Homework 2 Report Problem Set

Professor Pei-Yuan Wu
EE5184 - Machine Learning

B04203058
蘇軒

Problem 1. (1%) 請簡單描述你實作之 logistic regression 以及 generative model 於此 task 的表現,並試著討論可能原因。

取所有 feature 在 logistic 方面用了 mini batch + normalization 的 training method 得到下面結果
Logistic : 0.82120
Generative : 0.8120
會發現其實 logistic 的表現會比 generative 好一些可能原因如下:
因為 generative model 是用 Gaussian distribution 去預估 這只能適用在 data 數量比較小的情況下，而且也不能確定真正的數據分布是照著 Guassian distribution，而相反的 logistic model 是一步一步去找到 cross entropy 最小的時候，所以照理說表現會比較好一些。

Problem 2. (1%) 請試著將 input feature 中的 gender, education, martial status 等改為 one-hot encoding 進行 training process,比較其模型準確率及其可能影響原因。

在未經過 one-hot encoding 之前的 public score 為 0.81760 而在經過 one-hot encoding 之後的 public score 為 0.82120 因為就 one-hot encoding 把原本只有單一變數的 weight 展開成有很多 weight 的情況可以很顯然地知道這樣的效果會比較好。

Problem 3. (1%) 請試著討論哪些 input features 的影響較大(實驗方法沒有特別限制,但請簡單闡述實驗方法)。

透過對每一個 input feature 對結果所做相關係數測驗 發現 pay_0, pay_2, pay_3, pay_6 的排名都相當前面，可是把他們一一挑出並且做 one-hot encoding 的結果 只有 pay_0 的結果比較顯著，而其他的 feature 反而沒有那麼明顯的結果。

Problem 4. (1%) 請實作特徵標準化 (feature normalization),並討論其對於模型準確率的影響與可能原因。

取全部 feature 拿去用 mini batch training method 得到下面結果
有 feature normalization: 0.8120
無 feature normalization: 0.7860
很明顯的我們可以發現有沒有做 normalization 將會嚴重的影響我們的結果，原因可能是因為如果沒有拿去做 normalization 的話整個 model 會被 scale 較大的幾個 feature 拖著造成誤差。

Problem 5. (1%)
Problem 6. (1%)