ML2017Fall - HW2 Report

學號：R05945012 系級：生醫電資碩二 姓名：張凱崴

※ 將 training set 的 10%隨機切分成 validation set。考慮到隨機性的問題，每組數據都做了三次並取平均。呈現如下：

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

Model	Generative model	Logistic regression
Training accuracy	0.843087	0.858750
Validation accuracy	0.835688	0.851556
Testing accuracy	0.842905	0.857360

Logistic regression 無論是 training, validation 還是 testing set，都比 generative model 有較高的準確度。因為 generative model 是建立在樣本是 Gaussian distribution 的假設上，而這不一定是正確的。因此 logistic regression 比較容易得到正確的預測。

2.請說明你實作的 best model，其訓練方式和準確率為何？

我採用 logistic regression model 作訓練，設定 batch size 與 epoch num 均為 1000，並使用 adam (learning rate=0.001) 作 gradient descent。首先我比較 1st, 2nd, 3rd order 三種模型，其中 2nd 有最好的 validation accuracy；接著對其做正規化，實驗結果如問題 4.所示，在 lambda=0.1 時有最佳的 validation accuracy。因此將其選為我的 best model。

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。

Generative model:

Model	w/ feature normalization	w/o feature normalization
Training accuracy	0.843087	0.842336
Validation accuracy	0.835688	0.832003
Testing accuracy	0.842905	0.842905

對於 generative model，標準化對於結果來說並沒有影響，因為 generative model 某種程度是套用公式解，標準化並不會改變其建立的模型的準確度。

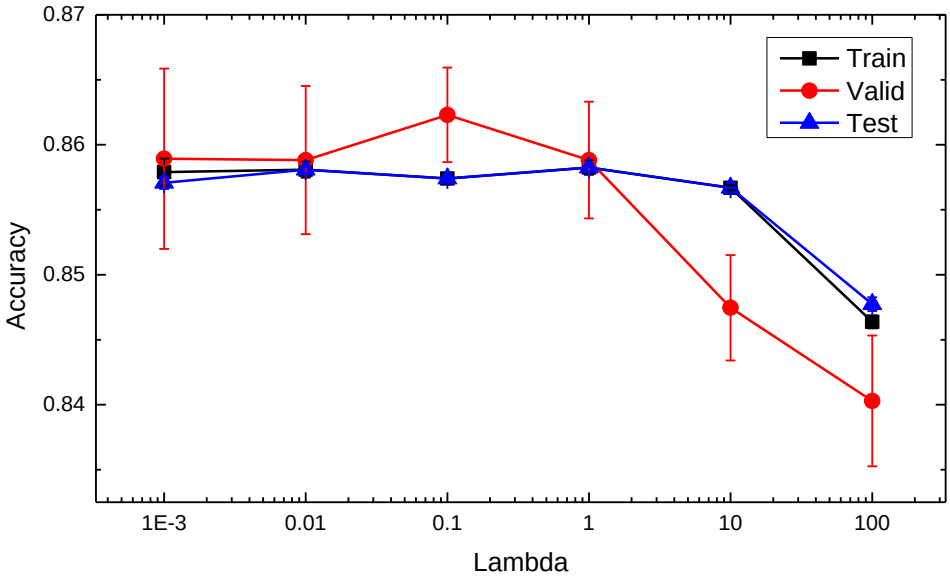
Logistic regression

Model	w/ feature normalization	w/o feature normalization
Training accuracy	0.858750	0.791139
Validation accuracy	0.851556	0.787572
Testing accuracy	0.857360	0.794750

沒有標準化的 logistic regression 結果非常差，原因有二。其一是因特徵的 scale 不同，做 gradient descent 時較不易走向 optimal。其二，在沒有標準化的情況下，sigmoid function 內的 exponential 很容易出現 overflow，因此產生偏差的結果。

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。

Model	λ=0	λ=0.01	λ=0.1	λ=1	λ=10
Training accuracy	0.858750	0.858090	0.857407	0.858249	0.856691
Validation accuracy	0.851556	0.858825	0.862300	0.858825	0.847461
Testing accuracy	0.857360	0.856766	0.856766	0.857400	0.855824



加上正規化後，對準確率的作圖如上。我們可以看到 training set 的準確率隨著 lambda 增加而下降，而 validation set 的準確率則隨 lambda 增加，先上升而後下降，因此我選擇 lambda=0.1 做為我的 best model。然而對於 testing set 來說卻並非如此，理論上 validation set 是用來估計 testing set 的分數，但以上圖來看，validation set 反而完全無法用來衡量 testing set 的分數，倒是 testing set 與 training set 有十分相似的趨勢，這是做訓練時完全意料之外的事。推測原因可能是我在切 validation set 時的比例太小，導致取出來的 validation set 容易有 bias，因此每次做出來的結果變動十分劇烈(標準差很大)，所以不能準確估計 testing set 的分數。

5.請討論你認為哪個 attribute 對結果影響最大？

分析訓練出來的 model 對於不同 feautre 的權重(取絕對值)並作圖如下。由大而小前五名分別為 capital gain (first order), age (first order), age (second order), never-married (first order), capital gain (second order)。由此分析結果，我認為 capital gain 和 age 這兩個 attribute 有前兩大的影響力。