學號：B03901030 系級： 電機三 姓名：蕭晨豪

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

我的訓練方式模仿老師投影片的方法，假定資料是 Gaussian distribution，並且對 Y=0 和 Y=1 的資料的 covariance matrix 取加權平均，兩種資料均使用加權平均後的 covariance matrix，並使用 Maximum likelihood 的方式決定，在 Kaggle 上的分數為 0.84128(84.128%的準確率)

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

首先我實作線性的 discriminative model，使用 gradient descent 和 Adagrad 方法去訓練，準確率為 0.85307，而最後效果最好的 model，我加入了前六個 attribute 的二次以及三次項當作 feature，用一樣的方法訓練，最好的準確率達到 0.85799

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：

特徵標準化對於正確率的影響非常大，public score 從 0.78194 直接進步到了 0.85307，主要是因為此次資料裡有連續的資料也有離散的資料，其中某些連續的資料範圍很大，直接訓練的話做出來的 weight 可能會有較大的誤差，因此作 normalization 之後會大幅提升正確率。

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：

我嘗試了四次 regularization(lambda 分別為 0.1,1,5,50)，不過對模型準確率都沒有太大的影響，且全部的準確率都比沒有作 regularization 還差，猜測當進行 feature normalization 之後，w 會被縮的很小以至於加 regularization 沒有用。為了證明這個猜想，我對沒有 feature normalization 的資料作 regularization，當 lambda = 50 左右時，準確率從 0.78194 進步到了 0.78354，因此我認為我的猜想有一定的可能是正確的。

5.請討論你認為哪個 attribute 對結果影響最大？

我另外計算了 X_train 中各個 attribute 和 Y_train 的相關係數，其中絕對值最高(即正或負相關最高)的前五個分別為:
Married-civ-spouse :0.44469
Husband:0.40103
Never-married:0.31844
age:0.23403
Hours_per_week:0.22968

可以看到即使是第一名的 attribute 相關係數都不是特別高，我只針對相關係數高的加高次項或是將相關係數低的 attribute 拿掉準確率都沒有進步，因此可能需要多項特徵才能較正確的辨別出該人是否年收入>50k。
在訓練完模型觀察我的 weight 之後，我認為對結果影響最大的 attribute 應為 capital_gain