學號：R05921077 系級： 電機所碩二 姓名：陳立杰

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

答：

我用所有 data 的 106 個 features 來做 106-D 的 Gaussian Distribution，最後在 training set 上的準確度為 0.840207610331378，在 public set 上的準確度為 0.84115。

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

答：

我用前面 8500 筆 data 的 106 個 features 來做 discriminative model，有加上 normalization 跟 regularization，最後在 training set 上的準確度為 rate= 0.8537，在 public set 上的準確率為 0.85332。

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：

原本我還沒做 feature normalization 的時候一直跑不出來，不是全猜 50K 以上就是 50K 以下，好像是因為有幾項數值太大導致的，但是那幾項又蠻重要的。後來先用 normalization 把資料標準化後，作起來就正常多了，而且數字都比較小，跑起來的速度跟沒有做 feature normalization 比快很多，最後在 public set 上可以跑到 0.85332 的準確率。

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：

當用前 8500 筆 data 的 106 個 feature 來 train 時，如果沒加 regularization 時在 training 跟 public set 的準確度分別是(0.85482,0.85086)，當 lambda=1,2,3 時的準確度分別是 (0.85470,0.85332),(0.85435,0.85344),(0.85423,0.85283)，從這些資料可以分析出當 regularization 的 lambda 越大時 training set 的準度卻會越差，但 public set 的準確度會越來越好，但是當 lambda 超過一定的值時還是會下降，但是準確度還是會比沒有加 regularization 時還要好。

5.請討論你認為哪個 attribute 對結果影響最大？

我覺得 capital_gain 的影響最大，因為經過 normalized 後通常是他的 weight 都會最大。