Homework 2 Report - Income Prediction

學號：b04901074 系級：電機三 姓名：吳倉永

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

Feature 不做任何變化：

Method	Kaggle
Logistic	79.67%
Generative	84.36%

Feature 做最佳變化：

Method	Kaggle
Logistic	84.84%
Generative	76.38%

可以發現，Logistic 需要多一點 feature 去進行 training，因為他的優勢就是在 function set 很大，給他多一點 feature 他也能進行判斷，而 Generative 受現在機率分布已經給定，因此準確率還因 feature 增加下降。

除此又對 training data 做了些實驗，發現 Generative 在 data 量少時，準確率是有機會可以衝起來的（在 feature 處理過的 data，我取 20% 資料準確率有從 78%->82%），稍微感受到 Generative 腦補的特徵

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

Ans: 我利用 logistic regression 得並且 learning rate 使用最基本的隨次數遞減的方法，以及使用 regularization 得到最佳的 model，其中準確率在 Kaggle 達 public:85.83%, private:85.55%

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)

Ans: 我本來有對 age, fnlwgt, capital gain/loss, working hours 這些連續性的 features 做標準化，當我拿掉對他們的標準化時，得到的準確率掉了近 10%(85.5 -> 76.6)。我去分析 loss 對 training 次數的圖，發現 loss 的擺動極劇烈，如果沒做 normalization，各個 feature 間的變化幅度不一樣，因此 learning rate 應該要有相同的 scaling 才能符合不同 feature 的變化。我便使用 Adagrad，來改善這個情況，變在準確率上有了 1% 的進步。

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。

Ans: 利用了 Regularization 後準確率從 85% -> 85.8 讓我過 Strong Baseline! 實際上，我去觀測沒做 regularization 跟做了的差別，發現做了的參數會收斂到一個較小的值，沒做的則是會往上收斂到一個較大的值可見 regularization，而這是在我 feature 數較多時有顯著的成長，我做了實驗，使用最一般的未處理過的 feature，得到了有做跟沒做完全一樣的結果。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

刪除的 Feature	刪除前準確率	刪除後準確率	Difference
Age	85.67%	85.22%	-0.45%
Fnlwgt	85.67%	85.68%	0.01%
Capital Gain	85.67%	83.81%	-1.82%
Capital Loss	85.67%	85.41%	-0.26%
Working Hours	85.67%	85.50%	-0.17%

觀看表殼可以發現，Fnlwgt 對 model 的貢獻是最少的，而 Capital Gain 則最大，差到了 1.82%，其實利用 Domain Knowledge 也可以發現，一個人的資本進額對一個人的收入有蠻顯著的影響。