

Homework 2 Report - PM2.5 Prediction

學號： B05602022 系級：工海三 姓名: 盧庭偉

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。

	Public Score	Private Score
logistic regression	0.81760	0.81980
generative model	0.81160	0.80860

備註: 1. 使用 20000 筆 training 資料

2. 對連續型參數做 min-max normalization

3. Generative model 使用 Gaussian distribution 做 model.)

logistic regression 由於是沒有做假設的狀況下，用 regression 自己去找規律，一般來說在資料量足夠的情況下，結果應該會比有做分布狀況假設(Gaussian distribution)的 generative model 好一些。為了確證我們的假設，我們將 training 資料量減半做測試。

	Public Score	Private Score
logistic regression	0.79720	0.79400
generative model	0.80780	0.80908

備註: 1. 使用 10000 筆 training 資料

2. 對連續型參數做 min-max scaling

3. Generative model 使用 Gaussian distribution 做 model.

可以發現比起 generative model，logistic regression 的結果明顯變差許多，符合我們的預期。

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。

	Public Score	Private Score
使用原始資料	0.81760	0.81980
使用 one-hot encoding	0.82040	0.82140

備註: 1. 對連續型參數做 min-max scaling

可以看到改用 one-hot encoding 後結果有變好。我認為是因為，當用同一個參數不同的離散值當作 input 時，regression 很難真的 train 到剛好能表達不同值的意義。以 History of past payment(X6 ~ X11) 為例，1 ~ 9 代表 delay 1 ~ 9 個月；-1 代表準時；9 代表 delay 9 個月以上，很明顯這三種狀況下參數的意義有很大的不同，但因為這些參數彼此的間距是一樣的(除了-1 但也沒差很多)，做 regression 時很難做出一個能讓這三種狀況明顯不同的切割。

而使用 one-hot encoding，由於每種離散的狀況都有自己的參數，因此可以分別 train 出不同情形對結果的影響，為處理離散資料較好的做法。

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）。

我們將 training set 去掉各項 feature 做，得到以下結果:

扣除項	Public Score	Private Score	平均
LIMIT_BAL	0.81760	0.81640	
SEX	0.81680	0.81320	
EDUCATION			
MARRIAGE			
AGE			
PAY_0			
PAY_2			
PAY_3			
PAY_4			
PAY_5			
PAY_6			
BILL_AMT1			
BILL_AMT2			
BILL_AMT3			
BILL_AMT4			
BILL_AMT5			
BILL_AMT6			
PAY_AMT1			
PAY_AMT2			
PAY_AMT3			
PAY_AMT4			
PAY_AMT5			
PAY_AMT6			

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。

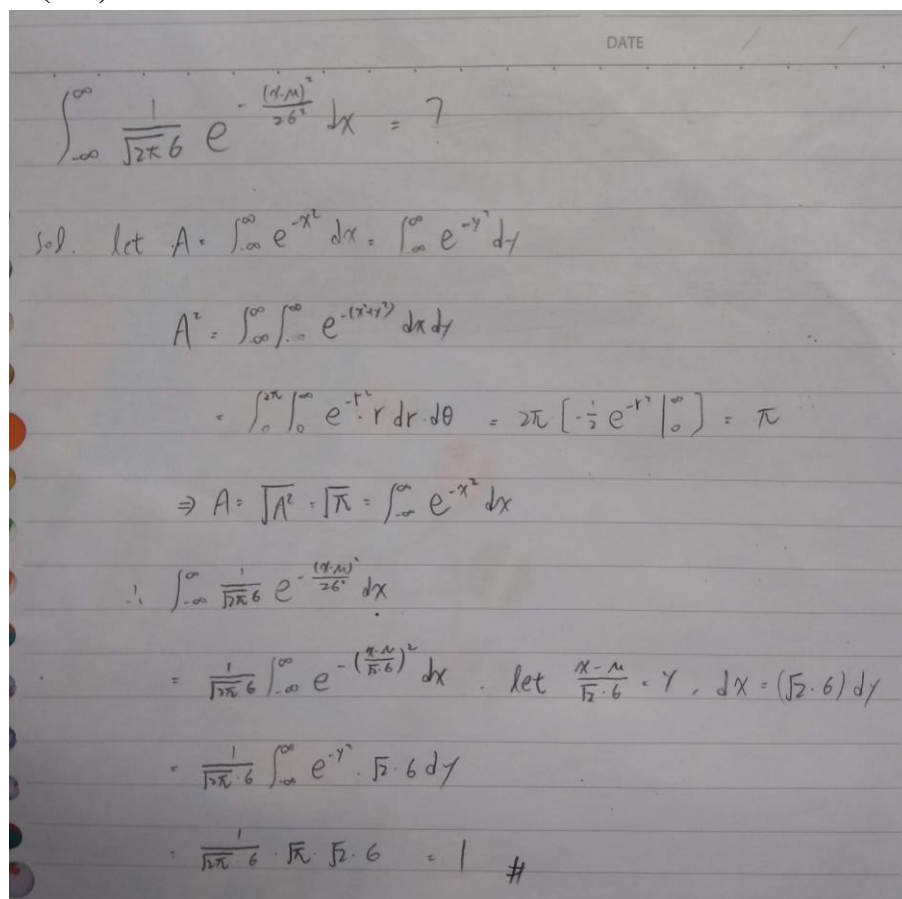
	Public Score	Private Score
No Scaling	0.78680	0.78460
Min-Max Scaling	0.81760	0.81980
Standardization (Z-score)	0.81200	0.81760

備註: 1. 對連續型參數做 scaling

理論上有做 scaling 的資料收斂速度會比較快，且由於本次 training 資料彼此間性質及數值範圍差距較大，沒做 scaling 幾乎不太會收斂。而 Min-Max 與 Standardization 的結果則沒有顯著的差異。

(Reference: https://sebastianraschka.com/Articles/2014_about_feature_scaling.html#about-standardization)

5. (1%)



DATE / /

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

Sol. let $A = \int_{-\infty}^{\infty} e^{-x^2} dx = \int_{-\infty}^{\infty} e^{-y^2} dy$

$$A^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy$$

$$= \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\theta = 2\pi \left[-\frac{1}{2} e^{-r^2} \right]_0^{\infty} = \pi$$

$$\Rightarrow A = \sqrt{A^2} = \sqrt{\pi} = \int_{-\infty}^{\infty} e^{-x^2} dx$$

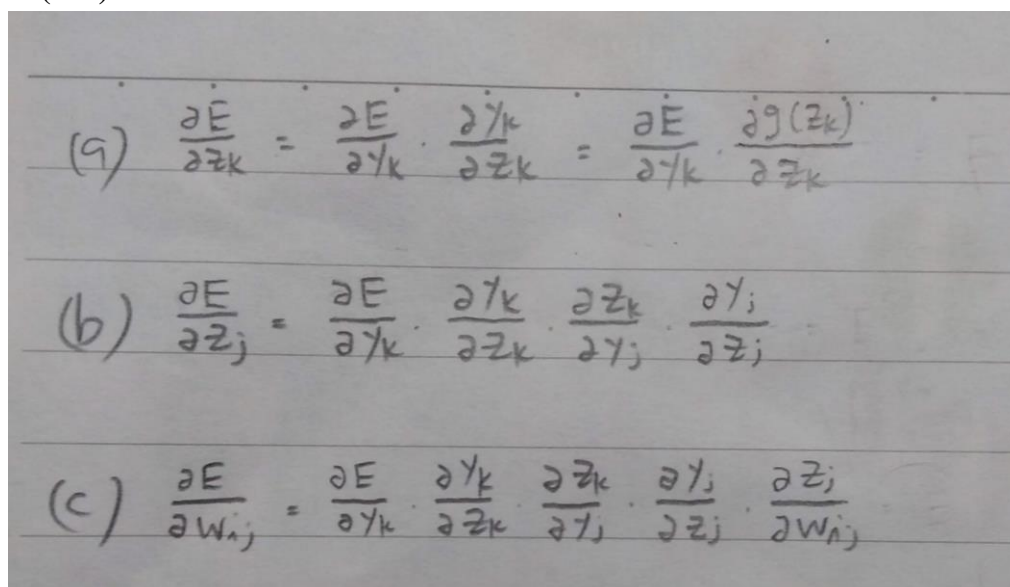
$$\therefore \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{\infty} e^{-\left(\frac{x-\mu}{\sigma}\right)^2} dx \quad \text{let } \frac{x-\mu}{\sigma} = y, dx = (\sigma \cdot 1) dy$$

$$= \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{\infty} e^{-y^2} \cdot \sigma dy$$

$$= \frac{1}{\sqrt{2\pi} \sigma} \cdot \pi \cdot \sigma = 1 \quad \#$$

6. (1%)



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

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

(c) $\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial z_k} \cdot \frac{\partial z_k}{\partial y_j} \cdot \frac{\partial y_j}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{ij}}$