

學號：B05902018 系級：資工三 姓名：張凱程

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

A: 以未實作 feature normalization 的情況下，generative model (0.84043 + 0.84668)表現比 logistic regression (0.79081 + 0.79533)好很多，其中可能原因是我是用 X_train 檔去做，而 feature 多數為 0 與 1 兩種，在做 logistic regression 會出一些問題。

2. 請說明你實作的 best model，其訓練方式和準確率為何？

A: Gradient Boosting Classifier, 其中估計函式數量為 500 個，準確率：0.87544 + 0.87592

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: logistic regression: 使用後可以比 generative model 高(0.85272 + 0.85135)。
generative model: 使用後沒有明顯的比較準確(0.84129 + 0.84594)。

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:

Lambda = 0.001: (0.85272 + 0.85135)

Lambda = 0.1: (0.85272 + 0.85135)

Lambda = 10: (0.85222 + 0.85147)

Lambda = 1000: (0.84436 + 0.84950)

本次實作 logistic regression 的 iteration 次數不多，但很快收斂。而這次看出加入 regularization 後發現：lambda 越大，正確率越低，可以看出正規化在此 model 並沒有帶來正面影響，可以推測出如果該 model 很快收斂，可能比較沒有過度訓練的情形，也就不必要加上 regularization。

5. 請討論你認為哪個 attribute 對結果影響最大？

A: 以 X_train 檔做討論，Prof-specialty 的結果影響最大，我在建立 model 時使用 feature_importances_ 的函式，可以知道該特徵的實用程度，而 Prof-specialty 與 capital-gain 均超過 0.1，多數為 0.01 至 0.00001 之間，有些則為 0 (無參考價值)。