Homework 2 Report - Income Prediction

學號：r06921081 系級：電機碩一 姓名：張邵瑀

1. **(1%) 請比較你實作的generative model、logistic regression的準確率，何者較佳？**

   1) Accuracy (kaggle)

	Pulic	Private	平均
Generative model	0.85171	0.84621	0.84896
Logistic regression	0.85724	0.84805	0.85264

   2) Logistic regression 較佳，與老師上課的討論結果相符。

2. **(1%) 請說明你實作的best model，其訓練方式和準確率為何？**

   1) 訓練方式

   使用 logistic regression,optimizer 使用 adagrad，動態調整learning rate 讓他有一定的隨機性，使得他有能力在收斂的時候有機會跳出local minimum並紀錄最高accuracy的weight(因為會不時晃動)，features 則是挑選助教處理過的資料中第1,2,3,4,5,6,7,8,9,11,12,13,14,15,16,17,18,19,20,32,35,36,38,41,43,44,45,49,50,51,52,53,54,56,57,58,59,60,61,62,63,64,67,68,69,70,73,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,102,106,107,120,項並把(age, capital_gain, capital_loss)取高次方項(2~7次)。

   2) Accuracy

Pulic	Private	平均
0.86154	0.85800	0.85977

3. **(1%) 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。**

   1) Accuracy (kaggle)

	Pulic	Private	平均
Normalization	0.86154	0.85800	0.85977
Without Normalization	0.82569	0.83120	0.82844

2) 討論

沒有 Normalization 的部份準確度明顯很低，而且在訓練的過程中，validation set的accuracy 振動相當大，而且訓練的收斂速度較慢。

4. (1%) 請實作logistic regression的正規化(regularization)，並討論其對於你的模型準確率的影響。

1) Accuracy (kaggle)

λ	Pulic	Private	平均
0.1	0.76268	0.76461	0.76364
0.01	0.78430	0.78771	0.78600
0.001	0.83134	0.83280	0.83207
0.0001	0.82827	0.82997	0.82912

2) 討論

可能是因為有做 early stopping 的關係,可以發現 regularization 用在自身的 model 沒有幫助,甚至 λ 太大,還會出現 underfitting 的現象。

5. (1%) 請討論你認為哪個attribute對結果影響最大？

Capital gain 對結果影響最大，因為在normalize 後，他的權重還是很大，遠大於其他feature，雖然 Capital gain 大多數為0，但看樣子他仍然是個非常重要的指標。