學號：B06705001 系級：資管二 姓名：楊力行

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

   generative 的為 0.84080、0.84656

   而 logistic regression 的為 0.85628、0.85429

   logistic regression 的較佳

2. 請說明你實作的 best model，其訓練方式和準確率為何？

對幾項數據是連續性的特徵，進行平方，當作新的特徵加到 DATA 之中，

接著標準化，再先從所有資料中挑選幾個當其不為 0 時 class 是 1 還是 0 分布的最為極端的特徵先用 logistic regression train 出個 W1 再拿所有的特徵去用 logistic regression train 個 W2，再 test 時先用 W1 去對 TEST 作判斷，當她的 p>0.65 時直接判斷為 class1，不是的話再用 W2 去判斷，準確率為 0.85628、0.85429。

但是，後來我再測試，發現純粹用 W2 判斷分數會較高，可知這方法有些多餘。

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

generative 無標準化 0.84080、0.84656

generative 有標準化 0.84080、0.84656

標準化對 generative 沒有任何影響 因為生成函數所考慮的為每點的出現機率，標準化後其機率不變，結果也不變。

logistic 無標準化 0.79449、0.79164

logistic 有標準化 0.85628、0.85429

在這之中，標準化對 logistic 的答案影響頗大，因為某些特徵的值域會遠大於其他的，在這種時候，若他的係數變化一點，對預測的 Y 的變化就會很大，收斂的步驟會頗為偏斜，整體的收斂要花很久，因此要標準化，讓他的等高線變的圓潤點加速收斂的速度。

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。

有正規化 0.85628、0.85429

沒有 0.85628、0.85405

正規化對程式的結果影響不大，就只差了 1 筆資料的判斷錯誤，這是因為我的模型並不算複雜，也並沒有 overfit 訓練資料，因此正規化的作用並不大，甚至可能降低正確率。

5. 請討論你認為哪個 attribute 對結果影響最大？

我認為 education 對結果的影響最大，隨著學習年分和學歷的增加，在訓練資料中，該學歷的所有人之中 薪水大於 50K 的比例是逐步上升的從 1-4 的 0.03 到博士的 0.74，並且在分別利用各特徵進行 gaussion 時，education 所的到的正確率為 0.78135、0.77874，是最高的，從這可以得出，education 對結果的影響最大，也很符合大眾的想法。