Machine Learning Homework 2 – Income Prediction

學號：B04901138 系級： 電機三 姓名：張景程

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

答：

	Private score	Public score	Average
Generative model	0.84215	0.84582	0.843985
Logistic regression	0.85124	0.85393	0.852585

由 kaggle 結果可得知無論是在 public set 上還是 private set 上，logistic regression 得到的結果準確率都比較好一點。

2.請說明你實作的 best model，其訓練方式和準確率為何？

答：

我並沒有額外使用 deep learning 的套件，所以 best model 和 logistic regression 的 model 大同小異，主要的參數設定為：

1. 取全部 106 項 feature 的一次方 + age, fnlwgt, sex, capital_gain, hours_per_week 的平方項 + age, hours_per_week 的三次方項，共 113 項
2. 做 feature normalization
3. Regularization (λ=0.1)
4. Iteraion 20000 次、learning rate = 0.2

⟹ Accuracy : private score : 0.85530 public score : 0.85810

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。

答：

Training 條件：

1. 取全部 106 項 feature + age & capital_gain & hours_per_week 的平方共 109 項
2. Iteration：20000 次

	Private score	Public score	Average
without normalization	0.80641	0.80749	0.80695
with normalization	0.85419	0.85823	0.85621

很明顯地，和沒做 normalization 相比，有做 feature normalization 得到的結果準確率提高不少，約為 5%左右，我推測可能是因為在 X_train 的 106 項 feature 中，大部分 feature 的值都為 0 和 1，只有少部分項 feature (ex: age, capital_gain …)有很大的數值，所以若沒做 normalization，在 training 的過程中，這幾項 feature 的 weight 會有很大的 gradient，影響整個 training 的穩定性。

若有輸入 feature normalization，除了可以加速 training 以外，也比較不容易卡在 saddle point 或 local minimum。

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。

答：

Training 條件：

1. 取全部 106 項 feature 的一次方 + age, fnlwgt, sex, capital_gain, hours_per_week 的平方項 + age, hours_per_week 的三次方項，共 113 項
2. Iteration：20000 次

λ	50	10	1	0	0.1	0.01	0.001
Private	0.85333	0.85505	0.85505	0.85517	0.85530	0.85517	0.85517
Public	0.85761	0.85749	0.85773	0.85823	0.85810	0.85823	0.85823
Average	0.85547	0.85627	0.85639	0.85670	0.85670	0.85670	0.85670

由上表可以觀察出，$\lambda \leq 1$時對準確度幾乎沒什麼差別，但當λ很大時，準確度就開始下降了。

5.請討論你認為哪個 attribute 對結果影響最大？

一開始先使用全部 106 項 feature 一次項做 training，透過 weight 的大小比較發現 age、capital_gain 和 hours_per_week 這三項數據影響準確率較大，以直覺和常理來看這也蠻合理的，通常年紀越大、工時越長的人收入會越高，年收較容易超過 50k。此外，收入較高的人通常也比較有多餘的錢可以來進行投資，所以 capital_gain 越高的人，較有可能年收超過 50k。