學號：R06922002 系級： 資工碩一 姓名：姚嘉昇

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

答：由下表可看出 Logistic regression 不管在 training set 還是 testing set 上都有比 generative model 還要高的準確率，可能是因為 generative model 是在假設資料是在高斯分布的前提下去運算的，所以沒辦法在 training set 上很 fit 本次作業的資料分布。

	Training Score	Private Score	Public Score	Testing Score
Generative Model	0.83732072	0.83822	0.83906	0.838640
Logistic regression	0.86109149	0.85665	0.86056	0.858605

備註：Training Score：此 model 在 training set 中的 accuracy。

備註：Testing Score：此 model 在 testing set 中的 accuracy，為 private 與 public 之平均值。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

答：首先把 X_train 內的 106 維資料，再多加一筆 train.csv 中的 education_num，如此就有 107 維的資料了。再來這 107 維的資料中，有 6 種 features 是 continuous 項（沒有被轉成 one hot 的 features），把這 6 種 features 分別取 log、然後兩兩相乘再取 log 後（共 21 項），我就有 129 維的資料可以給後面做 logistic regression 了。會這麼做是因為 106 維下去怎麼 fit training set 都只能到 91%的準確率，若靠這些方法新增更多資料後，則可以在 training set 上 fit 到 95%以上。如此先 overfitting 後再來處理類似 dropout 與 early stopping 會比較快達到高準確率。

	Training Score	Private Score	Public Score	Testing Score
Best Model	0.86026228	0.85505	0.86130	0.858175

備註：Training Score：此 model 在 training set 中的 accuracy。

備註：Testing Score：此 model 在 testing set 中的 accuracy，為 private 與 public 之平均值。

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：在 Logistic regression 中，沒有做 normalization 的話，整個 model 是 train 不起來的，因為每個 feature 的分布不一樣，會造成 train 出來的 weights 差距會很大，這樣會造成 model 很不穩定，而且若有做 regularization 的話，同一個 lambda 對大小不一的 weights 會很不公平。

而在 Generative model 中，沒有做 normalization 的準確率卻比較高，讓我驚訝了一下。我覺得可能的原因是在推導 mu 與 sigma 時，是每個 feature 分開算的，互相影響不大，而且可能不做 normalization 的話，原本的機率分布會更像高斯分布。

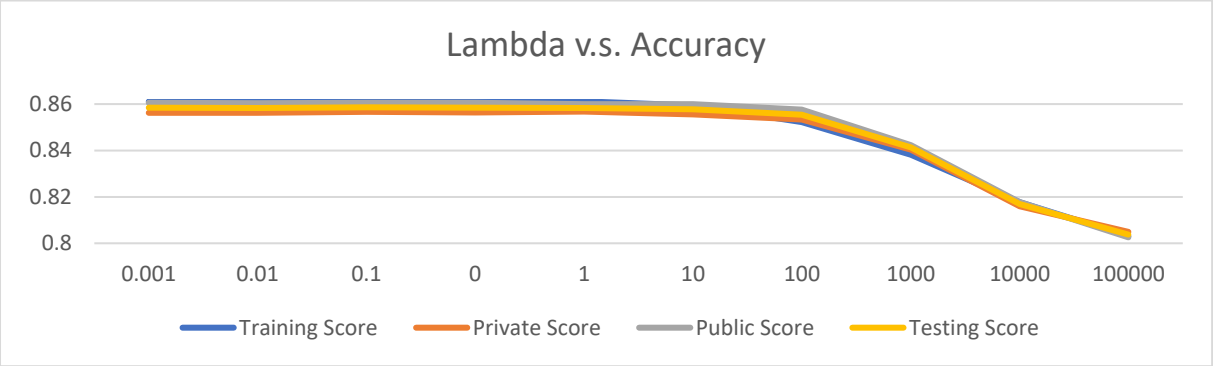
	Normalization	Training Score	Private Score	Public Score	Testing Score
Logistic regression	Before	0.56054789	0.52794	0.52309	0.525515
	After	0.86109149	0.85665	0.86056	0.858605
Generative model	Before	0.84257240	0.84240	0.84520	0.843800
	After	0.83732072	0.83822	0.83906	0.838640

備註：Training Score：此 model 在 training set 中的 accuracy。

備註：Testing Score：此 model 在 testing set 中的 accuracy，為 private 與 public 之平均值。

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：適當的 lambda 可以讓 logistic regression 求出來的曲線更加平滑，減少 weights 變化太快而造成 overfitting 的機會。但 lambda 太大會讓曲線過度平滑，而使準確率又下降了。以下圖來看的話，lambda 選超過 100 之後就開始壞掉了，而選小於 1 的結果都不錯，但看表的話選擇 0.1 是最合適的。



Lambda	Training Score	Private Score	Public Score	Testing Score
100000	0.80304659	0.80493	0.80257	0.803750
10000	0.81775744	0.81599	0.81781	0.816900
1000	0.83814993	0.84031	0.84226	0.841285
100	0.85221584	0.85333	0.85761	0.855470
10	0.85937164	0.85554	0.85995	0.857745
1	0.86118362	0.85677	0.85995	0.858360
0	0.86099936	0.85640	0.86056	0.858480
0.1	0.86109149	0.85665	0.86056	0.858605
0.01	0.86096864	0.85628	0.86044	0.858360
0.001	0.86099936	0.85628	0.86056	0.858420

備註：Training Score：此 model 在 training set 中的 accuracy。

備註：Testing Score：此 model 在 testing set 中的 accuracy，為 private 與 public 之平均值。

5. 請討論你認為哪個 attribute 對結果影響最大？

答：下表是 best model 前二十大(絕對值最大)的 weights 的 feature 名稱，與其對應的 attribute。前三名都是 capital_gain，可見其影響是最大的。

Attribute	Feature	weight	ABS(weight)
capital_gain	capital_gain^4	-1.18E+01	11.8208828
capital_gain	capital_gain	7.10E+00	7.10410778
capital_gain	capital_gain^3	6.27E+00	6.27093842
marital-status	Married-AF-spouse	3.58E+00	3.57964355
occupation	Priv-house-serv	-3.45E+00	3.4543021
workclass	Without-pay	-3.30E+00	3.30353658
race	Amer-Indian-Eskimo	-2.89E+00	2.88572785
relationship	Wife	2.84E+00	2.83586868
marital-status	Married-civ-spouse	2.65E+00	2.64792331
race	Other	-2.62E+00	2.62215811
race	Black	-2.50E+00	2.49893698
race	White	-2.25E+00	2.24837884
race	Asian-Pac-Islander	-2.17E+00	2.16735351
age	age.log	2.13E+00	2.13419132
education	Assoc-acdm	-2.11E+00	2.11269468
relationship	Not-in-family	1.97E+00	1.97190395
hours_per_week	hours_per_week^3	-1.96E+00	1.95647014
native-country	Outlying-US(Guam-USVI-etc)	-1.94E+00	1.93938965
workclass	Self-emp-not-inc	-1.94E+00	1.93825741
workclass	State-gov	-1.86E+00	1.86261748