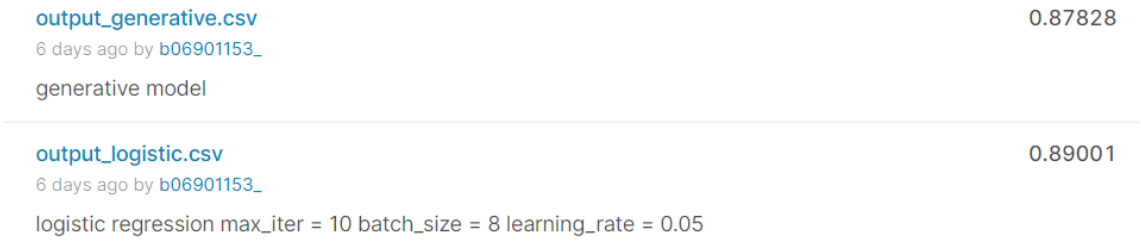
學號：B06901153 系級：電機三 姓名：林瑩昇

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 的結果較好。我覺得原因可能有幾項

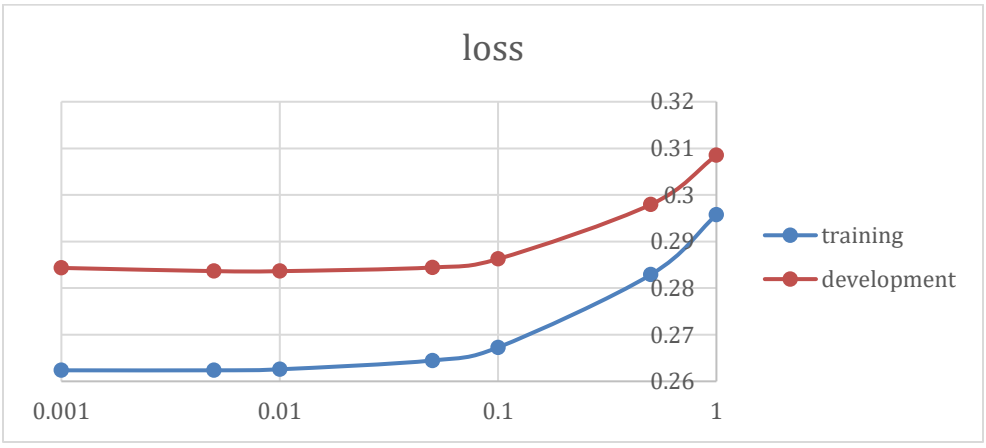
(1) Generative model 的前提是各個 feature 之間是獨立的，但如果我們去看 train.csv 中就可以看出，其實很多 feature 之間並不獨立。像是 marital state 與 age 之間就不一定是獨立的。

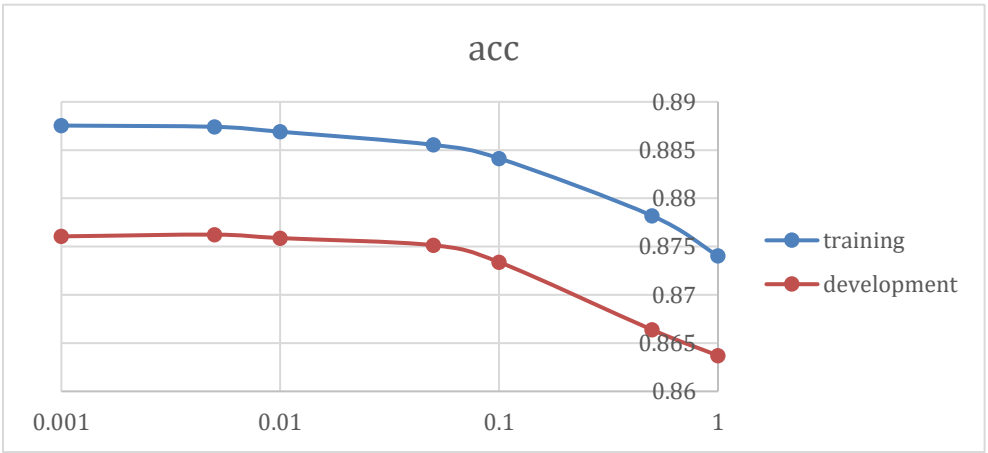
(2) 當 data 少的時候，generative model 比 logistic regression 有優勢，因為 logistic regression 容易 overfit。但我們今天的 data 其實 feature 很多，甚至許多是不相干的，因此 generative model 表現也就比較差。

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)

   Regularization 能夠讓 weight 往小的地方收斂，所以也就會比較 smooth。可以從下圖中看到 lambda = 0.001、0.05、0.01 都能在 training 和 dev set 上表現不錯，但其實在 testing set 上，0.1 表現會最好。原本如果直接用 codelab 上的 code 跑，loss curve 即使 lr 調小後仍會有上上下下的狀況。使用 regularization 後，loss 收斂的情形變得比較好，可能是因為 weight 比較小，所以不會因為更新參數而導致 loss 上升。

   以下是 max_iter = 40、batch_size = 8、learning_rate = 0.1





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

(1) 用的是 logistic regression

(2) Regularization：lambda 經過實驗之後決定用 0.01

(3) max_iter = 50、batch_size = 2048、learning_rate = 0.05。其中我每 10 個 epoch 就會把 lr 減半。

(4) feature expanding：

我原本打算要多加二次方及三次方(總共 1530 個 feature)，但是結果並不理想。之後我發現 data 中有幾項大小相差非常多，所以就決定用 0.3 跟 0.7 試試看，結果竟然非常好。

後來我在想，或許有些 data 的極值會造成 model 對這些 feature 不敏感。像是其中有一個相關係數很高的 feature 是 capital gain，數字的範圍是 0~9999。假若其實 capital gain 超過 100，最後>50k 的機率就很高，但因為 feature 數字範圍到 9999 的緣故，導致 model 對 100 不太敏感而輸出錯誤的結果。

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。

   原先我認為這次 data 很多 one hot encoding 的 feature，所以應該只 normalize 那些不是 one hot encoding 的 feature，像是 age。結果卻與我的想法相反，如果只挑選特定 feature 去標準化，最後不管在 training、development 或是 testing set 上的表現都變差。

   Normalize 特定 feature：

   (1) Training loss：0.29405

   (2) Development loss：0.30280

   (3) Training acc：0.87397

   (4) Development acc：0.86509

   Normalize 所有 feature

   (1) Training loss：0.26727

   (2) Development loss：0.28625

   (3) Training acc：0.88416

   (4) Development acc：0.87330

   以上數據用的是 max_iter = 40、batch_size = 8、lamda = 0.1

   norm_column = [0, 126, 210, 211, 212, 358, 507] (不是 one lot encoding 的 feature)