學號：B03901076　系級： 電機三　姓名：林尚謙

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

在 generative model 中使用 Gaussian distribution 的模型來預測，計算出 train data 中分別屬於>50k 和<50k 時，所有 feature 各自的 mean(u)和 covariance matrix(sigma)，之後便可直接由公式求得 w 和 b，之後便可由$y = \sigma(w \cdot x + b)$，直接預測 test data 是>50k(y>0.5)或者<50k(y<0.5)。

準確率部分使用原本的 106 項 feature 訓練結果可以達到 84.4%，再加入五項連續資料(age、fnlwgt、capital_gain、capital_loss 和 hours_per_week)的平方項後可達 85.4%。

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

在 discriminative model 中是使用 logistic regression 來做訓練，先對連續的 feature 作 feature normalization，對每項 feature 設定初始的 w，再透過 gradient decent 的方式來找到使誤差最小的 w 和 b，並使用 adagrad 和 batch 的方式來加速訓練過程。最後由$y = \sigma(w \cdot x + b)$，找出 test data 是>50k(y>0.5)或者<50k(y<0.5)。

準確率比起 generative model 較為準確，使用原本的 106 項 feature 時最佳可達 85.5%，加入五項連續資料(age、fnlwgt、capital_gain、capital_loss 和 hours_per_week)的平方項後可達 85.9%。

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：

這次 feature normalization 主要是針對這次資料中的五項連續特徵(age、fnlwgt、capital_gain、capital_loss 和 hours_per_week)，因為相較於其他不連續資料，這些連續資料的數量級與變動幅度都較大，因此需要做 normalization 來降低變動幅度和誤差，而其他的不連續特徵(0 或 1)則維持原樣。

準確率的影響上在 discriminative model 的部分影響較大，如果沒做很可能因為影響太大根本訓練不出正確的模型，而對於 generative model 的影響較不明顯，大約僅有 0.1%的影響，但若加入了連續資料的平方項，如果沒做 normalization 會因為資料數值過大而有 overflow 的現象。

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：

正規化是在 gradient decent 的每個 iteration 要更新 w 時，加上一個 lambda 乘上 w，可以避免參數的值變得過大而導致 overfitting，在這次 logistic regression 中 lamda 的值大約選擇 0.1~0.01 左右，對於準確率大約可以提升 0.1%。

5. 請討論你認為哪個 attribute 對結果影響最大？

答：

我在觀察 logistic regression 做完 gradient decent 後的 w 參數時，發現大多數的參數大致都介於-1~1 之間，有少數參數會超過這個範圍，因此這些超過範圍的參數對應到的 attribute 應該就是對結果影響較大的，包括了 age、capital_gain、一部分的 education 和 education num、Own-child 和 Wife 等，而其中影響最大的是 age 和 capital_gain 這兩項 attribute。