學號：B03902016 系級： 資工四 姓名：周聖筌

1.請比較你實作的generative model、logistic regression的準確率，何者較佳？

答：generative model : public: 0.84508 private: 0.84240
logistic regression: public: 0.85503 private: 0.84940

logistic regression的準確率較佳，原因應該是generative model直接假設資料間的分佈是normal distribution，事實上卻不一定如此。而logistic regression沒有做事先的假設，直接從隨意起始值做更新，將model越train越接近實際的結果。

2.請說明你實作的best model，其訓練方式和準確率為何？

答：我的最佳model是使用kera這個framework來訓練的。總共兩層layer，activation function是使用sigmoid，而每層layer有600個neuron，最後output使用softmax跑出兩個值，其中一個代表>50k，另一個則是<50k。而訓練時的loss function使用課程所教的cross entropy，batch size設為100，epochs為20。

準確率：public: 0.86044 private: 0.85333

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：logistic regression:

normalization前：public: 0.79398 private: 0.79167

normalization後：public: 0.85503 private: 0.84940

generative model:

normalization前：public: 0.84508 private: 0.84240

normalization後：public: 0.84533 private: 0.84242

normalization對logistic regression影響很大，但generative model就沒什麼影響。我想因為是這個方法原本的目的就是要加快learn的效率，所以對於不用learn的generative model來說，就沒有什麼意義，而且做normalization對generative model中的機率分佈應該還是大致相同。

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：regularization前：public: 0.85503 private: 0.84940

regularization後：public: 0.85552 private: 0.85026

不同的lamda值會有不同的效果，我最後是將其設為0.2，發現結果會好一點。

5.請討論你認為哪個attribute對結果影響最大？

（此題的測試結果皆使用generative model的naive bayes）

我認為教育程度應該影響最大，當我只使用教育程度當feature去預測時，就得到的差不多七成的準確率。而在實際測試其他attribute後，我發現只使用人種、母國家和工作時數作為預測feature得到的準確率最差，都低於五成，可見這三項應該是最無相關的attribute。其他都差不多有六成以上的正確率。