Homework 2 Report - Income Prediction

學號：b03901109 系級：電機四 姓名：陳緯哲

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

	Training accuracy	Test accuracy
Generative model	0.58576	0. 80970
Logistic regression	0.8570	0.85749

Logistic regression 不管在 Training 或是 Testing 上的表現都比 Generative model 好上許多，而 Generative model 在 Test 的表現較 Training 好很多，推測是因為 Test Data 的 bias 小且 variance 也較小，而 Training data 的數量太多，variance 也大，才讓 training 的準確度降低許多。

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

本次作業的 best model 主要是使用 logistic model 來進行實作，並利用 feature scaling 將 feature 進行 normalization，用以訓練的 feature 如下：

一次式：所有 feature

二次、三次、四次式、對數：age、fnlwgt、capital gain、capital loss

在訓練過程中使用 adagrad 與 regularization(lambda = 1)，最後 training accuracy 為 0.8570，test accuracy 為 0.85749，此值較 kaggle 上的結果小，主要是因為 train 的過程中含有隨機的變數，當起始點不同時便有可能到達不同的 Local minimum。

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。

使用 Logistic regression 作為訓練方式，有無使用標準化的結果分別如下：

	Training accuracy	Test accuracy
有標準化	0.855276	0.85700
沒有標準化	0.560913	0.52788

沒有進行標準化的訓練結果並不理想，主要是因為本次作業的 data 大部分為 one hot encoder，非 0 即 1，而少部分的 data 為連續值，最大可到十萬以上，訓練過程中模型無法公平評估這些 data，而會偏重數值較大的 data，如 fnlwgt、capital_gain、capital_loss、hours_per_week，必須要將這些 data 都進行標準化才能讓 model 公平衡量不同 data 的重要性。

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。

Lambda 大小	Training accuracy	Test accuracy
0	0.859951	0.85982
1	0.855412	0.85958
10	0.849918	0.85503
100	0.844219	0.85368
1000	0.823335	0.83083

在沒有使用 regularization(lambda = 0)的情況下，Accuracy 是最好的；隨著 Lambda 增加，Accuracy 會逐漸下降，這與一般認知 regularization 可以改善 test 的結果不同，我認為是主要是因為此 model 的 overfitting 問題並不嚴重，才使得 regularization 看不出相對的成效。

5. (1%)請討論你認為哪個 attribute 對結果影響最大？
為了知道哪個 attribute 的影響最大，分別單獨使用各種 attribute 來做 generative model(註)，以下的表格為各參數進行 generative model training 後的 valid data accuracy。
從表格中可以看出 workclass 、education、education_num、capital_gain、capital_loss 、hours_per_week 這 6 個 attribute 的 accuracy 最高，意味著根據這些參數的分布，generative model 可以較準確的找到每個人對應的分類。

註：之所以使用 generative model 而不是 logistic regression 主要是因為若分別取用各個參數，對 logistic regression 來說其 feature 的數量太少，完全無法分辨何者較為重要，因此採用 generative model 來觀察 feature 分布與結果的關係。

age	workclass	fnlwgt	education	education_num
0.540650	0.632830	0.466190	0.628322	0.628322

marital_status	occupation	relationship	race	sex
0.514845	0.594901	0.519509	0.304058	0.411907

capital_gain	capital_loss	hours_per_week	native_country	All feature
0.713819	0.727654	0.601585	0.281051	0.575470