Homework 2 Report - Credit Card Default Payment Prediction

學號：B06209027 系級：大氣二 姓名:李冠勳

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。
將LIMIT_BAL、AGE、BILL_AMT1、BILL_AMT2、BILL_AMT3、BILL_AMT4、BILL_AMT5、BILL_AMT6、PAY_AMT1、PAY_AMT2、PAY_AMT3、PAY_AMT4、PAY_AMT5、PAY_AMT6進行stardard normalization其餘項目進行one-hot encoding得到總共90維的資料。
其中generative model使用Gaussian Distribution，得到的結果如下表。
結果差上許多，因為使用one-hot encoding 時會導致feature的數值為0或1，這個作法會導致Gaussian Distrubution無法正常運作，但若是沒有使用one-hot encoding原始數據那種類似填表格的資料也會導致Gaussian Distrubution無法正常運作。在Logistic Regression方面，它沒有受到既定的model影響，導致在這個分析中可以得到較佳的成績。

model	trainging	public score	private score
Generative Model	0.66310	0.65400	0.64440
Logistic Regression	0.82170	0.82080	0.82100

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。
如第一題將上述的資料都進行stardard normalization，其餘一個有做one-hot encoding一個沒有，結果如下。
有做one-hot encoding的model比較準，因為那些項目中資料不同時，兩兩間的差異應該相同這樣在進行個參數weight計算時才不會有偏差，這也是為甚麼沒做one-hot encoding時即使iteration數十萬次參數依然浮動(找不到極值)。下表是我在各種找不到極值的情況下挑出最好的結果。

model	trainging	public score	private score
one-hot encoding	0.82170	0.82080	0.82100
without one-hot encoding	0.79010	0.78260	0.78260

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)。
資料分別以feature normalization 和 one-hot encoding處理後分別將各項分別刪除後查看結果

刪除項	training	public score	private score
limit_bal	0.82170	0.82180	0.82280
sex	0.82120	0.82080	0.82160
education	0.82090	0.82080	0.82040
marriage	0.82105	0.82040	0.82100
age	0.82120	0.82060	0.82060
pay_0	0.80580	0.80600	0.80040
pay_2	0.82080	0.82160	0.82080
pay_3	0.82110	0.82060	0.82100
pay_4	0.82085	0.81980	0.82040
pay_5	0.82140	0.81960	0.82000
pay_6	0.82075	0.81940	0.82060
BILL_AMT1	0.82085	0.82040	0.82160
BILL_AMT2	0.82090	0.82060	0.82160
BILL_AMT3	0.82090	0.82020	0.82200
BILL_AMT4	0.82095	0.82040	0.82180
BILL_AMT5	0.82070	0.82060	0.82160
BILL_AMT6	0.82075	0.82100	0.82180
PAY_AMT1	0.82055	0.82040	0.82220
PAY_AMT2	0.82055	0.81840	0.82200
PAY_AMT3	0.82115	0.81840	0.82160
PAY_AMT4	0.82070	0.82060	0.82160
PAY_AMT5	0.82115	0.82020	0.82160
PAY_AMT6	0.82095	0.82040	0.82180

觀察上表可以發現每一項的數值都查不多惟獨pay_0項與其他項差異最大，刪除它後只得到0.8左右的成績，因此推測此項的影響最大，其餘項目的影響差不多

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。
兩者都有使用one-hot encoding，其中一個有對參數標準化另一個沒有，結果如下表。
觀察參數Loss的數值，可以發現沒有參數標準化達到極值得速率比較慢，而且需要參數標準化項目的weight的數值都很小，這表示需要在極端的資料（極正或極負）才會對output產生影響，反觀做了feature normalization後較大的數值會被收縮到一個範圍內，因此該項目的數值對於output的影響效果會比較平均，得以呈現較佳的結果。

model	trainging	public score	private score
feature normalization	0.82170	0.82080	0.82100
without feature normalization	0.78015	0.77780	0.77980

5 (1%)

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

$$因為 x = \mu + \sigma n \quad dx = \sigma dn \quad n \in Z$$

$$=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(\sigma n)^2}{2\sigma^2}}\sigma dn$$

$$=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-n^2}{2}}dn$$

$$令I=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-n^2}{2}}dn=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-m^2}{2}}dm$$

$$I^2==\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-n^2}{2}}dn\frac{1}{\sqrt{2\pi}}e^{\frac{-m^2}{2}}dm$$

$$I^2=I^2=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{\frac{-(n^2+m^2)}{2}}dndm$$

$$令n=rcos\theta,m=rsin\theta$$

$$I^2=\frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty}e^{\frac{-r^2}{2}}rdrd\theta$$

$$I^2=e^{\frac{-r^2}{2}}rdr=\frac{1}{2}\int_{0}^{\infty}e^{\frac{-r^2}{2}}2rdr=\frac{1}{2}\int_{0}^{\infty}e^{\frac{-r^2}{2}}dr^2=\int_{0}^{\infty}e^{-r^2}dr^2=1$$

$$所以\Rightarrow I=1\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx=1,-\infty<x<\infty得證$$

6 (1%)

(a) $\frac{\partial E}{\partial z_k}$

$$\frac{\partial E}{\partial z_k}=\frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial z_k},\quad y_k=g(z_k)$$

$$因為g(x)是一種可微分函數，因此\frac{\partial E}{\partial z_k}=\frac{\partial E}{\partial y_k}g'(z_k)$$

(b) $\frac{\partial E}{\partial z_j}$

$$\frac{\partial E}{\partial z_j}=\frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial z_j}$$

分為前後項分別討論

$$前項\Rightarrow\frac{\partial E}{\partial y_j}=\frac{\partial E}{\partial z_k}\frac{\partial z_k}{\partial y_j}\quad 又z_k=\sum_j w_{jk}y_j因此\frac{\partial z_k}{\partial y_j}=\sum_j w_{jk}$$

$$\frac{\partial E}{\partial y_j} = \frac{\partial E}{\partial y_k} g'(z_k) \sum_j w_{jk}$$

$$\text{後項} \Rightarrow \frac{\partial y_j}{\partial z_j} = g'(z_j)$$

$$\text{因此} \frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_k} g'(z_k) (\sum_j w_{jk}) g'(z_j)$$

(c) $\frac{\partial E}{\partial w_{ij}}$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial z_j} \frac{\partial z_j}{\partial w_{ij}}$$

$$\text{前項與}(b)\text{相同，後項} = y_i = g(z_i)$$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_k} g'(z_k) (\sum_j w_{jk}) g'(z_j) g(z_i)$$