學號：R05942056 系級： 電信所碩一 姓名：時丕澔

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

答：

我實作了Joint Gaussian distribution的 generative model。此種機率模型在高維度空間中也十分好表達成數學式，且其函數完全可以簡單的由兩個參數：期望值m與共變異矩陣Σ決定，故其訓練方式非常直觀：直接將機率模型中的參數用樣本平均和樣本共變異矩陣代入，再將機率分布函數代入loss function就可以得出分辨兩個class的函數。

　在使用助教給的feature set 實驗後，此方法在training set上的分辨正確率約84.25%

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

答：

　我實作了logistic regression，使用sigmoid function 將feature 線性組合之後的值從打到(0,1)。其訓練方式為batched Gradient descent，同時試著使用feature normalization和Adam來加速收斂。並且試著使用regularization且觀察其效果。

　使用這方法得出的正確率約為85.2~85.6%

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：

　輸入特徵標準化在logistic regression中對模型準確率理論上不會有任何影響。

　假設一組MxN維的 feature data 有經過標準化(一個行向量為一組feature)，則有：

　考慮此原始資料的logistic regression最佳解w,b有以下關係式：

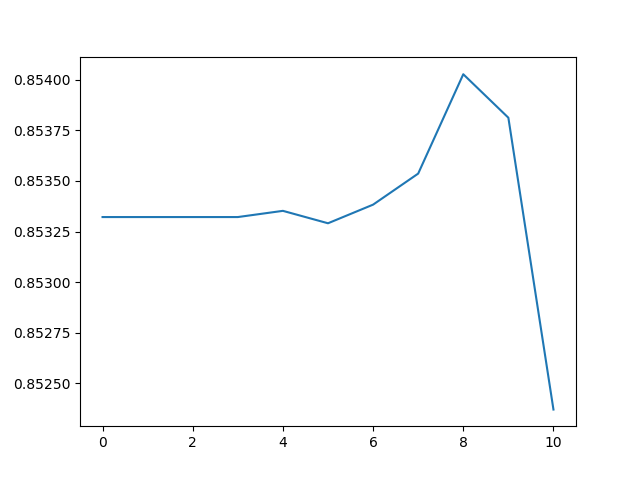
　如果令, 則因為這個關係為線性轉換，故兩者的解空間相同，因此在域上的最佳解應和在域上的最佳解有對應關係，因此最終得出來的函數也會是同一個。

　實作上，標準化加速了Gradient descent的效率。也衍伸出了一些實作上的問題例如浮點數運算的誤差等，只要控制好這些變因，輸入特徵標準化對這種簡單logistic regression的影響就趨近於零。

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：

　以下為實作regularization準確率與參數λ的關係圖，其中λ分別取了0.01, 0.03, 0.05, 0.1, 0.5, 1,5, 10, 20, 40, 100等十個數字，可以發現在這個問題中λ取在20附近會有較高的準確率。



5.請討論你認為哪個attribute對結果影響最大？

　我認為第二個feature, “fnlwgt”是影響比較顯著的attribute，如此判斷的原因是有作了一個實驗將所有attribute逐一去除，最後去掉此項的準確率下降得比其他attribute顯著的多。