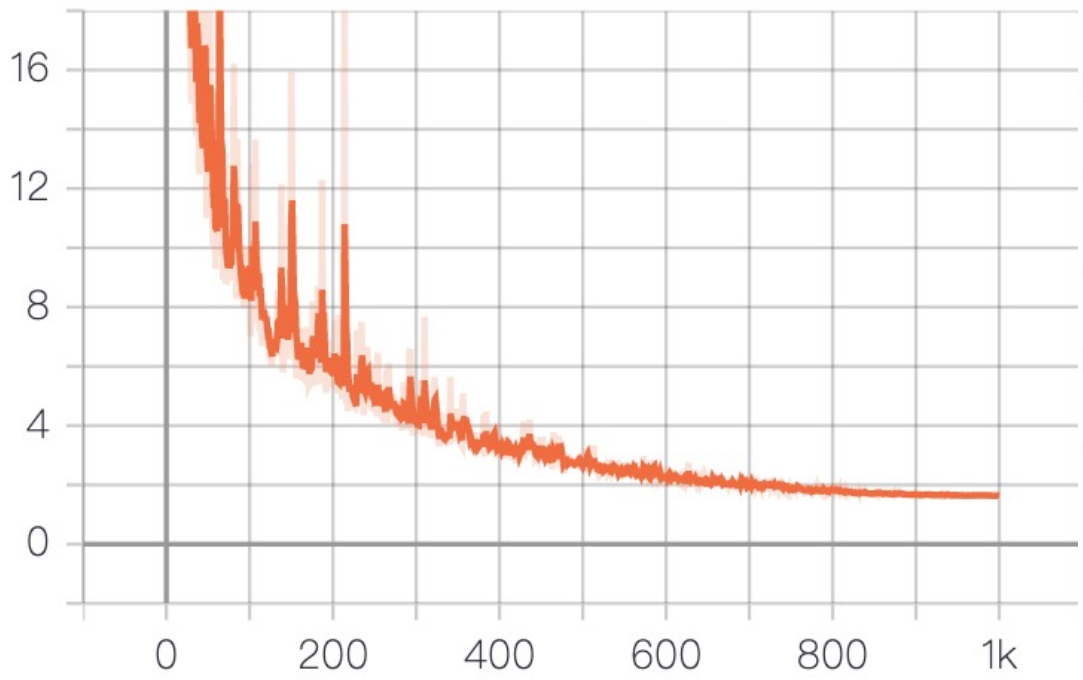


(b) Learning rate

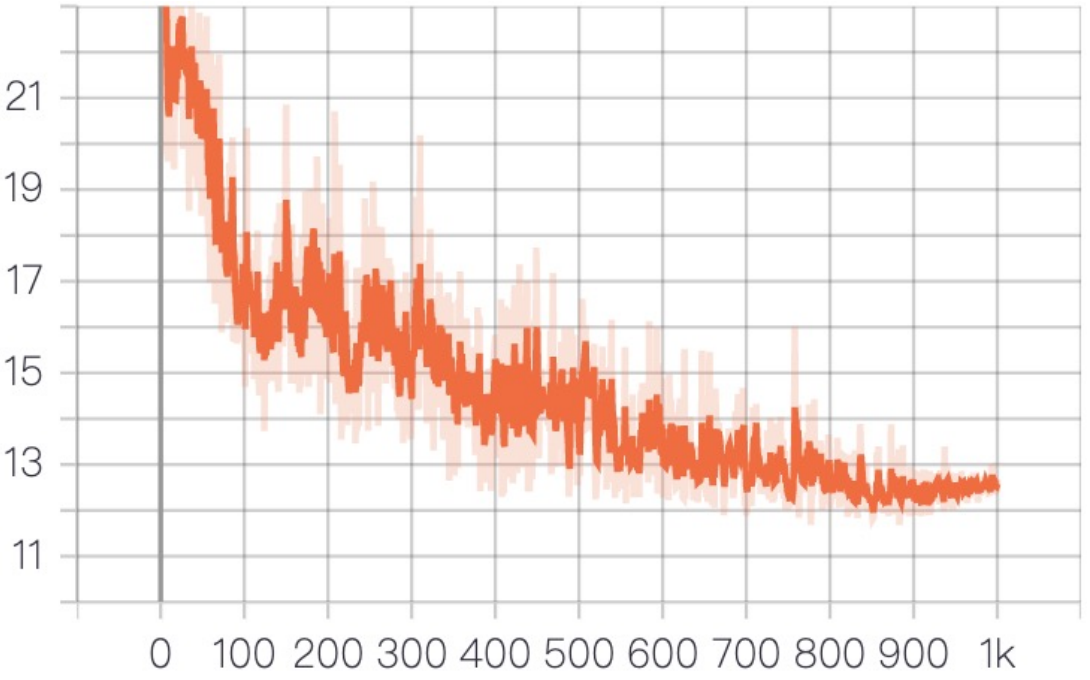
四、Tensorboard訓練圖示(P.2)

B.驗證資料(Validation)

註：Epoch數為1,000



(a) Loss function

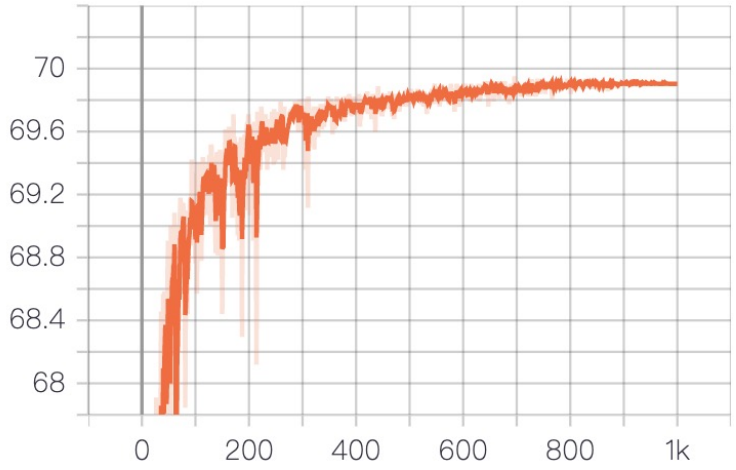


(b) y值, $y = \text{Max}(|O_{\text{目標值}} - O_{\text{預測值}}|)$

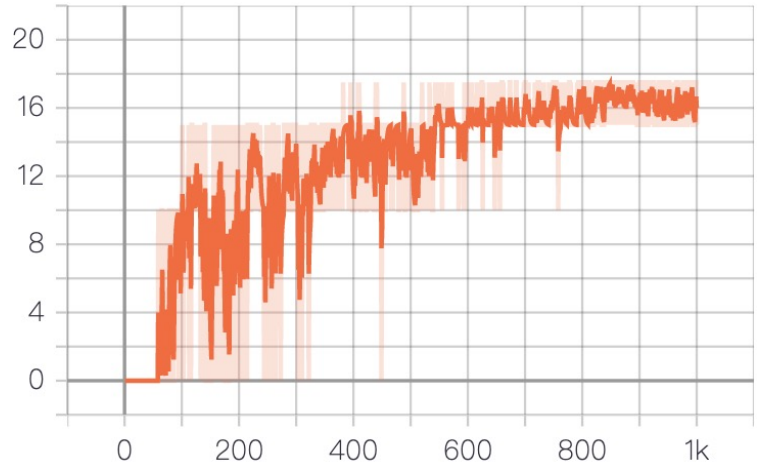
四、Tensorboard訓練圖示(P.3)

B.驗證資料(Validation)

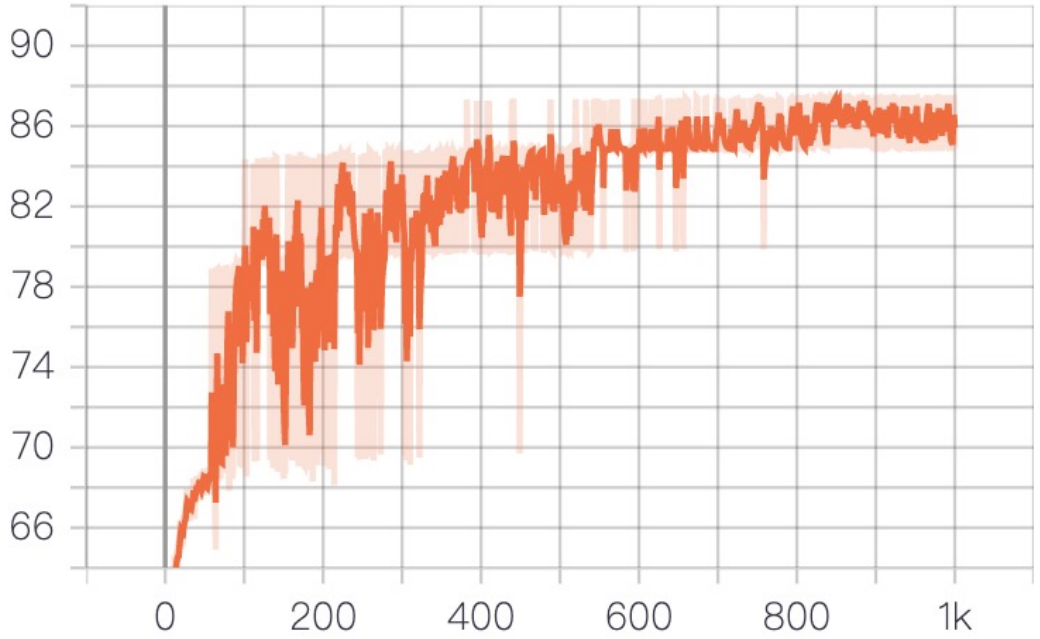
(a) Score A



(b) Score B



(c) Total Score



註：Epoch數為1,000