Machine Learning HW2 Report

學號：b06901007 系級：電機二 姓名：戴子宜

2019.3.22

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

   在資料的前處理方式相同的情況下（都是用已經分好的 X_train Y_train），Logistic regression 的在 validation set 準確率較高。下圖的 generative model 是利用 Gaussion distribution 的假設，而 Logistic regression 的 model 則是利用 adagrad，learning rate 初始值為 0.001 訓練 50000 個 epoch 的結果。

	Train Accuracy	Validation Accuracy
Generative Model	0.843372507	0.839189189
Logistic Regression	0.848613898	0.845577396

2. 請說明你實作的 best model，其訓練方式和準確率為何？

   Best model 是用 fully connected feed forward network，input 的 data 是把 'train.csv'的內容先處理過，每個 x 是一個 80 維度的 vector。Network 有三層 hidden layer，第一層有 1000 個 neurons，第二層有 500 個 neurons，第三層有 200 個 hidden layer。訓練方式是 learning rate 初始值為 0.001，batch size 32，且利用 Adam 的方法更新參數。最好的 validation accuracy 為 0.8563，是第 22 個 epoch。

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。

		Training Accuracy	Validation Accuracy
Logistic Regression	Original data	0.2413448952879581	0.2397058823529411
	Normalization	0.8528550392670157	0.8501225490196078
Generative Model	Original data	0.2413906064452725	0.2390663390663390
	Normalization	0.8433725072683347	0.8391891891891892

   無論是哪一種 model，利用最原始的資料訓練出來的準確率大約都在 0.24 左右，但只要加上 feature normalization 後都會得到顯著的提升。

   Generative model 是利用 Gaussion distribution 的假設，而 Logistic regression 的 model 則是利用 Adam，learning rate 初始值為 0.001，batch size 為 32，訓練 100 個 epoch 的最佳結果。

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。

λ	Training Accuracy	Validation Accuracy
0	0.8432906105401089	0.8406633906633907
0.1	0.8432906105401089	0.8407862407862408
1	0.8432496621759961	0.8406633906633907
100	0.8385406003030179	0.8162162162162162

在原本的 Loss function 加上 $\lambda \sum w_i^2$，對於準確率的影響看起來不明顯，但可以看到當λ越大，training accuracy 和 validation accuracy 都會下降。

5. 請討論你認為哪個 attribute 對結果影響最大？

我認為最重要的 attribute 是 hours per week 或是 capital gain。如果只利用 hours per week 作為訓練資料的話，利用 logistic regression 可以得到 81.56%的準確率，如果只利用 capital gain 作為訓練資料的話，可以得到 80.32%的準確率。

如果只利用其他的 attribute 作為訓練資料的話，準確率大約都在 75-77%之間，而所有的資料量中 ≤50k 的比例大約就是 75%，因此如果只看其他的 attribute（一次只有一種 attribute）其實對於訓練的幫助很小。