Machine Learning HW 3

B04705003 資工三 林子雋

1. 請比較你實作的 generative model、logistic regression 的準確率，何者較佳？

   Ａ：結果發現，logistic regression 的表現比較好，不論是在 feature 怎麼選取的情況下，推測是因為 logistic regression 可以 fit 上不是 gaussian distribution 的資料上，而且其實這次的 feature 有許多是 binary encoding 的，其實很明顯的，不適合被 model 成 gaussian distribution，因此可能因為這樣 generative 表現比較差。

Figure 1. 為 Regularization = 0，跑過 5000 epochs 的情形，有 Normalize		
Model/Feature 選取	所有的 feature(106 維)	所有的 feature+index=[0,1,3,4,5]平方項
Logistic Regression	0.852158	0.856028
Generative Model	0.843682	0.843682

2. 請說明你實作的 best model，其訓練方式和準確率為何？

   A：最好的 model 是使用 Xgboost classifier，裡面實際運作方式是使用 gradient boosting 的方法。準確率為 0.87617。

3. 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。

   A：下圖為沒有做 feature normalization 的情形，發現沒有做 normalization 的效果對於 logistic regression 的結果是變差的。對於 logistic normalization，推測是因為 epoch 不夠久以至於還未走到最佳解便結束；而 generative model 則是相對穩定， 甚至加上平方項的還超過沒有 normalization 的實驗，猜測是因為 generative model 在 normalize 之後，distribution 變形反而比較容易失去原來機率分布的樣子(可能原本的分布比較像 normal distribution，經過 normalization 之後機率分布就會稍稍變形)

Figure 2. 為 Regularization = 0，跑過 5000 epochs 的情形，但沒有 Normalize		
Model/Feature 選取	所有 feature(106 維)	所有的

		feature+index=[0,1,3,4,5]平方項
Logistic Regression	0.78551	0.79245
Generative Model	0.84411	0.85375

4. 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。

   A：從下面數據中可以發現，regularization 大致落在 0.0001 和 0.001 中的效果會比較好，而 regularization 太小或太大都會造成表現下降，尤其是太大時，會使 model performance 劇烈下降。

Figure 3. 下圖為各種 λ 對 testing set accuracy 的影響，皆使用 106 維的 feature					
Model\λ	0	0.0001	0.001	0.01	0.1
Logistic Regression	0.85215	0.85234	0.85191	0.84884	0.83176

Figure 4. 下圖為各種 λ 對 testing set accuracy 的影響，皆使用 106 維 +index=[0,1,3,4,5]平方項的 feature					
Model\λ	0	0.0001	0.001	0.01	0.1
Logistic Regression	0.856028	0.85639	0.8555	0.84914	0.83072

5. 請討論你認為哪個 attribute 對結果影響最大？

   A：根據前面幾項數據推測，我認為是 normalization 的影響最大，從上面可以看出，有無 normalization 會影響到準確率收斂的速度，而好的 normalization 可以讓收斂的狀況更為準確。