學號：B05902010　系級：資工四　　姓名：張頌平

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？

A:
在同樣都有做feature selection的情況下，以下為 generative model 及 logistic regression在development set (generative model 是尋找最佳解，所以可以直接用training set)和public test set 的分數：

generative:
training set acc: 0.87292
public test set acc: 0.88262

logistic regression:
development set acc: 0.88002
public test set acc: 0.89233

從上面的分數比較可以看到，logistic regression的準確率相較於generative model好上1%左右，其中的原因，我覺得是因為generative model主要是利用訓練資料的distribution來預測分類，logistic regression則是在一次又一次的錯誤中學會如何預測(discrimination model)，其實兩者是從相同的function set尋找最佳解，然而，distribution某種程度在對於尋找weight時就已經給予其一些限制(老師在影片是說腦補xD)，因此，在這次作業的資料量夠多且並沒有太多noise的情況下，logistic regression的方法，直接在訓練中學習最好的weight，並且透過development set 來防止overfitting的結果才能有更好的準確率。

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。(有關 regularization 請參考 https://goo.gl/SSWGhf p.35)

A:
regularization對於模型影響不大，推測是因為模型不夠複雜，因此overfitting的情況並不嚴重，反倒在訓練過程中產生些許noise讓模型有點小退步，其比較結果如下：

Without regularization:
development set acc: 0.88002
public test set acc: 0.89233

lambda=0.01
development set acc: 0.87836
public test set acc: 0.89146

lambda=0.001
development set acc: 0.87983
public test set acc: 0.89233

lambda=0.0001
development set acc: 0.87992
public test set acc: 0.89240

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

A:
在沒有使用sklearn套件的訓練時，我最好的model是使用logistic regression再加上一些feature engineering的技巧：

- 移除太多缺失值或是和其他特徵相關性過高的特徵，ex: 父母出生地
- 移除太過general或是沒有特別意義的one hot特徵，ex: 有包含?的特徵、包含Do not know的特徵、出生地在美國等等...
- 將數值型的特徵做binning，像是每小時薪資，因為很多人的資料為0，但其實應該是他們沒有填此項資料，而不是真的每小時薪資為0，為了縮小因為不對等的資料或是在數值中有小差距但實際上並無太大影響(像是年齡)，而導致模型訓練時可能有偏差，因此將特定的數值型特徵做binning達到更好的結果

在使用完上述技巧後，準確率如下：
Settings:
batch_size:8, learning rate: 0.18, dev size: 0.2

development set acc: 0.88022
public test set acc: 0.89233

後來在討論區助教說明了可以使用sklearn套件訓練，因此後來使用gradient boosting classifier來訓練，在做相同的feature engineering的情況下，分數進步非常多，準確率如下：
Settings:
n_estimators:200, learning rate: 0.18, dev size: 0.2, random_state: 0

development set acc: ***0.88785***
public test set acc: ***0.89979***

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。

A:
一般常用的feature normalization有兩種，分別是min-max normalization和standard deviation normalization，而我在這次作業的程式中實做的 Feature normalization是第二種normalization的方法，將特徵的平均值縮放為0，標準差為1，標準化能夠讓每個特徵在訓練時不會因為資料數值大小的不同，而使得某一特徵過於影響訓練，因而能夠讓結果更為準確，而實際的實驗結果如下(使用development set 比較)：

無normalization:
development set acc: 0.81597

有normalization:
development set acc: 0.88022