學號：B03902059 系級： 資工三 姓名：紀典佑

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$\frac{\mathrm{x}-\min(\mathrm{X})}{\max(X)-\min(X)+10^{-8}}$，分母加上$10^{-8}$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。