

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

答：

在 **generative model** 中我所採用的模型為 Gaussian，並且採用算 prior probability 的方式實作，首先將 training data 依據 label 區分為兩個 class 後，計算其 mean 值。接著透過加權平均算出他們的 covariance matrix。接著就可以利用上述參數對照老師投影片中的公式計算出 y ($y = wx + b$)，然後將其套入 sigmoid 函數即可算出分類的機率，最後取 0.5 做為分類的依據。透過這個方式所得到的結果在 Kaggle 上顯示的正確率為 0.84165。

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

答：

在 **discriminative model** 中利用 logistic regression 方式，比照作業一的 gradient descend 方式從 function set 中找出一組最好的 function。這次作業我試過 $y=wx+b$ 以及 $y=ax^2+wx+b$ 兩種，由於這次沒有針對 feature 抽去，因此在二次方的 model 估計受到 noise 干擾造成 performance 表現不佳，其結果在 Kaggle 上所顯示最好的成績為 0.78194。最後選擇採用一次方的 model，其結果在 Kaggle 上之分數為 0.85307。

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：

在 $y=wx+b$ 的模型中，我針對不同的 feature 做了 normalization，其結果如下表。如果沒有做 feature normalization，其運算結果會產生 overflow。之後針對第二個欄位的值做 normalization，其在 Kaggle 上表現得分數為 0.83575，之後將前六個數值較大的欄位都做 normalization 之後，在 Kaggle 上的表現分數達到 0.85307。

準確率	Non-normalization	Normalized 'fmlwgt'	Normalized first six columns
$y = wx + b$	overflow	0.83575	0.85307

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：

我在不同的 model 上皆有試過 regularization，如下表所示。

準確率	$\lambda = 0$	$\lambda = 0.75$	$\lambda = 10$
$y = wx + b$	0.85307	0.85307	0.85307
$y = ax^2 + wx + b$	0.78194	0.69324	0.77727

5.請討論你認為哪個 **attribute** 對結果影響最大？

一開始我觀察到'fnlwgt'分佈的情況較為明顯，在針對其 normalization 後所得到的分數已達到 83%，接著針對前六個 attribute 一起做 normalization 後，分數更是趨近 85%。不過如果只拿前六個 attribute 下去 train 分數並沒有比較好，因此最後拿所有 attribute 並且針對前六個做 normalization 後下去 train，得到此次的分數。雖然在 public 沒有達到 strong baseline，不過在 private 有勉強達標。因此我認為前六個 attribute 對結果有最顯著的影響。