Homework 2 Report

學號：B06901031　系級：電機二　姓名:楊宗賢

1. (1%) 請說明你實作之 logistic regression 以及 generative model 於此 task 的表現，並試著討論造成此差異及可能原因。

使用 generative model 時直接根據 Gaussian distribution 的假設得出 w,b，在 training data 上即有 0.8114 左右的 accuracy，在 Kaggle 上可得到 0.812 的分數；使用 logistic regression 時會發生 cross entropy 下降、accuracy 卻下降的現象，最好的 accuracy 未必發生在最小的 cross entropy。一開始 logistic regression 的表現都未超越 generative model， 直到使用 mini-batch learning 後，才在 Kaggle 上得到更高的 0.8128。

generative model 終究是基於 Gaussian distribution 的 mean 和 covariance 而得出，已經對資料的分佈型態作了假設，沒有什麼可以改進的空間。如果資料實際的分佈型態不似於 Gaussian distribution 的話，則不作此假設的 logistic regression 很有機會表現得更好。

2. (1%) 請試著將 input feature 中的 gender, education, martial status 等改為 one-hot encoding 進行 training， 並比較其與原本的差異及其可能原因。

在 generative model 中做 one-hot encoding 會讓 covariance matrix 變得 too sparse，於是礙於 NumPy 有限的 precision，算不出正確的 Σ^{-1}(即 $\Sigma\Sigma^{-1}\neq I$)，也就無從比較。

在 logistic regression 中同時對 gender, education, martial status 做 one-hot encoding，得出的成效如下：

	training set accuracy	Public Score	Private Score
without one-hot encoding	0.80770	0.80840	0.80520
with one-hot encoding	0.80905	0.81000	0.80620

如果再加上對 PAY_0 ~ PAY_6 也做 one-hot encoding，則 accuracy 可以超過 0.82。

當 input feature 中存在著無順序、非量化的屬性編號值時，先改成 one-hot encoding 再 training 比較能得出更準的預測。另外有些 feature 雖然是有順序的量化值，卻相當的離散(PAY_0 ~ PAY_6)，也可以試一試 one-hot encoding。

3. (1%) 請試著討論哪些 input features 的影響較大、哪些 input features 的影響較小（方法不限）。

為了探討 23 個 feature 對 accuracy 的影響，我一次移除一個 feature，將 23 組實驗組的 accuracy 跟保留全部 feature 的對照組做比較，其餘參數不變。

在 Training Data 上的 accuracy (這些資料沒有送 Kaggle)			
對照組	LIMIT_BAL	SEX	EDUCATION
0.8077	0.8073	0.8081	0.80815
MARRIAGE	AGE	PAY_0	PAY_2
0.80755	0.80765	0.7938	0.8087
PAY_3	PAY_4	PAY_5	PAY_6
0.8078	0.8079	0.80765	0.80805
BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4
0.80755	0.80775	0.80785	0.808
BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2
0.80775	0.808	0.80795	0.80775
PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
0.8077	0.80785	0.8081	0.80785

假設拔除影響大的 feature 將使 accuracy 下降，拔除影響小或無關的 feature 將使 accuracy 不變或上升。則可明顯看出 PAY_0(一個月前的還款延遲狀況)有極為重大的相關性，反倒 SEX(性別)、EDUCATION(教育程度)、PAY_6(六個月前的還款延遲狀況)等 feature 的存在對 accuracy 無影響或有負面影響。如果要對銀行做出資料屬性的結論，可以說「不需要重視持卡人的性別或教育程度，而應重視其最近一個月的還款情況」。

4. (1%) 請實作特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。(有關 normalization 請參考：https://goo.gl/XBM3aE)

Wikipedia 提到了四種 normalization 的方法，在此實作其中兩種：

	training set accuracy	Public Score	Private Score
without normalization	0.77830	0.78160	0.78120
min-max normalization	0.80770	0.80840	0.80520
standardization	0.81130	0.81200	0.80540

沒有作特徵標準化時，根本連 local minimum 都不容易到達，作了特徵標準化後，便快速地達到了好的 accuracy，而其中 standardization(減去 mean 再除以 standard variance)又比 min-max normalization(減去 min 再除以(max-min))表現略佳。

5. (1%)

5\. for $\int e^{-x^2}dx$, known as Gaussian integral,

its antiderivative cannot be expressed by elementary functions,

but $\int_{-\infty}^{\infty} e^{-x^2}dx$ is calculable :

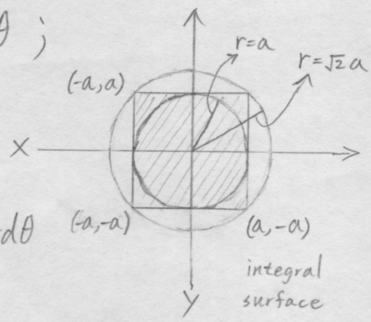
$$\left(\int_{-\infty}^{\infty} e^{-x^2}dx\right)^2 = \lim_{a\to\infty}\left(\int_{-a}^{a} e^{-x^2}dx\right)\left(\int_{-a}^{a} e^{-y^2}dy\right)$$

$$= \lim_{a\to\infty}\int_{-a}^{a}\int_{-a}^{a} e^{-(x^2+y^2)}dxdy \quad ,$$

calculate this double integral by converting Cartesian coordinates to polar coordinates :

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta \end{cases} \Rightarrow \text{Jacobian matrix } J(r,\theta) = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$$

$$\Rightarrow dxdy = |J(r,\theta)|\,drd\theta = r\,drd\theta \ ;$$



by the squeeze theorem,

$$\int_0^{2\pi}\int_0^{a} re^{-r^2}drd\theta \le \int_{-a}^{a}\int_{-a}^{a} e^{-(x^2+y^2)}dxdy \le \int_0^{2\pi}\int_0^{\sqrt{2}a} re^{-r^2}drd\theta$$

$$\Rightarrow (1-e^{-a^2})\pi \le \left(\int_{-a}^{a} e^{-x^2}dx\right)^2 \le (1-e^{-2a^2})\pi \ , \quad \text{apply } a\to\infty$$

$$\Rightarrow \pi \le \left(\int_{-\infty}^{\infty} e^{-x^2}dx\right)^2 \le \pi \Rightarrow \int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi} \ ;$$

$$\int_{-\infty}^{\infty} f(x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2}dx = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-t^2}dt \ \begin{pmatrix} t=(x-\mu)/\sqrt{2}\sigma \\ dx=\sqrt{2}\sigma\, dt \end{pmatrix} = 1 \ ※$$

reference : Wikipedia, textbook of Probability and Statistics

6. (1%)

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

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

(c) $$\frac{\partial E}{\partial W_{ij}} = \frac{\partial E}{\partial z_j}\frac{\partial z_j}{\partial W_{ij}} = W_{jk}\, y_i\frac{\partial E}{\partial y_k}\frac{\partial g(z_k)}{\partial z_k}\frac{\partial g(z_j)}{\partial z_j}$$