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