學號：R05944043　系級：　網媒碩一　姓名：宋焱檳

1.請說明你實作的 generative model，其訓練方式和準確率為何？

答：

generative model 主要通過先分類 yhead 是 0 和 1 的情況，分別去計算 sigema 和期望 u，算出共用的 sigema 簡化了計算。然後套用下面的公式進行計算 z。之後帶入 sigmoid algorithm。最後對參數進行 update，就得到了最後的結果。然後帶入 Test 進行測試。準確率在 public score 得分 0.8411。

$$\sum share = \frac{cnt1}{cnt}\sum 1 + \frac{cnt0}{cnt}\sum 0$$

$$\Sigma_1 = \Sigma_2 = \Sigma$$

$$z = \underbrace{(\mu^1 - \mu^2)^T\Sigma^{-1}}_{\boldsymbol{w^T}} x \underbrace{- \frac{1}{2}(\mu^1)^T\Sigma^{-1}\mu^1 + \frac{1}{2}(\mu^2)^T\Sigma^{-1}\mu^2 + ln\frac{N_1}{N_2}}_{\mathrm{b}}$$

2.請說明你實作的 discriminative model，其訓練方式和準確率為何？

答：

discriminative model 使用的是 logistics regression，就是把全部 data 讀入 X 後，做 feature normalization 計算 z=wx+b，把 z 通過 sigmoid algorithm 進行處理為 0-1 的 array，之後使用 Adam 演算法設計 m 和 v 對參數進行 update，選取在迭代中 Loss 最低的一組作為 model 存下來。最後在通過 test 把>=0.5 的劃分為 1 第一類，其餘的為 0 第二類。public 上的 正確率在 0.853 左右，使用了各種優化算法，均提升不大。主要公式如下：

***Logistic Regression***

Step 1: $f_{w,b}(x) = \sigma\left(\sum_i w_i x_i + b\right)$

Output: between 0 and 1

Training data: $(x^n, \hat{y}^n)$

Step 2: $\hat{y}^n$: 1 for class 1, 0 for class 2

$$L(f) = \sum_n C(f(x^n), \hat{y}^n)$$

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：discriminative model 上：

| 有無 normalization | Loss | Validation 準確率 | Public score |
|---|---|---|---|
| 無 | 0.321 | 0.849 | 0.849 |
| 有 | 0.320 | 0.850 | 0.850 |

這裡的 feature normalization 是對於 X_train 中前六 column 的 data 做的（因為只有前六 column 超過 1），公式：$\frac{x - x\min}{x\max - x\min}$。feature normalization 會影響一開始 Loss 的位置，無的時候大概從 7 開始收斂，有做的時候從 1 左右，最後得到的結果還是很相似的。在 validation 上的結果也比較相似，因為最終總會收斂到一起，只是 normalization 加速了這個過程。

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：

| 有無 regularization | Loss | Validation 準確率 | Public score |
|---|---|---|---|
| 無 | 0.3147 | 0.852 | 0.851 |
| 有 | 0.3188 | 0.850 | 0.853 |

logistic regression 中使用 regularization 可以減少 model 的 variance 使得 data 比較集中，但是這樣也會導致整體的 loss 變大，因為不加 regularization 的時候 data 可以有一定的浮動空間，有機會達到我們需要的最好的結果。在 validation 上準確率也不是太高，所以在 public 上的得分可能沒有不加的時候好，但是最後的得分一般比不加的好。

5.請討論你認為哪個 attribute 對結果影響最大？

經過多次的測試，保存 model，每次都有做 feature normalization。發現 capital_gain 在最後得到 model 中，有最大的 weight，遠遠超過其他的 weight，所以可以認為它對結果的影響是最大的。

實驗心得：

這次的實驗，需要注意對數據的觀測，對 data 的 column 進行 feature normalization。一般情況下，需要先在自己設定的 validation 上得到不錯的 Loss，選擇好 model，在經過所有 data train，即使在 PublicScore 上可能比分不那麼高，但是綜合 private 後，應該會有很大的提升。畢竟不會因為一半的 Score 影響整體 model 的選擇。這次的 data 在 train 的過程中，發現一層處理的瓶頸。怎麼 train 都在 0.853 附近。說明可能需要對輸入的 data 進行處理，或者開始使用 DNN。由於時間有限，嘗試過的效果都不是很理想於是就沒有加入文檔裏面。