Homework 2 Report - Income Prediction

學號：b04901020 系級：電機三 姓名：解正平

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

Logistic regression 的準確率較佳兩者都只經過 normalized 的 data feature 來比較，因為本次作業 data 很多有三萬多筆資料，較適合使用此方法，前者是使用在 data 較少來猜測 gaussian 模型。另外 gernerative model 使用不同函式分別為 linalg.inv 及 linalg.pinv 計算 invert covariance，結果卻大不相同。

	Public score	Private score
Generative model (pinv)	0.84545	0.84191
Generative model (inv)	0.66535	0.66306
Logistic regression model	0.85712	0.84805

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

我增加了幾項 feature 因為這些特徵較能區別出 result， 分別是平方項及三次方項的 age、fnlwgt、capital_gain、capital_loss and hours per week 還有四次方及五次方項的 age 及 hours per week，另外我 learning rate 設為 0.1，iteration 設為 30000。另外為了怕 overfitting，有做 regularization ，將 lamda 設為 0.1。

Public score	Private score
0.86093	0.85665

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)

Normalization 準確率較精準，error 較小。原因我覺得是整個圖形較接近圓形，不需要過多的 iteration 即可做完 gradient descent，較迅速也較正確；相較於未 normalize 的 model，因為我們有 123 維度的 vector，這樣想逼近最低點非常困難，因為要 feat 的腳步很難控制，容易受到數值較高的 feature 影響造成整個 training model 像是只用到少數 feature 而無法完整每個 feature 都平均呈現。

	Public score	Private score
Unnormal lr = 30000	0.82948	0.81930
Normal lr=30000	0.85712	0.84805

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)

	Public score	Private score
Lamda = 0	0.85884	0.85566
Lamda = 0.1	0.85872	0.85579
Lamda = 1	0.85761	0.85665
Lamda =100	0.85491	0.85382

在 train model 的時候會發現當具有 lamda 且數值越高，會使得 training error 較快收斂不在變化，也就是會相對 error 較高。另外在看 public score 的變化也差不多，具有 Lamda 的 model 使得 public score 有些微降低，這是因為我們使 model 並沒有那麼受 training data 影響，故 error 較高。但是觀察 private score 會發現，有 lamda 可以有效的增加 private score。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

我認為 hours_per_week 對 income 影響最大，因為只要是 income 大於 50K 的數據，幾乎 hours_per_week 都有大於 40hr，而且從一般常理判斷，年收入多通常那個人工作時數也較長，可以獲得較多薪水，因此我在處理特徵特別加入此項的平方項以及三次方項，很明顯除了使得 accuracy 有變高的趨勢，error 也有些微下降，較原本的 model 更接地 gradient descent 的最低點。