Homework 2 Report - Income Prediction

學號：r06922143 系級：資工所碩一 姓名：台大盧俊澎

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

答：從 Kaggle 的 private score 和 public score 來看，generative model 的得分都只能達到 77 左右，但是 logistic regression 的得分均可達到 84 左右，因此我實作的 logistic regression model 更加準確。

2. (1%) 請說明□實作的 best model，其訓練方式和準確率□何？

答：我實作的 best model 是 logistic regression，和任務 1 一樣。Private score 為 84；Public score 為 84

3. (1%) 請實作輸入特徵標準化(feature normalization)，並討論其對於□的模型準確率的影響。(有關 normalization 請參考：https://goo.gl/XBM3aE)

答：我使用的輸入數據是使用 one-hot-spot 處理後的數據，特徵標準化會對 w 有巨大影響。根本原因是各個特徵的範圍差別很大，一方面是 one-hot-spot 處理後的特徵會有很多列是只有{0,1}；另一方面連續型的特徵也有不同的數量級，如年齡是兩位數而年薪是五位數。因此若不是用特徵標準化，會導致範圍大的特徵的係數太大，影響了模型準確性。在保證其他參數不變的情況下（如 gradient dissent 的步長和步數），使用特徵標準化均會使 Kaggle 的 private score 提高 6% 左右。

4. (1%) 請實作 logistic regression 的正規化(regularization)，並討論其對於□的模型準確率的影響。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)

答：使用正規化可以使 logistic regression 的參數更加平滑，雖然 in sample error 會略微升高，但是 private score 卻有 2%的提高。實際上是限制了 logistic regression 去學習數據裡的噪聲，從而提高實際的準確程度。

5. (1%) 請討論□認□□個 attribute 對結果影響最大？

答：從 logistic regression 訓練後的特徵的係數來看，hours_per_week 的參數最大。猜測原因大概是多勞多得。