學號：B07902006 系級：資工二 姓名：林則仰

1. (2%) 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？

	Training accuracy	Development accuracy
Logistic regression	0.885644	0.876152
Generative regression	0.872496	0.863803

在這次的作業中，我們可以看見以hw2助教給的training set而言，logistic regression 是優於 generative regression的。會有這樣的結果，我認為與dataset的大小有關。由於在generative regression 的過程中，是以分開的出現機率相乘來計算，因此會有類似"腦補"的情形出現。因此，在data較小的時候，generative regression 的表現可能會更好。而logistic regression 是採用梯度下降法一步一步找出最適合的向量，一定要有看見資料才會更新，而不會有類似"腦補"的情形，所以給的資料越大理論上最後使用logsitic regression train出的model也會相較之下在預測上越準確。在這次的作業中，由於資料量較大，故logistic regression 的準確率較高。

2. (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	Training accuracy	Development accuracy
0	0.885848	0.876336
0.1	0.880544	0.874493
0.25	0.876899	0.870622
0.5	0.873417	0.866568

在上表中可以發現，當在原本的loss function上加上正歸化的項之後，當lambda越大，在準確率上的表現反而是越來越不好。我認為其原因是因為我的model可能沒那麼複雜，曲線本來就已經比較圓滑了，所以本身就不太會發生

overfitting的現象。而其實加入較小的lambda時，對於準確率的影響也不大，我認為是因為這份data大部分資料都是0與1，因此曲線不太會有不平滑的現象，故使用正規化的技術可能對於這次作業沒什麼太大幫助。

3. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？

   關於我實作的best model，我的訓練方式是feature engineering 以及 one-hot encoding 的技巧。利用one-hot encoding，將原本連續的資料也變為0與1的資料，如此一來不但不需要normalize，而且也能更準確的預測。舉例來說，由於年齡對於收入的關係並非線性，但若是沒有使用one-hot encoding ，也沒有採取添加二次項等技巧，就會有年齡越大或年齡越小預測出收入大於50000美金機率越大的情形存在，但這是不合理的，因此one-hot 可以解決這樣的問題。

4. (1%) 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。

Generative regression	Training Accuracy	Development Accuracy
Normalization	0.872496	0.863803
No normalization	0.876326	0.866384

Logistic regression	Training Accuracy	Development Accuracy
Normalization	0.885644	0.876152
No normalization	0.804443	0.798562

從以上兩個表格可以得知，對於generative regression 的 model 而言，可能是因為沒有使用gradient descent 的關係，有沒有做特徵標準化並沒有什麼影響，甚至沒做標準化的在準確率上的表現還略勝一籌。但是反觀logistic regression 的 model ，有沒有做normalization 就差非常多，我想原因是因為在資料中有7筆是連續性的資料，而非由0跟1組成，因此若沒有做標準化則可能導致不同feature間的數值差距過大，也就是連續性的資料與非連續性的可能需要不同的learning rate 來訓練模型，而標準化則可以順利解決此問題，使準確率大幅提升。