

# Homework 2 Report - Credit Card Payment Prediction

學號：B05505004 系級：工海三 姓名：朱信霆

## 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。

這題我 logistic regression 實作的方法是先把資料 one hot encoding，接著考量到性別、婚姻與預測會不會繳信用卡費較為無關，因此把這兩個 feature 直接的捨棄掉。接著使用 stochastic gradient descent 的精神，把每次訓練一個 mini-batch，並且使用 adagrad 調整 learning rate。

generative model，因為如果使用 one hot encoding 會導致 covariance matrix 為 singular，無法取反矩陣。因此我僅使用投影片的推導，假設兩者為 gaussian distribution 並且按照 label 的比例 share covariance。

由於我在 logistic regression 有使用 one hot encoding，可是在 generative model 並沒有，因此在接下來的討論為了公平起見，都將所有的資料、feature 下去訓練，藉此來比較兩者的表現。

	Training score	Public score	Private score
Generative model	0.81245	0.81200	0.80600
Logistic regression	0.79505	0.79980	0.79560

可以看到在沒有 one hot encoding 的情況下 generative model 表現得比 logistic regression 好，不管是在 training 還是 testing set 上，我認為這是因為在 logistic regression 中，當我們使用 integer label 標註離散的資料時，事實上暗示的是數字越高越好，然而像是性別及婚姻並沒有所謂的優劣，因此在這裡就會出現問題。

然而像是 Generative model 他其實是算你在這個 class 的機率是多少，而機率又是根據每個 feature 的 distribution 而決定，假設平均數是  $\mu$ ，不管你的數值是  $\mu + x$ ，還是  $\mu - x$ ，你從這個 class 出來的機率都是相同的。因此 Generative model 在處理離散而沒有優劣的數據中反而會比 logistic regression 來的適合。

w/o sex & marriage	Training score	Public score	Private score
Generative model	0.81195	0.81200	0.80780
Logistic regression	0.7980	0.80020	0.79940

為了驗證我剛剛的說法，因此我把資料的性別和婚姻狀態都拿掉不考慮進去，如此產生的結果可以看到 Logistic regression 在三個 score 中都有微幅的上升，而 Generative model 則是掉了一點點的 score，如此可以驗證我剛剛的說法，從 Generative model 的分數下降可以推知其實這兩個 feature 還是有一點點的資訊量存在，可是拿掉這兩個 feature 反而使 Logistic regression 的分數上升，代表說這種 integer label 不僅不會幫助，反而還會誤導我們的 model，造成錯誤的預測。

## 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。

<b>Logistic regression</b>	<b>Training score</b>	<b>Public score</b>	<b>Private score</b>
w/ one-hot encoding	0.82215	0.81920	0.82000
w/o one-hot encoding	0.79505	0.79980	0.79560

在上方的實驗中我是用stochastic gradient descent用adagrad實作logistic regression，batch size設為100下去分別訓練有one-hot encoding以及沒有one-hot encoding的model來做比較。

若沒有加入one-hot encoding，則feature為23維，若加入則高達90維。我們可以看到有one hot coding的model準確率大幅的領先沒有加入的model，我認為這是因為這個資料集中的feature隱含的許多離散，同時各據意義的資料，這先各個類別彼此並沒有優劣的差別。而這也呼應我上一題討論說的，如果使用原本的integer label則會隱含著數字代表優劣的關係，進而影響到model的準確性。

### 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）。

我的實驗方始是先使用最原始的Logistic regression加上one hot encoding，接著再實驗如果我們不看任一feature，它的結果會是如何，如果score和原本比下降很多，那就代表這個feature對於model的判斷非常最重要。為了比免表格太大，如果有同一大類有不同欄的我只實驗最先和最後的欄位。

<b>Logistic regression</b>	<b>Training score</b>	<b>Public score</b>	<b>Private score</b>
original	0.80055	0.80380	0.80140
w/o LIMIT_BAL	0.80535	0.80400	0.80560
w/o SEX	0.8	0.80360	0.80120
w/o EDUCATION	0.8007	0.80360	0.80140
w/o MARRIAGE	0.8002	0.80260	0.80100
w/o AGE	0.80065	0.80360	0.80240
w/o PAY_0	0.79515	0.80020	0.79280
w/o PAY_6	0.8005	0.80040	0.80080
w/o BILL_AMT1	0.8019	0.80300	0.80320
w/o BILL_AMT6	0.80175	0.80240	0.80340
w/o PAY_AMT1	0.8036	0.80300	0.80480
w/o PAY_AMT6	0.80025	0.80320	0.80120

從上表我們可以看到如果沒有PAY\_0的話，整個預測的準確度下降最多，因此我們合理的推斷這一項是最重要的feature，再來還可以發現少掉LIMIT\_BAL甚至分數還上升，代表這一項其實對預測並沒有幫助，甚至還會使的結果變得更糟。

### 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。

Logistic regression	Training score	Public score	Private score
w/o feature normalization	0.80055	0.80380	0.80140
w/o Standardization	0.8202	0.81920	0.82160
w/ Minmax scaling	0.81265	0.81160	0.81160

這題的實驗我是用logistic regression，並且使用全部的資料做one hot encoding後下去搭配不同的feature normalization來做比較。從上面的表格可以看出不論是哪一種標準化的方法，都使的model的預測準確度上升，然後Standardization的分數上升幅度又大於Minmax scaling。

我認為這可以從這兩個標準化的特性下去分析，Standardization過後的資料事實上會有standard normal distribution，如果原始的資料沒有normal distribution的分布的話，就有可能有比較差的表現。而Minmax scaling則是將所有數值縮到0到1之間，而如此做的特性是它會使的outliers有較小的影響力，同時導致標準差變得比較小。

根據以上的特性，我認為導致Standardization分數較高的原因是因為在one hot encoding後這個dataset剩下比較多的是屬於金錢相關的資訊，這些資訊其實是符合normal distribution，再者如果這些資訊有較大的標準差，勢必會使的資料點在高維空間拉的越開，也比較容易區分的開來，這也是為什麼我認為Standardization會有較高準確率的原因

## 5 (1%)

- The Normal (or Gaussian) Distribution is a very common continuous probability distribution. Given the PDF of such distribution  
please show that such integral over  $(-\infty, \infty)$  is equal to 1.
- Ans:

First, I want to compute the value of  $\int_{-\infty}^{\infty} e^{-x^2} dx$

since

$$\begin{aligned} (\int_{-\infty}^{\infty} e^{-x^2} dx)^2 &= \int_{-\infty}^{\infty} e^{-x^2} dx \int_{-\infty}^{\infty} e^{-y^2} dy \\ &= \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\theta = 2\pi \int_0^{\infty} e^{-r^2} d(\frac{r^2}{2}) = \pi(e^{-0} - e^{-\infty}) = \pi \end{aligned}$$

Therefore,  $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$

set  $z = \frac{(x-\mu)}{\sqrt{2}\sigma}$ ,  $dz = \frac{1}{\sqrt{2}\sigma} dx$ ,  $dx = \sqrt{2}\sigma dz$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-z^2} \sqrt{2\sigma} dz = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-z^2} dz = 1$$

## 6 (1%)

- Given a three layers neural network, each layer labeled by its respective index variable. Derive the general expressions for the following partial derivatives of an error function E.

$$(a) \frac{\partial E}{\partial z_k} = \frac{\partial y_k}{\partial z_k} \cdot \frac{\partial E}{\partial y_k} *$$

$$(b) \frac{\partial E}{\partial z_j} = \frac{\partial y_j}{\partial z_j} \cdot \frac{\partial E}{\partial y_j} \quad \left( \frac{\partial E}{\partial y_j} = \frac{\partial z_k}{\partial y_j} \frac{\partial E}{\partial z_k} = \frac{\partial z_k}{\partial y_j} \cdot \frac{\partial y_k}{\partial z_k} \cdot \frac{\partial E}{\partial y_k} \right)$$
$$= \frac{\partial y_j}{\partial z_j} \cdot \frac{\partial z_k}{\partial y_j} \cdot \frac{\partial y_k}{\partial z_k} \cdot \frac{\partial E}{\partial y_k} *$$

$$(c) \frac{\partial E}{\partial w_{i,j}} = \frac{\partial z_i}{\partial w_{i,j}} \cdot \frac{\partial y_j}{\partial z_i} \cdot \frac{\partial z_k}{\partial y_j} \cdot \frac{\partial y_k}{\partial z_k} \cdot \frac{\partial E}{\partial y_k} *$$