Homework 2 Report - Income Prediction

學號：b04902063　系級：資工三　姓名：陳昱儒

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

	Train data set	Test data set
Generative	0.84232	0.84545
Logistic	0.85359	0.85737

由上表可知，Logistic 的 model 不管在 train data set 上還是 test data set 上的準確率均比 Generative model 還要高，推斷可能是這次的薪資資料分佈和 generative model 原先假設的 Gaussian distribution 有落差，所以訓練出來的 model 準確率較沒有預先假設分佈的 logistic 還要來得低。（以上兩者的訓練資料均一致）

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

	Train data set	Test data set
準確率	0.85734	0.85835

我是將助教提供的 feature 中，在我認爲字詞中有 work 比較有用的 feature 取出，除了原本的 feature 再加上其的一次方、二次方、二點五次方以及三次方，並做 normalization，之後每次隨機選取七成的 training data 進行訓練，找出 performance 最好的 model 儲存。

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)

	Train data set	Test data set
無 normalization	0.79613	0.63452
有 normalization	0.85703	0.85712

由表中可以觀察到，有做 normalization 的 model 的準確率不管在 train data set 還是 test data set 上均較沒有做 normalization 的高，並且在沒有做 normalization 的 model 訓練過程中，發現其準確率呈現非常不穩定的跳動，甚至出現從 70% 降到 20%再跳回 70%的狀況。（以上兩者的訓練資料，訓練次數，初始值均相等）

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。(有關 regularization 請參考：https://goo.gl/SSWGhf P.35)

	Train data set	Test data set
λ = 0	0.85531	0.85675
λ = 0.1	0.68700	0.67960
λ = 0.01	0.84226	0.84348
λ = 0.001	0.85024	0.84975
λ = 0.0001	0.85571	0.85810

由上表可見，當 λ 值越大時，其所訓練出的 model 準確率越低，但當 λ 值低到一個程度時，其訓練出的 model 甚至能比 λ=0 時好一點。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

	Train data set	Test data set
[0] age	0.851289	0.85503
[1-9] workclass	0.851812	0.85540
[10] fnlwgt	0.852702	0.85761
[11-26] education	0.853562	0.85737
[27-42] education_num	0.853562	0.85737
[43-49] marital_status	0.853839	0.85724
[50-64] occupation	0.848034	0.84987
[65-70] relationship	0.852886	0.85614
[71-75] race	0.853316	0.85712
[76-77] sex	0.852948	0.85724
[78] capital_gain	0.838329	0.83832

[79] capital_loss	0.850982	0.85429
[80] hours_per_week	0.851136	0.85503
[81-122]native_country	0.852149	0.85687
None	0.853593	0.85737

由上表可知，Capital_gain 此項 feature 對於整體準確率影響最大，將其抽出後準確率下降了近 1.5%，另外也可注意到，occupation 對於整體準確率的影響也很大，下降了近 0.6%。（以上資料除了變動訓練的 feature，其 learning rate 以及初始值，訓練方式均相同。）