Homework 2 Report - Income Prediction

學號：b05902127 系級：資工二 姓名：劉俊緯

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

Generative model 的 acc 較為差，我認為 Logistic regression 會比較適合這次作業的原因是因為，logistiv 本身會有懲罰極端特例的點使得回歸線不會往極端特例篇(因為所有都會被壓成[0,1])，而這次作業：薪水分類，根據我國悲慘的現狀，明顯得知貧富差距會非常大。意思也就是極端的例子一定是特別多，因此 logistic regression 較能處理這類分類。以下為準確率:

generative model: pri-scores : 0.84203 pub-scores: 0.84557
logistic regression: pri-scores 0.85603 pub-scores: 0.85712

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

根據"機器學習基石"的課程，我們知道就經驗法則來說，(training data 數目)N ~>= 10 * model VC dimension 的情況就已經比較不會 overfitting 了。再加上我們知道一個預測 task 是不可能用線性組合就能夠做的好的。換句話說：我們只要强置讓它轉為非線性並且增加維度，準確率就有機會變好。(事實也是如此。)至於要挑何種 feature? 我們知道 age,capital_gain,capital_loss , hours_per_week , fnlwgt 只有這五個 feature 不是 one-hot，因此其他 feature 無法轉換，因為只能分成 0 和非 0，而這個依舊可以和 bias 線性組合出來，只是徒增維度。因此把上述五種 feature 從 x, x^2 , x^3 … 加到 x^100 次方就可增加維度，在用一般的 logsitic regression，就能衝到前面去。

pri-scores:0.86893 pub-scores:0.87199

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。

在這之前，我們先想想為何要 normalization，理論上來說，不管測資分佈怎麼樣，normalization 也只是平移與 scale 某個數字。對 regression 系列(wx+b)來說，scale 可以合併在 w，平移可以合併在 b，看起來 normalization 並沒有什麼意義。我認為 normalization 的意義在於一部分在於 IV(initial vector)，大部分我們做 regression 時 IV 都是全 0 或者 unifrom/normal，但是不管怎麼樣，都是介於[0,1]之間。因此如果 avg(feature)距離 0/1 很遠，不做 normalization 收斂的速度會比較慢。此外，因為 w 會多乘東西，所以 regularization 也會大受影響。

No-normalization：pri-scores: 0.78000 pub-scores: 0.78710
normalization： pri-scores: 0.83601 pub-scores: 0.83685

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。

regularization 的意義在於强置讓曲線平滑，這意味著用"能力"換取"不要 overfitting"。但是就第二點我所提出的理論，在測資數量很多且 model 的 VC dimension 的情況下，並不容易 overfitting。所以有沒有做 regularization 對我所實行的 model 或者對原本的測資做都沒有很顯著的效果，反而會更差。因為它就幾乎沒有 overfitting。

No-regularization : pri-scores: 0.84768 pub-scores: 0.85466

regularization with lamba 0.01: pri-scores: 0.84129 pub-scores: 0.84373

5. (1%) 請討論你認為哪個 attribute 對結果影響最大?

根據 logistic regression 出來的 w 的結果，age,capital_gain,capital_loss , hours_per_week , fnlwgt 這五個非 one-hot 的影響很大，另外 sex 和婚姻狀態的影響也很大。這樣個結果不怎麼稀奇，很符合社會的狀態。例如年齡越高→ 工作可能越久→ 薪資水平高；男性比女性比較容易賺得到錢因為性別刻板...；工時越長越多錢這也是廢話。