學號：B03902059 系級： 資工三 姓名：紀典佑

ENV: python3.5

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

答：

此次實作generative model是採用上課教過的Gaussian Distribution來當作機率分布的模型，在各個維度求得平均值與變異數，最後帶入上課最後導出的公式中（sigmoid(WX + b)）以0.5為界做分類。Kaggle上的準確率為：0.84165

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

答：

此次實作discriminative model是以logistic regression的方式實作，先將各個feature做Max-min的normalization，，並且將x不為Boolean的feature加上二次項。再以cross entropy 當作lost function，使用gradient descent求得最佳解，其中在gradient descent時，有加上mini-batch(十份)與adagradient，iteration次數為500次，相當於更新500x10次W的值。Kaggle上的準確率為：0.85602

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：

特徵標準化在cross-validation上的準確率為: 0.8531986112824438，而沒有做normalize在cross-validation上的準確率為: 0.6800742070951652，normalize是採用max-min的normalize，即，分母加上的原因是怕產生除以0的錯誤。而準確率的差異可能原因是有做normalize過的資料回介於0跟1之間，各個feature的weight初始值也在0跟1之間，所以收斂較快，而沒有做normalize過的data分佈較廣，使用0跟1之間的weight當初始值要較久才能收斂，也就是比較多的iteration，所以相同iteration在兩筆資料間會有此差距。

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：

實作正規化分別給於不同的lambda(0, 0.01, 0.1)，出來的validation error分別為0.8531986112824438, 0.8225793451841357, 0.7598969750915858，正規化並未使得結果更好的原因可能是因為資料分佈本來就比較趨於沒有正規化的結果，所以將正規化的term加上去gradient的式子後，反而降低了準確率。

5.請討論你認為哪個attribute對結果影響最大？

將結果的weight輸出發現capital gain的weight特別大，不只是在linear上有30.8左右，在quadratic上也有8.9左右的大小，可見這個feature對判斷所得有密切關係，而應用在常理上判斷，國家所得越高，國民平均所得也會越多，因此此樣feature對結果影響大。