

機器學習 HW2

學號：R04522631 系級：機械碩二 姓名：盧玄真

1.請說明你實作的 generative model，其訓練方式和準確率為何？

答：

訓練方式：

利用助教幫我們抽出來的 feature 和 label 進行訓練。在這邊我選擇不切出 Validating data，因為收入大於 50K(class 2)的資料量較少，如果在 Validating data 中切出過多的 class 2 會讓其平均以及標準差失真，訓練就會失敗。然後將兩個 class 的平均值跟標準差算出來之後代回公式就可以得到 model。

訓練準確率：84.1159%

Kaggle 準確率：84.189%

2.請說明你實作的 discriminative model，其訓練方式和準確率為何？

答：

訓練方式：

同樣利用助教幫我們抽出來的 feature 和 label 進行訓練。資料處理方面，在此 model 中有隨機切出 1/5 的資料作為 Validating data，並對連續資料作特徵標準化。而在演算法的部分，我以 5000 筆資料為一批訓練 20000 個 epoch 實作批量梯度下降法，並且使用 adagrad 優化演算過程，最後加上正規化減少 overfitting 產生。

訓練準確率：85.3305%

Validation 準確率：85.2%

Kaggle 準確率：85.319%

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。

答：

標準化方法：

本次特徵標準化的目標主要是對連續資料作標準化，因為其他的離散資料為 0 或 1，因此連續資料標準化時最好將其縮減在 0 到 1 之間，所以標準化方法就如下

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

其中X為某一項特徵， X_{max} 以及 X_{min} 分別代表此特徵中的最大值以及最小值。這樣就能確保標準化區間落在 0 到 1 之間。

準確率影響程度：

當沒有作特徵標準化時，訓練準確率波動比較大如圖 1(a)因此在 10000 個 epoch 後可能會出現壞掉的訓練結果(50%或更低)，而最好的訓練結果大約是在 81%左右。

而有作特徵標準化時，訓練準確率波動小很多如圖 1(b)，而在 10000 個 epoch 後訓練結果大約為 85.5%明顯比起沒有做特徵標準化的訓練表現還好，因為標準化之後的訓練路徑會比沒標準化的還要平順很多，因此能在同樣的訓練次數下達到比較好的表現。

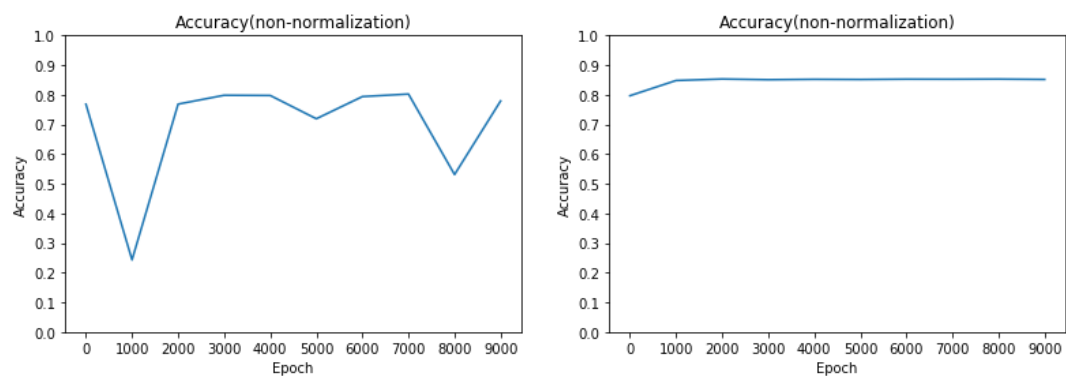


圖 1 訓練準確率(a) 無標準化 (b) 有標準化

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。

答：

由下表可以看出正規化參數在 0.0001 時模型準確率會有最佳表現，太小會 overfitting，太大則 underfitting。

Lamda	Training Accuracy	Testing Accuracy
0	85.261%	85.172%
0.00001	85.180%	85.332%
0.0001	85.265%	85.418%
0.001	85.261%	85.270%
0.01	85.161%	85.233%
0.1	85.180%	85.111%
1	84.651%	84.324%

5. 請討論你認為哪個 attribute 對結果影響最大？

我想用一個一個去除來比較那個影響最大，從下表可以看出不考慮性別的時候 testing accuracy 變小許多，所以依照我目前的測試性別對結果影響最大。

	Training Accuracy	Testing Accuracy
全部考慮	85.261%	85.172%
不考慮婚姻情況	85.21%	85.258%
不考慮年紀	85.11%	85.270%
不考慮性別	85.52%	85.091%