Homework 2 Report - Income Prediction

學號：b03505052 系級：工海四 姓名：李吉昌

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

	Private	Public
generative	0.70126	0.70773
logistic	0.84829	0.85737

比較經過 standard normalize 的 generative model、logistic regression，實驗結果得 logistic 比較好，其原因為其 class 的機率分布未必是高斯分布，因為由 logistic 經過 gradient descent 的結果會更 fit 實際的狀況。

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

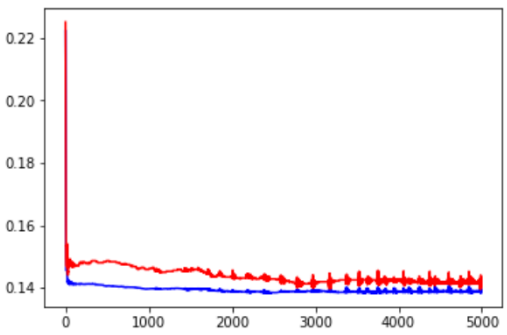
best model 的實作為採用 logistic 加上 sklearn 的 MinMaxScaler 套件，搭配 feature selection，optimizer 採用 Adam 的 full batch gradient，選擇 feature 上，取 gain 平方、gain – loss、gain – loss 三次方以及原本 raw data 的 education num 一次項還有二次項，另外再加入一些條件機率極高或是極低的 label 作為 feature，例如是否三十歲以下，是否具有 gain，gain 在四千以下還是四千到七千或是七千以上，以及是否具有 loss，並加入一個年紀的高斯機率分布。

Train 的時候切成 train 以及 valid，取 valid 的方式是設 numpy 的 random seed 以重現其結果，iterator 次數停在 valid 的最低點，後其 loss 收斂時會有震盪，有時候會比收斂前的最低點還要低，但本人認為其乃誤差並非真正低點，所以取停損在收斂前 valid loss 的最低點，準確率最高達 0.86044。

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。(有關 normalization 請參考：https://goo.gl/XBM3aE)

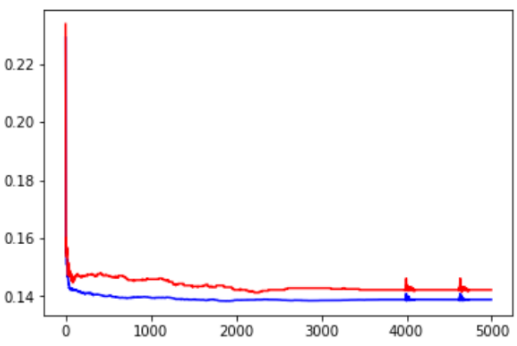
Standard loss 下降曲線:

```
at: 2396 min. train loss: 0.13833816754819997
at: 4988 min. valid loss: 0.13943488943488944

[<matplotlib.lines.Line2D at 0x1c2b2ef5c50>]
```



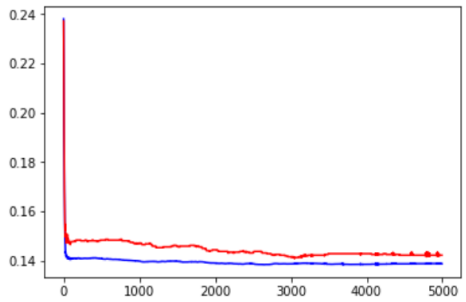
MinMax loss 下降曲線:

```
at: 1887 min. train loss: 0.13833816754819997
at: 2219 min. valid loss: 0.14127764127764128

[<matplotlib.lines.Line2D at 0x1c2b2f41d30>]
```



Mean loss 下降曲線:

```
at: 2645 min. train loss: 0.1383722914178468
at: 3053 min. valid loss: 0.14097051597051596

[<matplotlib.lines.Line2D at 0x1c2b0b5e0f0>]
```



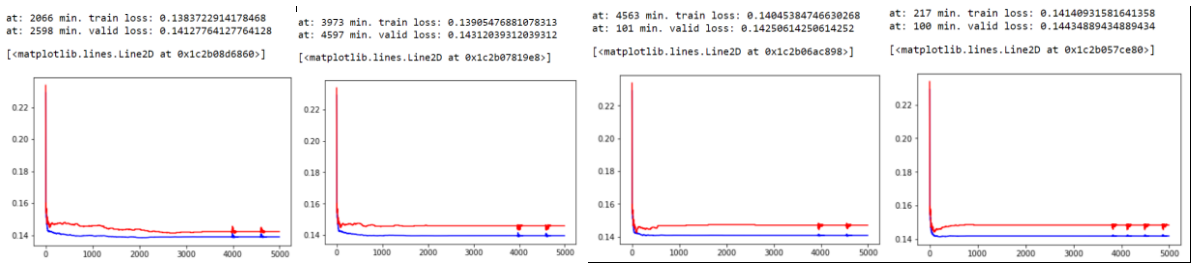
	Private	Public
standard	0.85714	0.85909
MinMax	0.85628	0.86044
Mean	0.85787	0.85995

由曲線圖可見，MinMax 的 scaling 下降至最低點的速度較快且較為平滑，考慮其中一個原因可能是，若 feature 為 encode 的結果(數值只有 0 和 1)，其代表的是數值為 1 時由 weight 決定貢獻多少數值來判斷其種類，但是 0 時不論 weight 大小為何皆無貢獻，在 MinMax 之後這類 feature 不會影響，但若是 standard scaling 則會造成原來 0 的值變成一個有貢獻的數值。

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。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)

	Private	Public
R = 0.01	0.85751	0.85921
R = 0.1	0.85505	0.85921
R = 1	0.85505	0.86031
R = 10	0.85787	0.85995

R = 0.01 R = 0.1 R = 1 R = 10



R 越來越大時，曲線明顯趨近平滑，收斂最低點趨勢也較為明顯。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

個人認為 gain 和 education_num 影響都很大，加了平方項或是三次方 score 都有明顯變好。