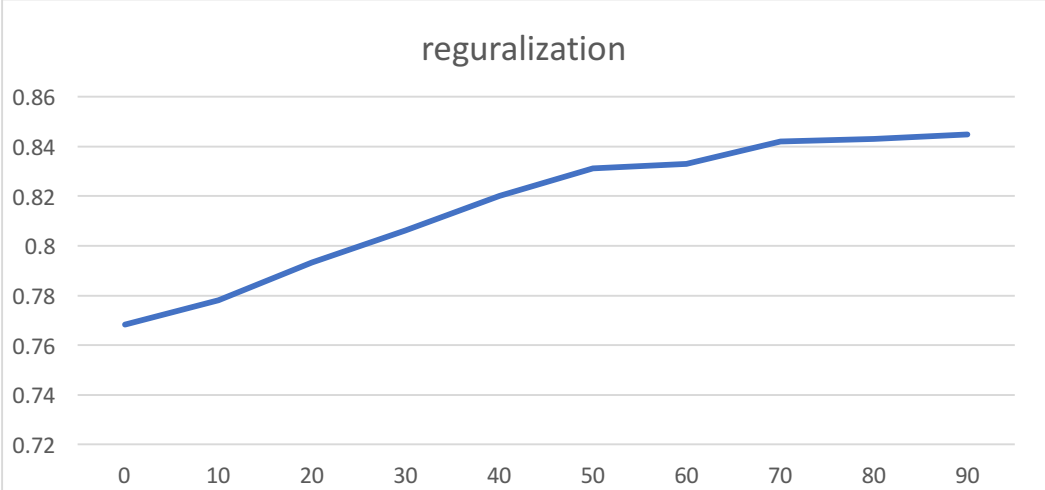
學號：R04921094 系級： 電機碩二 姓名：葉孟元

1.請說明你實作的 generative model，其訓練方式和準確率為何？
答：在計算 covariance 和 mean 是利用 label 將資料分成 0 和 1 的兩個 np array，然後用 np.mean 和 np.cov 來計算。一開始使用的是老師上課投影片的方式，但是只能勉強超過 simple base line，然後使用助教的利用 covariance 和 mean 來計算 w 和 b 再通過 sigmoid function 來計算的方式會更好，也許是 np.clipe 起的作用。在自己切割的 validation 上準確度為 0.83951,但是在 kaggle 上的準確度是 0.84658

2.請說明你實作的 discriminative model，其訓練方式和準確率為何？
答：利用 2d np.array 來表示不同 order 的 x，一開始就先計算出來儲存，避免重複計算。同樣創建一樣緯度的 array 表示 w。通過 feature normalization 和 regularization。最後 for 迴圈嘗試不同的參數，每隔一段時間利用 pickle.dump 將 w，b 和 validation score 全部存起來。最後選最好的 validation score 作為模型。由於 learning rate 有點難調整，在 train 的時候使用偏大的固定 learning rate。讓模型不斷震盪反而會找到相對較好的模型。自己 validation 準確度為 0.85697，在 kaggle 上為 0.85628

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：一開始使用老師課上講的(x-mean)/variance，連同所有的 vector 都進行 normalization，不過助教課之後發現其實真的沒用必要對 0/1 進行 normalization，只要對連續的數值做就好了。但是，normalization 對於 generative model 並沒有太大的影響，於是沒用採用，而是對 discriminative model 有比較大的影響。從原本的 0.84616 上升到 0.85697。

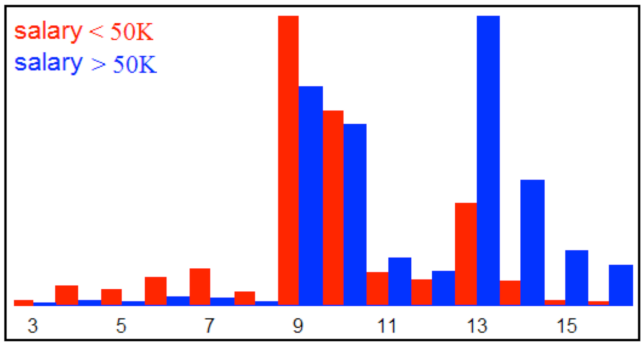
4. 請實作 logistic regression 的正規化(regularization)，並討論其對於你的模型準確率的影響。
答：為了能夠直觀察地表示 regularization 的影響在環境為 adagrad learning rate = 1，訓練 5000 次中選取最好的準確度 model，order = 8(為了確定可以發生 overfitting)。橫軸為 lambda，縱軸為準確度以下為圖表。
可以看到準確度隨著 lambda 增大而增大，當然，我認為 lambda 再增加下去會導致 under fitting。但是由於時間關係並沒有繼續做下去。



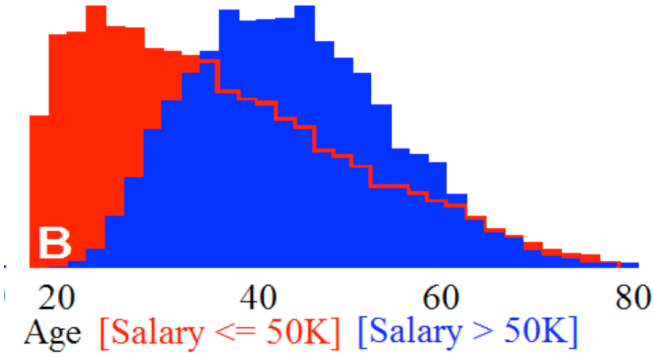
5.請討論你認為哪個 attribute 對結果影響最大？

答：我認為教育水平對結果影響最大（10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th）當然後面也有另一部分 Bachelors, Doctorate, HS-grad etc. 雖然也是教育水平，但是和上面所提供訊息類似。在網路上搜索到一篇名為 Predicting earning potential on Adult Dataset 的報告，裡面介紹了一部分 attribute 的影響。在此就引用其作為討論。在這份報告中提到了，contrary 有 90%為 U.S.A 所以沒用太大的可比性。fnlwgt 雖然有很大的關係，但是大部分資料都為 0. 可以從圖表上看到 age 對 salary 的影響並不如 education level。

（圖片都引用於報告）



education level vs. salary



age vs. salary.