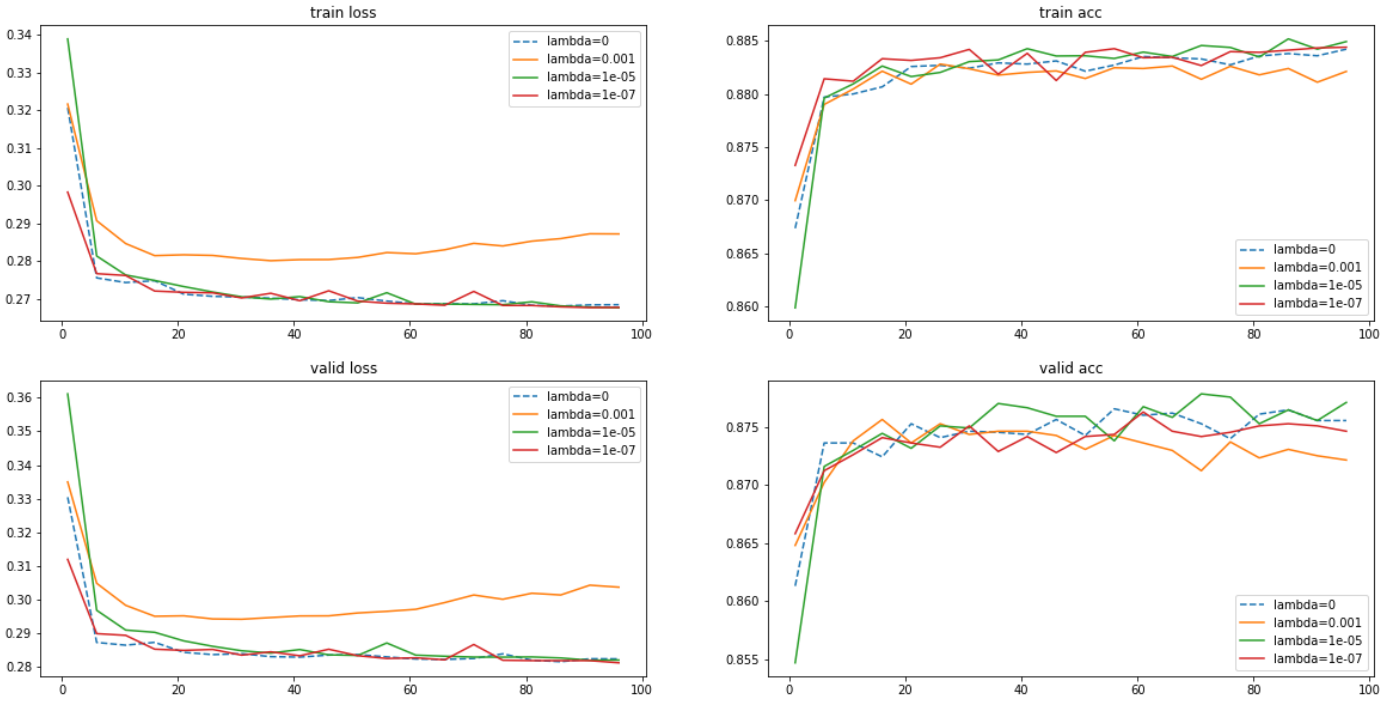
學號：B07902047 系級：資工二 姓名：羅啟帆

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？

   Logistic regression 有較佳的準確率。可能是我們的資料量夠多，使得 discriminative model（也就是 logistic regression）表現的比 generative model 好。為了驗證這個想法，我拿原本 training data 的 5% 當新的 training data(2713)，剩下的 95% 當作 validation data(51543)。結果 discriminative model(83.55%) 的表現略差於 generative model(84.64%)。

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

   (Validation data = 20% of train data)



   首先，可以觀察到加上 regularization term 會讓 train 的表現變得較差，而 valid 的表現則有些比沒加 regularization 來得好 (ex. lambda=1e-5)。

   再來，我們可以觀察到太大的 lambda 會造成較高 loss，而恰當的 lambda 會有好的 validation 表現。

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

   - 移除這些 feature: [' Not in universe', ' Not in universe.1', ' Not in universe.2', ' Not in universe.3', ' Not in universe.4', ' Not in universe.5', ' Not in universe.6', ' Not in universe.7', ' Not in universe.8', ' Not in universe.9', ' Not in universe.10', ' Not in universe.11', ' Not in universe.12', ' ?', ' ?.1', ' ?.2', ' ?.3', ' ?.4', ' ?.5', ' ?.6', ' ?.7', ' Do not

know', ' Not in universe or children', ' Not in labor force', ' Not identifiable', ' Not in universe under 1 year old', ' Foreign born- Not a citizen of U S ']

- 新增這些 feature:
  - pd.get_dummies(pd.cut(X['age'], [-np.inf, 20, 40, 50, 70, np.inf], labels=['<20', '20-40', '40-50', '50-70', '70+'])),
  - (X['weeks worked in year'] > 0).astype(int).rename('B - weeks worked in year'),
  - (X['wage per hour'] > 0).astype(int).rename('B - wage per hour'),
  - (X['capital gains'] > 0).astype(int).rename('B - capital gains'),
  - (X['dividends from stocks'] > 0).astype(int).rename('B - dividends from stocks')
- 移除 unary 的 feature
- 對有三種以上 value 的 feature 做 normalization
- 切一小塊 data 當 validation data(1%)
- 做 Logistic gradient descent, lr=0.001, epochs=10000, batch_size=512

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。

Valid loss/accuracy (valid ratio=0.1)	With feature normalization	Without feature normalization
Logistic regression	0.3258/0.8752	2.3404/0.7934 (overflow in sigmoid)
Generative model	x/0.8590	x/0.8664

對 Logistic 模型來說，沒有 feature normalization 的話會造成 exp 計算的時後溢出，導致結果爛掉。而對 Generative 來說，feature normalization 會有比較好的結果