學號：F05942036 系級： 電信博一 姓名：羅翊展

1.請比較你實作的 generative model、logistic regression 的準確率，何者較佳？
答：
Logistic regression 的準確率較佳，判斷原因是因為 logistic 有較多的參數可以調整，並且可以用 iteration 的來暴力解，參數比較多的話 bias 可能會較小，但 variance 會比較大，不會 variance 比較大的情形可以用 regression, normalization 和 training data 增加來解決。

2.請說明你實作的 best model，其訓練方式和準確率為何？
答：
我是用 logistic regression model 挑選二次的，用 adagrad 當 gradient descent，learning rate 為 0.001,regularization = 0.001, momentum=0.01 在 kaggle 上的成績為 8.547(public), 8.5309(private)，相較於一次的 model 我的方法很容易就超過 stronge baseline 了，而且並沒有一直丟 kaggle 來衝高分數，所以不怕 overfit，我想這是因為這次的資料量比較大，所以用比較複雜的 model bias 小的同時 variance 也不大

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：
我使用的是助教所提供的 normalize model 比較有加和沒加的差別(方法是算出 training data+testing data 的平均和標準差，然後將每一項減掉平均除上標準差，代表取出 scaling 後，變化的部分)，可以發現準確度有所提高，這是因為如果 feature 之間的差距過大在算 gradient 時，比較小的一方很容易較會被當成誤差，變得對 model 沒貢獻，normalized 之後就可以讓每個 feature 對 model 有所貢獻

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：
我只用的方法是在微分後的 loss 加上 regularization term 我發現加入 regression 的結果並沒有讓 model 的準確率上升很多，反而我加入的 momentum 能讓準確度提升，推估原因是因為 logistic regression 會有 local minima 所以加入 momentum 有機會能讓 training 的過程跳出 local minima，但我還是有比較不同 lamda 對 training 結果的影響:

RMSE\lambda	0.1	0.01	0.001	0.0001
RMSE	0.84371	0.84692	**0.848524**	0.846750

5.請討論你認為哪個 attribute 對結果影響最大？
我認為 age, working class, Education, relationship, race, sex 會是預測收入的幾個重要的 feature，經過 decision tree 判斷每一個 feature 的影響力，年齡(age) 對結果影響最大