學號：T10902303 系級： 計資一　姓名：梁高浚

1. (1%) 請比較說明 generative model、logistic regression 兩者的異同為何？再分別列出本次使用的資料中五個分得正確/不正確的 sample, 並說明為什麼如此？

Logistic regression 是 discriminative 的方法，而 generative model 是 generative 的方法。在 logistic regression 中, 我們可以用 gradient descent 去迭代，算出 optimal w and b, 使到 loss function 達到 minimum. 而 generative model 的話，就是先算 u^1, u^2, 和 Σ ,然後算出 w 和 b.

兩者算出的 w 和 b 是不一樣的。

五個分的正確的 sample:

1. i:1028, y:0, yhat: 0.00670425
2. i:1029, y:0, yhat: 0.28461435
3. i:1030, y:0, yhat: 0.00968792
4. i:1031, y:0, yhat: 0.02517578
5. i:1032, y:0, yhat: 0.23454912

五個分得不正確的 sample:

1. i:3764, y:0, yhat: 0.7456637
2. i:3762, y:1, yhat: 0.28589513
3. i:3753, y:1, yhat: 0.44111478
4. i:3749, y:1: yhat: 0.15594321
5. i:3743, y:0, yhat: 0.66042516

他們在 logistic regression 下是分的不正確的，原因可能是 logistic regression 這個 model 的 complexity 不足，沒辦法 capture 到 feature 之間的重要性。而且，logistic regression 需要對 data 進行假設，他假設了 feature 之間是 independent 的，但這個在實際情況未必會發生。就比如在這幾個 sample 裡面，有一些 feature 可能中間有關聯，例如是 never-married 和 age。最後 logistic 的準確率只有大概 84%,過不了 strong baseline. 還有就是看了一下其他的 sample, 發現當 y=1 的時候，準確率不足，所以可能是最後算出的 weighting 和 bias 偏向 y=0 的 sample

2. (1%) 請實作兩種 feature scaling 的方法 (feature normalization, feature standardization), 並說明哪種方法適合用在本次作業？

如果使用 normalization, accuracy 是 0.8483979936533934

如果使用 standardization, accuracy 是 0.8288463507011977

normalization 把 value 會 scale 到[0,1]區間裡面，而 standardization 則是把 data rescale 至 mean 0, std = 1 的 normal distribution 裡面。從上述結果可見，本次作業

適合使用 normalization,而 standardization 的 accuracy 比較低，可能是因為 data 本身並不服從 normal distribution, 應該用 normalization 方法比較好。

3. (1%) 請說明你實作的 best model 及其背後「原理」為何？你覺得這次作業的 dataset 比較適合哪個 model？為什麼？

我的 best model 還是使用了 logistic regression，準確率只有 84%，過不了 strong baseline。這個 model 的原理就是使用 sigmoid function 去預測最後 P(Y=0)和 P(Y=1)的概率，並且使用 gradient descent 去更新每次的 w 和 b, 最後使用 w 和 b 給出預測的概率。我最後使用了 grid search 去尋找最佳的 threshold，發現最佳 threshold 在 53%左右，並不是 50%。 不過，這個是在 validation set 的結果，有可能會導致 overfit, 因此在最後提交的 result 並沒有超過 private leaderboard 的 strong baseline。我覺得這次作業的 Dataset 比較適合一些 deep learning model，比如 neural network。因為 feature 有很多，需要一個比較 complex 的 model 去 generalize data 中間的規律，而 logistic regression 相比還是稍微簡單，沒辦法超越 85%以上的準確率。