Homework 2 Report - Income Prediction

學號：r06922086 系級：資工所 姓名：林凡煒

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

   Generative model Kaggle's private/ public score: 0.84215/ 0.84508
   Logistic regression Kaggle's private/ public score: 0.84645/ 0.85724

   明顯看得出來，不論是 public score 或是 private score 上，都是 logistic regression model 表現得比較好。

2. *請說明你實作的 best model，其訓練方式和準確率為何？*

   利用 keras 實作 deep learning ，structure 採用 fully connected 1 level: 50 units/ level ，加上 10 % dropout ， 500 batch size ，30 epochs。

   DNN model Kaggle's private/ public score: 0.85272/ 0.85405

   整體上與 logistic regression 的結果差不多，其 public score performance 略差。
   但 private score performance 卻比 logistic regression 來得好。

3. *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。*

   DNN model feature scaling(standard): 0.85272/ 0.85405
   DNN model no feature scaling(standard): 0.82262/ 0.82076

   Logistic regression feature scaling(standard): 0.84645/ 0.85724
   Logistic regression no feature scaling(standard): 0.84535/ 0.85245

   Generative model feature scaling(standard): 0.84215/ 0.84508
   Generative model no feature scaling(standard): 0.84215/ 0.84508

   可以由上面結果發現，在 DNN model 與 logistic regression 這兩種方法中， 加上 feature scaling 的確會對 performance 有幫助，推測是在算 gradient 時會產生影響。
   至於 generative method 則是基於統計的方法，過程中也不會去計算 gradient ，所以做不做 feature scaling ，都不會產生影響。

4. *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。*

   Regularization training acc: 0.8524308221491969
   Regularization Kaggle's score: 0.84645/ 0.85724

No regularization training acc: 0.8524922453241608
No regularization Kaggle's score: 0.84657/ 0.85712 (parameter: 10.0)

上述結果觀察，加上了 regularization 後 對於 performance 的影響並不大。由此可知，目前的選用的 model 不夠好， model complexity 低，沒有 overfitting 的產生，所以就算加上了 regularization ， testing result 並沒有變好。

5. *請討論你認為哪個 attribute 對結果影響最大？*

個人認為 marital status 對於結果的影響是最大。我觀察資料得來的結果：經過 training 後， model 在做 prediction 時， 只有屬於 已婚 的狀態，才比較有機會被判別為 > 50k 。