Homework 2 Report - Income Prediction

學號：r06522828 系級：機械碩一　姓名：王榆昇

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

	Private	Public
Logistic model	0.84817	0.85712
Generative model	0.84178	0.84582

由上表的分數可看出，在同樣的條件下(對 age、fnlwgt、capital gain、capital loss、hr_per_week 做標準化)，我認為在這個問題上 logistic model 會比 generative model 還要好，因為:

(1) 有足夠的 data 讓 logistic model 尋找最好的 w、b

(2) generative 大多數的資料型態轉換成 one-hot encoding 後只有 0、1，會較符合 Bernoulli distribution，而不是我們預設的 gaussian distribution

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

	Private	Public
Best model	0.85763	0.86007

我以 logistic 為基底，對 5 個 continuous 的資料點(age、fnlwgt、capital gain、capital loss、hr_per_week)做 gaussain normalization；接著我將 training data 的 70%作為 training set，剩下 30%做為 validation set，測試後發現 age、fnlwgt、capital gain、hr_per_week 是重要的 features，將 logistic model 再加上這四項的二次、三次和絕對值取 log 項。再者，education 和 education_num 其實是兩組相同的資料，還有 work_?、occupation_?、naïve_country_?三個特徵我認為是壞的 features，將 education 和三個問號的資料從模型中刪除。

同樣以上述 validation set 繼續做測試，最後採用 learning rate = 0.1、batch size = 300、epoch = 200，不加入 regularization term 做訓練。

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)

以 logistic 模型為例:

Feature normalization	Private	Public
Yes	0.84817	0.85712
No	0.79769	0.80909

在實作 feature normalization 後分數大幅的提高了，從資料點的數值分布來看，continuous 的 5 項 features 平均值大約分別是 38.58、189778.36、1077.65、

87.30、40.44，量值與其他 118 種 0、1 的 features 差異非常大，算 gradient 時這 5 項的 gradient 量值也會與其他 118 個比例非常大，導致整個模型很難收斂至 minima，而在做完 feature normalization 後就能解決這個問題，使模型更能順利的收斂、提高準確率。

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)

lambda	Private	Public
0	0.85763	0.86007
0.000001	0.85714	0.86093
0.00001	0.85726	0.86093
0.0001	0.85763	0.86117
0.001	0.85345	0.85749

由上表可見，加入適當的 regularization term 能增加模型的準確率，是因為 regularization term 能把 weight 的變化也考慮進模型的 gradient，使整個模型不要太 fit training data，而能在 test data 上有更好的表現；但若權重太重(ex: lambda = 0.001)也會導致考慮太多 weight 而失去本身資料 gradient 的真實性，導致模型準確率下降。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

從 weight 來觀察，weight 的絕對值平均大約是 0.323，但 education_num 的絕對值平均卻高達 0.763，甚至有 5 項 weight 的參數高達 1 以上，因此我認為 education_num 是很重要的 attribute。

以第 2 題的 best model 為基底改變 attribute，與同樣 weight 參數值頗大、直覺影響嚴重的 age 做 kaggle 上的分數比較:

	Private	Public
Best model	0.85763	0.86007
Without education_num	0.84768	0.85557
Without age	0.85087	0.85712

由上表可見，當拿掉 education_num 時，acuuracy 直接少掉 0.1~0.2，當拿掉直覺影響嚴重的 age 時，acuuracy 改變值只在 0.1 內，因此我更能認為 education_num 是影響模型的最重要因子。