學號：B04902051 系級： 資工二 姓名：林承豫

1.請說明你實作的 generative model, 其訓練方式和準確率為何？
答：將 data normalize 後和調整輸入的 data 後，每個運用老師上課所教的 gaussian distribution model 算出平均跟變異數，在套入公式算出 testdata 的 output，在 kaggle 上準確率為 0.84582

2.請說明你實作的 discriminative model, 其訓練方式和準確率為何？
答：抽取前 65 項 model（即把國家別去掉，只留美國），跟前一題抽取的 attribute 基本上是一樣的，加上前第三到五項的平方和三次方作為 feature，並且經過 normalization，運用 logistic regression 加上 gradient descent 建出 model。在 kaggle 上準確率為 0.85639

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：如果同時用未經抽取過的 data 建 generative model 來測試（即 106 筆 data），在 kaggle 上表現為 0.83649，而經過標準化的 data 正確率為 0.83710 ，而在 discriminative model 上若未經過標準化，測試後發現需要跑大約三萬次到四萬次後才會收斂，因此便沒有傳上 kaggle 做測試。

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：沒做正規化之前 kaggle 上成績為 0.85639

正規化

lambda 設 0.1 正確率為 0.85639
lambda 設 1 正確率為 0.85614
lambda 設 10 正確率為 0.85676
lambda 設 100 正確率為 0.85160
可發現正規化對正確率影響不太顯著，而最高點是 lambda 為 10 的時候。

5.請討論你認為哪個 attribute 對結果影響最大？
答：寫了一支程式來判斷每個 attribute（只取 data 含 1 or 0 的 feature）去計算跟最後 flag 的相同程度，發現最相同的是 Prof-school 這個 attribute，而在測試時也曾刪除所有包含國家的 attribute, 後來發現留下美國這個 attribute 會讓結果更好一點，因此也選擇留下。