學號：B086502158 系級：機械三 姓名：陳柏元

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?

   表格的數值為 Before normalization / After normalization

	Training Accuracy Before normalization	Training Accuracy After normalization
generative	0.87435	0.87435
logistic	0.76549	0.88361

表一

由表一可以看出在標準化之後的 logistic model 大幅提升了準確率和效能,得到了較好的分數,但是 generative model 在標準化後的分數卻沒有什麼改變,因此我認為雖然 generative model 的參數容易因為資料偏移的影響而改變,但是在同一筆資料下 generative model 受到標準化的影響卻微乎其微,準確率並沒有任何改變，可能原因是或許 generative model 對於遠偏離平均的異樣樣本影響較小。而 logistic model 對於異常遠離平均的值，如 feature 210,212 等，失準會大幅提高，故 normalize 對 logistic model 有顯著成效。整體來說,如果要達到較好的準確率,logistic model 會比 generative model 更靈活,較能透過調整訓練參數、標準化等方法逼近最佳解。但若資料少且分歧，generative model 則據較能取得最佳解。

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)

λ	0.001	0.01	0.1	1
train	0.88361	0.88357	0.88334	0.88291
validation	0.87338	0.87433	0.87658	0.87938

在這邊我們可以發現隨著 λ 提高,validation 的準確率會有些許提升,train 的會有些許下降,但是幅度並不高,我認為是因為 Logistic model 沒有嚴重的 Overfitting 情況,因此 regularization 在此 Dataset 有得到些許微量的進步 。但或許再更偏差的 DataSet 進步幅度可能更大。

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

在這次實作上，我沿用了 logistic regression 進行了再優化，首先因為 normalize 值以後仍在 feature 210,212 等等中有許多的極偏差值，為了不讓這些異常值影響 logistic 劃分而值的線也跟著偏差，我將所有 feature 中標準差大於 3 的值進行 scaling，而經過一番測試以後，發現當加一並取二為底 log 時有明顯的準確率增加，因為 log 能將極偏差的值的縮小。接著為了再提升準確率，我將 feature 中影響力較大的 200 個 weight 記錄下來並取二次式，加到 feature 當中，再 development set(validation)得到更優化的準確率，但 training set 儘微幅增加，影響不大。最後修改了 learning rate =0.059 以及 batch_size=7，發現在第 8 個 iteration 時達到最高並過了 strong baseline。

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。

Accuracy logistic	no normalization	normalization	log After normalization	log Before normalization
Training	0.76549	0.88361	0.88640	0.88588
Validation	0.76409	0.87338	0.88260	0.87946

由上表一可以看出在標準化之後的 logistic model 大幅提升了準確率和效能,得到了較好的分數,但是 generative model 在標準化後的分數卻沒有什麼改變,因此在實作優化 logistic model 中，我仍使用 feature normalization 。但有趣的是，在我的 best.sh 中，如果我是先取 log10 再 normalize 的話，則完全不會優化，我想可能的原因是在偏差值以內的值如果先取 log10 了將導致其失真，故 normalize 後更加失真導致準率下降。