Homework 2 Report - Income Prediction

學號：r06944051 系級：網媒碩一 姓名：郭柏辰

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

答：我使用助教提供的feature，並在training data中依據年收入大於50k且擁有完整資訊的項目，統計出大於50k項目中workclass, occupation, native_country集合中各元素出現的機率，並依照此機率指定資料不齊全的項目。若小於50k者，則將？部分指定該集合機率最低的元素，大於50k者則依照該集合機率指定對應元素。最後再重新統計workclass, occupation, native_country集合中各元素出現的機率，再依據此機率指派test data中資料不齊全的項目。

	Public score	Private score
Generative model	83.267%	82.25%
Logistic regression	85.921%	85.787%

根據實驗結果，我認為generative model之所以表現得較差，原因在於使用gaussian distribution，但在真實情況下可能是其它種類的機率分佈，反而造成結果不佳。

2. (1%) 請說明你實作的best model，其訓練方式和準確率為何？

答：實作方法跟logistic regression一樣，資料預先處理方式如上題，並增加連續features的二次方，三次方及取log，learning rate: 0.05 並執行30000次。Public score: 85.921%，Private score: 85.787%。

3. (1%) 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。(有關normalization請參考：https://goo.gl/XBM3aE)

答：使用logistic model 在沒有實作feature normalization時，Public score: 79.238%，Private score: 78.737%，有實作feature normalization則Public score: 85.921%，Private score: 85.787%，明顯有實作時準確率較高，主要原因在於這次的data中，連續的feature各項data值與值間的差距非常大，如果尚未做feature normalization，feature之間的差異會造成model不易進行訓練。

4. (1%) 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。(有關regularization請參考：https://goo.gl/SSWGhf P.35)

答：

Lambda	Public score	Private score
0.01	85.909	85.824
0.1	85.884	85.812
1	85.872	85.861
10	85.835	85.702
100	85.626	85.087

根據實驗結果，可以發現在 λ = 1時準確率是最高的，當 λ 太大時，可能造成function太過平滑，導致準確率逐漸下降。

5. (1%) 請討論你認為哪個attribute對結果影響最大？

答：我把weight列出並取絕對值，找出 | W | > 1的feature，發現capital gain的影響最大，weight高達2.35721102，其餘feature weight差不多為0 ~ 1之間。而capital gain代表投資，其中包含股票基金等等，這反而是影響年收入的重大原因之一。