Homework 2 Report Problem Set

Professor Pei-Yuan Wu
EE5184 - Machine Learning
b05902109 資工三 柯上優

Problem 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。

Features	Training	testing public	testing private
logistic regression	0.821700	0.82000	0.82140
generative	0.821350	0.82060	0.82200

觀察：
此實驗的gender, education, martial status,and pay1~pay6都有做one-hat, 這些以外的feature都有做Normalize。可以觀察到logistic regression比起generative在training accuracy有更好表現，我認為是因為logistic regression能做gradient使的結果較能fit原本的traininig data。
然而，在tesing accuracy的部份，generative的結果都比較好。我認為是因為此次作業的資料dataset不夠大，因為generative model在少一點的資料上表現會discriminative model比較好。若資料量更多，可能有完全不同的結果。

Problem 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。

Features	Training	testing public	testing private
no one-hot	0.821500	0.81960	0.82100
one-hot	0.821700	0.82000	0.82140

觀察：
此實驗的pay1~pay6有做one-hat，gender, education, martial status,and pay1~pay6以外的feature都有做Normalize。可以觀察到沒做one-hot明顯有較低的正確率。我認為

one-hot能將非連續性的feature各自討論，舉例來說，age的feature有1和2，這或許是指男性與女性，在計算時不該把它當作1小於2這種數學關係，而one-hot能使得他們分開計算，在regression時當作某種加權特徵。

Problem 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)。

delete	sex	marrage	age
Acc	0.821600	0.821200	0.821300

delete	pay0	pay2	pay3	pay4	pay5	pay6
Acc	0.805250	0.820950	0.821450	0.821200	0.821500	0.820950

delete	LIMIT_BAL	AGE
Acc	0.821950	0.821550

delete BILL_AMT	1	2	3	4	5	6
Acc	0.821600	0.821650	0.821650	0.821600	0.821700	0.821500

delete PAY_AMT	1	2	3	4	5	6
Acc	0.821400	0.821700	0.821650	0.821650	0.821450	0.821450

觀察：
此實驗固定將gender, education, martial status,and pay1~pay6做one-hat，除此以外的模型都有做Normalize。將每一項各自刪除做logistic regression得到以上training accuracy。
觀察發現，最高的training accuracy是在刪除LIMIT_BAL項，最低的training accuracy是刪除pay0項。這是在training set所做的觀察，testing set的觀察可能完全不同。

Problem 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。

Features	Training	testing public	testing private
no normalize	0.779550	0.78160	0.78100
normalize	0.821700	0.82000	0.82140

觀察:
此實驗固定將gender, education, martial status,and pay1~pay6做one-hat。發現沒有做Normalize的結果準確率明顯較低，我認為是，沒有做normalize的feature可能數量及差距過大，像是AGE會在百位數內，PAY_AMT卻會有破萬的數值，而regression最好都將input放在接近的範圍內，會有比較好的結果。

Problem 5. (1%)

collabarator: b04902131 黃郁凱

First we proof

$$2\int_0^\infty e^{-t^2}dt = \sqrt{\pi}$$

pf>

$$\begin{aligned}
I^2 &= 4\int_0^\infty \int_0^\infty e^{-(x^2+y^2)}\,dy\,dx \\
&= 4\int_0^\infty \left(\int_0^\infty e^{-(x^2+y^2)}\,dy\right)\,dx \\
&= 4\int_0^\infty \left(\int_0^\infty e^{-x^2(1+s^2)}x\,ds\right)\,dx \\
&= 4\int_0^\infty \left(\int_0^\infty e^{-x^2(1+s^2)}x\,dx\right)\,ds \\
&= 4\int_0^\infty \left[\frac{1}{-2(1+s^2)}e^{-x^2(1+s^2)}\right]_{x=0}^{x=\infty}\,ds \\
&= 4\left(\frac{1}{2}\int_0^\infty \frac{ds}{1+s^2}\right) \\
&= 2\Big[\arctan s\Big]_0^\infty \\
&= \pi.
\end{aligned}$$

Back to the problem,

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

let $z = \frac{x-\mu}{\sqrt{2}\sigma}$

then

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(\frac{x-\mu}{\sqrt{2}\sigma})^2} dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-(\frac{x-\mu}{\sqrt{2}\sigma})^2} dz = \frac{2}{\sqrt{\pi}} \int_{0}^{\infty} e^{-z^2} dz = \frac{1}{\sqrt{\pi}} \sqrt{\pi} = 1$$

Problem 6. (1%)

(a.) $\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k} \frac{\partial g(z_k)}{\partial z_k}$

(b.) $\frac{\partial E}{\partial z_j} = (\sum_k \frac{\partial E}{\partial y_k} \frac{\partial g(z_k)}{\partial z_k} \frac{\partial w_{jk} y_j}{\partial y_j}) \frac{\partial g(z_j)}{\partial z_j}$

(c.) $\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_k} \frac{\partial g(z_k)}{\partial z_k} (\sum_j \frac{\partial w_{jk} y_j}{\partial y_j} \frac{\partial g(z_j)}{\partial z_j} (\sum_i \frac{w_{ij} y_i}{w_{ij}}))$