ML2017FALL-HW2-Report

學號：B03801039　系級： 電機四　姓名：楊福恩

1.請比較你實作的 generative model、logistic regression 的準確率，何者較佳？
答：此題實作的兩個 model 皆有做 feature normalization
(1) generative model accuracy:
　　public:0.84508 ;private:0.84129 ;average:0.84319
(2) logistic regression:
　　public:0.84926 ;private:0.83847 ;average:0.84387
由以上數據的 average accuracy 來看，logistic regression 的 model 較佳，但兩者相差並不大。

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：利用 keras 實做有一層隱藏層的神經網路，其中隱藏層含有 54 個 units，activation function 為 sigmoid function;輸出層有 2 個 units，activation function 為 sigmoid function，整個 model 以 batch size=32，epoch 數目為 30 來訓練
準確率:
Public:0.85786 ;private:0.84817 ;average:0.85302

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：
(1) generative model:
　(a)沒有 feature normalization:
　　public:0.76523 ;private:0.76231 ;average:0.76377
　(b)有 feature normalization:
　　public:0.84508 ;private:0.84129 ;average:0.84319

(2) logistic regression:
　(a)沒有 feature normalization:
　　public:0.79213 ;private:0.78847 ;average:0.79030
　(b)有 feature normalization:
　　public:0.84926 ;private:0.83847 ;average:0.84387
由以上數據可知，是否做特徵標準化對結果影響甚大，原因大致有以下兩點:
(a)如同課程所述，有做特徵標準化後的 loss 對 feature 的圖較接近正圓形，沒有做特徵標準化的圖則可能接近扁長的橢圓形，因此在 gradient descent 的過程中，有做特徵標準化較容易收斂而有較佳的結果。

(b)本次作業的 feature 若沒有做特徵標準化，在通過 sigmoid function 時會 overflow，故若無在 sigmoid function 的程式碼中加上輸出的限制或是做特徵標準化，就會對準確度造成很大的影響。

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：

	λ=10	λ=1	λ=0.1	λ=0.01	λ=0
public accuracy	0.81400	0.84850	0.85012	0.84324	0.84926
private accuracy	0.81451	0.84277	0.84277	0.83761	0.83847
average accuracy	0.81426	0.84564	0.84645	0.84043	0.84387

由上表可知，當 λ=10 時，準確率反而較不加 regularization 時下降，原因應為原本沒加 regularization 的模型就沒有 overfitting 的問題，加上較大的 λ 後讓模型複雜性降低太多，反而使得準確率下降。當 λ=1 和 0.1 時，準確率較原先的 model 略好，但效果並非非常顯著。而當 λ=0.01 時，準確率又些微下降，但和原先 model 相差不多。故可看出這個模型加入 regularization 並無發揮使準確率明顯上升的作用。

5.請討論你認為哪個 attribute 對結果影響最大？

我認為 capital-gain 的影響最大，原因為當我們用全部 feature 且有 normalization 下做 logistic regression ，所得到的 weight 中 capital-gain 所對應的 weight 是最大的，代表 capital-gain 這個 attribute 對結果影響最大。