Homework 2 Report - Income Prediction

學號：b04502139 系級：電機三 姓名：戴瑋辰

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

   Logistic regression 表現較好，我在使用 generative model 時除非出現太明顯的不同（年紀很小），不然基本上分不出來。

   在 kaggle 上 generative model 正確率為 0.79975；而 Logistic regression 可以做到 0.85552

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

   用 Keras 疊 neural network 來 train。

   經過多次測試之後我發現 network 越深，效果並不會變好，反而非常容易出現 overfitting 的情況。最後，我只用了兩層 layer，activation function 都使用 sigmoid，改變 neuron 的數量 train 了五個 model

Number of neurons	100	256	512	1024	2048
Accuracy (training set)	0.8803	0.8814	0.8843	0.8937	0.9013
Accuracy (validation set)	0.8566	0.8563	0.8578	0.8603	0.8575
Score on Kaggle	0.85442	0.85466	0.84914	0.85565	0.85651

   從表中可以看到單獨一個 model，2048 個 neuron 表現最好，但是實際上分數卻沒有過 strong baseline。最後，我想了一個辦法把五個 model 合併，其作法有點像是在投票，分別紀錄每個 model 在相同的 input 時對應到的 output，若這個 output 為 class0 得到 3 票以上，則將其歸類為 class0；反之，則歸類為 class1。在這樣統計完之後再上傳結果，準確率提高到了 0.85786，剛好超過 strong baseline！

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)

我對四種連續的 feature (age, fnlwgt, capital_gain, capital_loss)做 Feature noralization. 試了很多次之後我發現對於 capital_gain 和 capital_loss 做 standardization、對 age 和 fnlwgt 做 mean normaliztion 後下去 train 所得到的效果會最好。

實際上在我實驗的過程中，不同種類的 normalization 其實對於結果的影響並不大，在 kaggle 上只有差 0.0XX，但是如果完全不做任何處理，效果會很差正確率大概只有 0.7XX，比起我 train 出來最好的結果少了大約 0.1 左右。

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。

在我的 model 加入正規化之後（我將係數設為 0.01），在 training data 上的準確率下降了，但是在 validation set 上的準確率稍微上升了一些，但是效果並不顯著。推測是因為 model 不夠複雜，因此做正規化沒有太大的效果。

在用ＮＮ訓練時我就有加入正規化，因為 model 夠複雜所以得到的結果就還不錯。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

我使用 logistic regression 來判斷選用的 features，並切出 1 成的 training data 作為 validation set。在固定其他參數後我發現年齡的影響很大，其次是 work_class、education、native country。我做了下表的整理：

	All of the features	Without “age”	All of the features	Without “work_class”
Accuracy on validation set	0.83169	0.77639	0.83169	0.81879
	All of the features	Without “education”	All of the features	Without “native_con”
Accuracy on validation set	0.83169	0.80934	0.83169	0.79914

從表中可以看出，若少考慮了年齡，其準確率下降了 0.06 左右，其餘的 3 個因素則是讓準確率下降差不多 0.03。至於剩下的那些 features 選不選用其實對結果影響並不大。