HW2 Report

學號：b07901039 系級：電機二 姓名：劉知穎

1. 請比較實作的 generative model 及 logistic regression 的準確率，何者較佳？請解釋為何有這種情況？

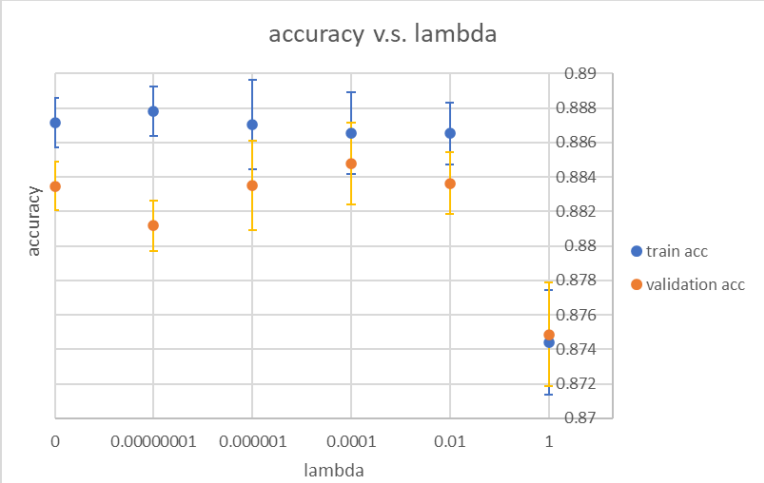
   兩者都沒做 feature engineering、shuffle data 並重複執行 10 次。

   Logistic Regression 不論是 training accuracy 或者 validation accuracy 都較 generative model 更好。因為 generative model 假設 probability distribution 為 Gaussian 分布，因此在這次作業中較無法符合真實的資料。

	Logistic Regression		Generative Model	
	train acc	val acc	train acc	val acc
0	0.885541	0.882049	0.871118	0.866384
1	0.885886	0.882602	0.870081	0.866845
2	0.884389	0.889882	0.868561	0.866568
3	0.88713	0.879285	0.871832	0.870254
4	0.884734	0.886657	0.8725	0.863988
5	0.886093	0.879008	0.870841	0.868688
6	0.885909	0.882879	0.870703	0.867674
7	0.885379	0.884169	0.872777	0.87523
8	0.8866	0.878179	0.871855	0.869886
9	0.88561	0.884353	0.873099	0.870899
avg	0.885727	0.882906	0.871337	0.868642
std	0.000766	0.003434	0.001305	0.002964

2. 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。接著嘗試對正規項使用不同的權重 (lambda)，並討論其影響。

	lambda = 0		lambda = 1e-08		lambda = 1e-06		lambda = 0.0001		lambda = 0.01		lambda = 1	
	train acc	val acc	train acc	val acc	train acc	val acc	train acc	val acc	train acc	val acc	train acc	val acc
0	0.8870381	0.883432	0.887821	0.882326	0.88713	0.885367	0.885264	0.888961	0.886163	0.884261	0.873076	0.879561
1	0.8867155	0.885182	0.887384	0.882418	0.887729	0.880667	0.887061	0.882234	0.886831	0.881865	0.874666	0.874862
2	0.8874067	0.884722	0.888467	0.878916	0.885656	0.88767	0.886854	0.885735	0.885955	0.88592	0.874251	0.875875
3	0.8871763	0.882787	0.887545	0.882234	0.887545	0.882142	0.88713	0.883616	0.887199	0.88122	0.874873	0.873756
4	0.8873837	0.88122	0.887844	0.87993	0.887107	0.881773	0.886393	0.883339	0.886462	0.884906	0.87515	0.870254
mean	0.887144	0.883468	0.887812	0.881165	0.887033	0.883524	0.88654	0.884777	0.886522	0.883634	0.874403	0.874862
std	0.0002539	0.001416	0.00037	0.001459	0.000729	0.002597	0.000688	0.002379	0.000449	0.0018	0.000725	0.003019



將 continuous features 加上二、三、四、五次項（因為原先只加上二、三次項時 regularization 的效果不明顯。）lambda 分別取 0（無 regularization）、1e-08、1e-06、1e-04、0.01、1，shuffle data 並各重複執行 5 次。

lambda 在 1e-06 到 1e-02 的區間能得到比未加 regularization 還好的結果，而最佳的 lambda 值大約是 1e-04。

3. 請說明你實作的 best model，其訓練方式和準確率為何？
   (1) Model：手刻的 logistic regression。
   (2) Feature engineering：對所有 continuous 的 features 加上二次、三次項和加加上"wage per hour × weeks worked in year"這一項。有做 feature normalization。
   (3) 加上 regularization，lambda 設為10^{-4}。
   (4) 重複執行五次，平均五次的 w 和 b 當做最終的 w 和 b。
   (5) 準確率：0.89066(kaggle 上的 public test)

4. 請實作輸入特徵標準化 (feature normalization)，並比較是否應用此技巧，會對於你的模型有何影響。

	With normalization		Without normalizatio	
	train acc	val acc	train acc	val acc
0	0.886854	0.882879	0.730624	0.730096
1	0.88584	0.885551	0.779053	0.776355
2	0.8869	0.883432	0.7262	0.72512
3	0.8866	0.881773	0.778937	0.778382
4	0.887199	0.880483	0.778661	0.786399
5	0.887107	0.882142	0.783407	0.777553
6	0.887107	0.88251	0.779951	0.790361
7	0.887568	0.878087	0.771565	0.775341
8	0.887522	0.88122	0.780896	0.775986
9	0.887153	0.879193	0.722929	0.726778
avg	0.886985	0.881727	0.763222	0.764237
std	0.000471	0.002028	0.024212	0.024613

將 continuous features 加上二次、三次項，shuffle data 並各執行 10 次。

加上 feature normalization 的模型的 training 和 validation accuracy 都較沒有用 feature normalization 的為好。因為 feature normalization 使各個 feature 有相同的 scaling，如此在做 regression 時，各個 feature 對於結果的影響力會較平均，能加快收斂的速度和準確率。