學號：B03505027　系級：電機三　姓名：劉亦浚

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

答：

抽取 X_train 中的所有資料（每個人的 feature 皆為 106 dim，共有 32561 人），因此為一個（32561, 106）的矩陣，再從 Y_train 中取其相對應 label。然後開始計算 mean1、mean2（計算時將 label 為 1 分為 class1；label 為 0 則分為 class2），以及 sigma1、sigma2，再由兩者的 sigma 推出共用的 sigma，最後由 mean1、mean2、共用的 sigma 求出 w、b，將 x（X_test 取 transpose）帶入，即可得到 z = w • x + b，把 z 帶入 sigmoid 函數，就可算出每筆 testing data 的機率，若機率大於等於 0.5 則視為 class1（輸出 1）；小於 0.5 則為 class2（輸出 0）。其結果準確率為 0.84177。

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x^n \qquad \Sigma = \frac{1}{N}\sum_{n=1}^{N}(x^n - \mu)(x^n - \mu)^T$$

$$\Sigma = \frac{N1}{N1+N2}\Sigma^1 + \frac{N2}{N1+N2}\Sigma^2 \quad (\text{共用的} sigma)$$

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

$$P(C_1|x) = \sigma(w \cdot x + b)$$

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

答：

抽取 X_train 中的所有資料之外，再加上 age、capital_gain、capital_loss、hours_per_week 等平方、三方和四方項，還有 fnlwgt 的平方項與 age 和 hours 相乘，此時所取 feature 的 dim 為 120（有對數據做標準化的處理），訓練方式採取一次丟 32561 筆的 data 量，iteration 為 2000、learning rate 為 0.5、正規化的 λ 為 1，先由取好的 data 與 weight 做 dot 得到 z，再帶入 sigmoid 函數，得到的結果分別與相對應 label 做處理，然後進一步求得每一個 weight 的 gradient，最後利用 adagrad 的方法來更新 weight。使用學習後的 weight 來判斷 X_test，得到的結果若大於等於 0.5 則寫出 1；小於 0.5 則寫出 0。其結果準確率為 0.85786。

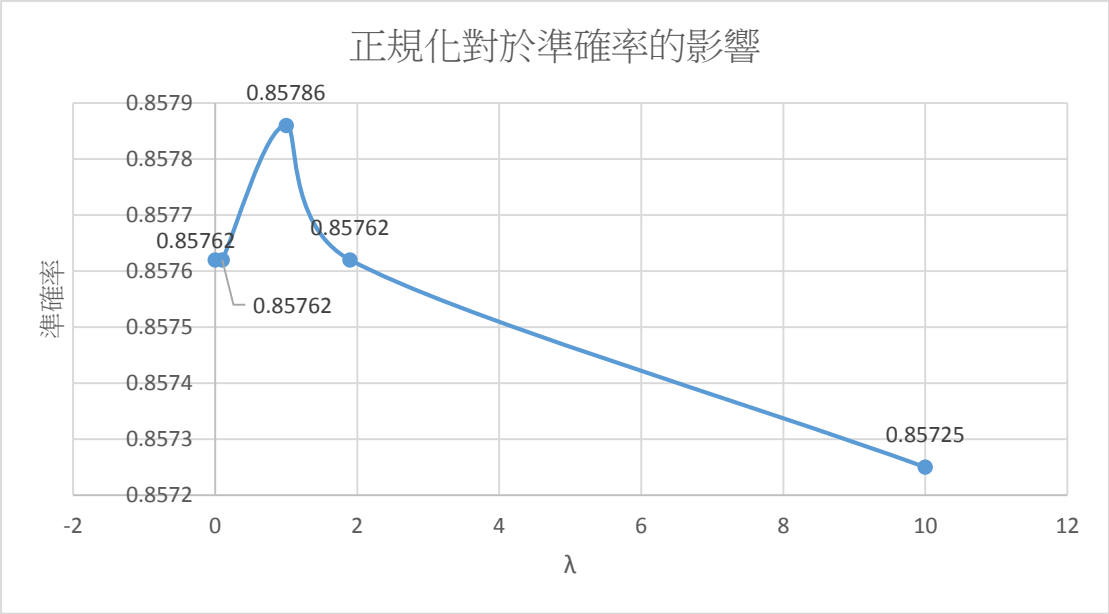
3. 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。

答：

在 logistic regression 中，使用特徵標準化會大大的提升準確率（以下分數建立在只取 X_train 中現有資料，dim 為 106），剛開始沒有標準化時，準確率為 0.80025，使用標準化後，其準確率提高至 0.85307，效果非常好，因為可以避免數據過於分散與出現很大的值（如 fnlwgt 若無標準化帶入 exp 中會 overflow），數據集中好處為減少極端點，model 就不會為了迎合某些點而 overfitting，且標準化也能加速其收斂速度；但在 generative model 中，沒有標準化的準確率為 0.84177，標準化後反而降成 0.82260，會使準確率下降的原因，應該是標準化後 feature 每項數值平均為 0，標準差為 1，使 Gaussian model 過於簡單而導致。

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：



此圖所抽取的 feature 與第二題相同（dim 為 120），λ 分別為 0、0.1、1、1.9、10，其相對準確率皆為 kaggle 上的分數，這次作業使用正規化對我的模型在準確率上好像沒有太大的影響，但 λ = 1 還是比 λ = 0 有較好的結果，採取正規化可使 weight 不會過大且有較 smooth 的 function，所以應該能增加準確率，最後在我的實作中也將 λ 取為 1。

5. 請討論你認為哪個 attribute 對結果影響最大？

在這次的作業中，我認為特徵標準化對結果影響最大，最開始還沒看到報告有討論標準化的影響，就直接將數據丟下去 train，出來的結果連 simple 都沒過，後來看完報告的題目後，馬上將數據標準化後再 train，出來的結果就直接超越 simple 了，其準確率直接提升了 5%，可見標準化的影響甚大且效果十分顯著，我認為原因應為 fnlwgt 這項指標差異過大，造成數據過度分散，在訓練的過程中，model 為了 fitting 某些點進而造成 overfitting 的現象，所以在 testing 的結果就不甚理想，但採取標準化就可以解決這個問題了。