Homework Report - Prediction

系級：資工三
學號：B05902022
姓名: 張雅信

Problem 1. (1%) 請簡單描述你實作之 logistic regression 以及 generative model 於此 task 的表現,試著討論可能原因。

實作結果：

Model	Data	Public Score	Private Score
logistic regression	All features	0.8130	0.8092
logistic regression	Selected features	0.8200	0.8190
generative	All features	0.8026	0.8002
generative	Selected features	0.7982	0.7946

理論上來說，兩者結果應該相近，logistic regression 會稍好一些，而 data 不足的時候則相反。就這次作業來說，對 data 做完 one-hot encoding 與 normalization 後，結果也與理論相符。

對於 Selected features, 篩掉較不相關的 feature 對 logistic regression 有幫助（減少 noise），但 generative model 就沒有進步。

Problem 2. (1%) 請試著將 input feature 中的 gender, education, martial status 等改 one-hot encoding 進行 training process,比較其模型準確率及其可能影響原因。

實作結果：

Data	Public Score	Private Score
Span nothing	0.8122	0.8080
Span sex & marriage	0.8130	0.8092
Span sex, marriage, pay	0.8204	0.8222

由於這些 feature 裡的數字沒有數值意義，對其做 regression 並不合理。將他們改成 one-hot encoding 相當於使他們變成 k 個不同的binary class（ k 為 one-hot encoding 前有出現的數值種類），一方面較為合理（類似於feature transformation），在結果上也表現的比較好。

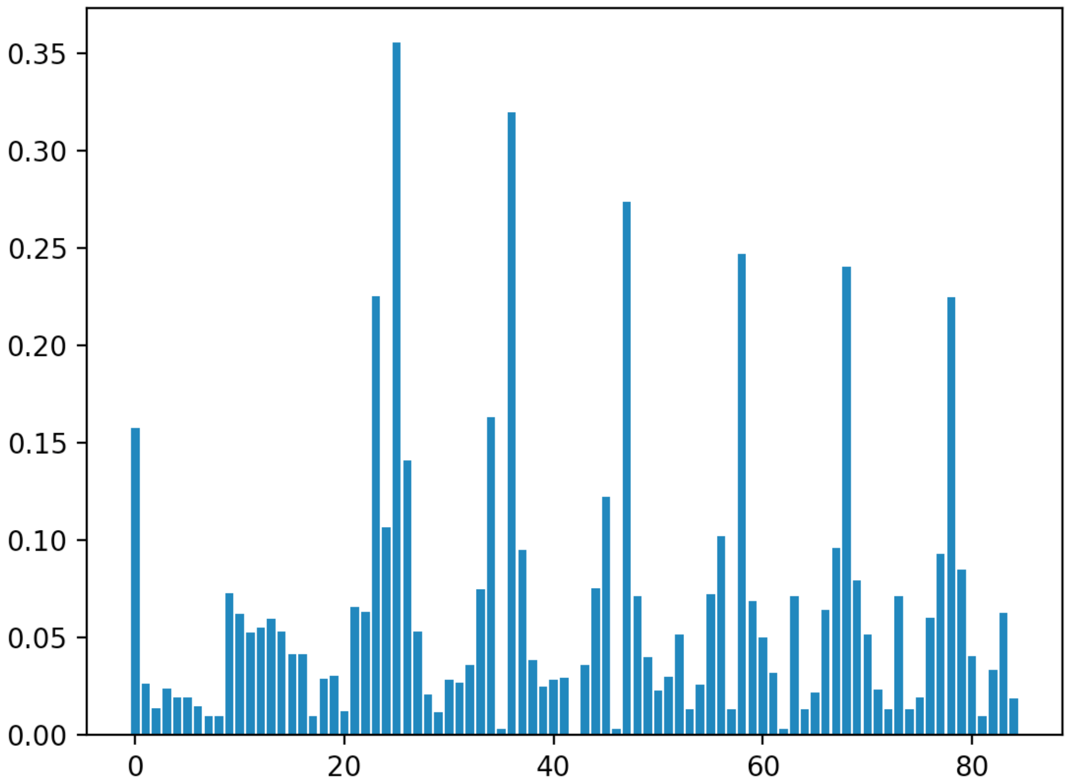
另外，參照下題的圖，也比較容易選擇我們要的feature。

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

(1) 對每個欄位做 logistic regression，loss 相近，無法分辨。

(2) 觀察 training data 與 label 之間的相關係數：

（某些欄位已經做完 one-hot encoding）



相關係數 > 0.15 的欄位有：

LIMIT_BAL, PAY_0 = 0, PAY_0 = 2, PAY_2 = -1
PAY_2 = 2, PAY_3 = 2, PAY_4 = 2, PAY_5 = 2
PAY_6 = 2

使用上述幾個欄位做 logistic regression 有表現較好，

其中最好的事只挑 PAY_x = 2的幾個欄位，

做出 Public score 0.8204。

Problem 4. (1%) 請實作特徵標準化 (feature normalization),討論其對於你的模型準確率的影響。

在做標準化之前，

(1) 參數初始化需做特別處理(使 $w^T x$ 接近 0~1)，

不然 loss 會太大，

無法用固定的 learning rate 做 logistic regression。

(2) 結果很差，比預測全 0 還不如。

一樣可能與固定的 learning rate 有關，

沒有針對每個維度調整參數改動幅度。

做完標準化，

用固定的 learning rate也很輕鬆的就達到超過 0.8 的準確率。

Problem 5. (1%)The Normal (or Gaussian) Distribution is a very common continuous probability distribution. Given the PDF of such distribution

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

please show that such integral over (−∞,∞) is equal to 1.

Ans:

$$let \quad a = \frac{1}{\sqrt{2\pi}\sigma}, \quad y = x - \mu, \quad z = \frac{y}{\sqrt{2\sigma^2}}$$

$$\int_{-\infty}^{\infty} f(x)dx = a\int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma^2}}dy = a\sqrt{2\sigma^2}\int_{-\infty}^{\infty} e^{-z^2}dz$$

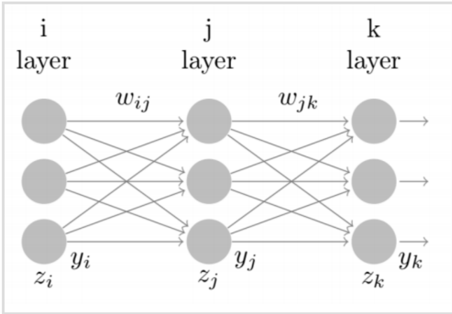
by Gaussian integral identity,

$$\int_{-\infty}^{\infty} e^{-z^2}dz = \sqrt{\pi}$$

Then we have

$$a\sqrt{2\sigma^2}\int_{-\infty}^{\infty} e^{-z^2}dz = a\sqrt{2\pi\sigma^2} = 1$$

Problem 6. (1%) Given a three layers neural network, each layer labeled by its respective index variable. I.e. the letter of the index indicates which layer the symbol corresponds to.



For convenience, we may consider only one training example and ignore the bias term. Forward propagation of the input z_i is done as follows. Where $g(z)$ is some differentiable function (e.g. the logistic function).

$$\begin{aligned} y_i &= g(z_i) \\ z_j &= \sum_i w_{ij} y_i \\ y_j &= g(z_j) \\ z_k &= \sum_j w_{jk} y_j \\ y_k &= g(z_k) \end{aligned}$$

Derive the general expressions for the following partial derivatives of an error function E in the feed-forward neural network depicted.

(a) $\dfrac{\partial E}{\partial z_k} = \dfrac{\partial y_k}{\partial z_k}\dfrac{\partial E}{\partial y_k} = g'(z_k)\dfrac{\partial E}{\partial y_k}$

(b) $\dfrac{\partial E}{\partial z_j} = \sum_k (\dfrac{\partial y_k}{\partial z_j}\dfrac{\partial E}{\partial y_k}) = \sum_k (\dfrac{\partial z_k}{\partial z_j}\dfrac{\partial y_k}{\partial z_k}\dfrac{\partial E}{\partial y_k})$

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

(c) $\dfrac{\partial E}{\partial w_{ij}} = \dfrac{\partial z_j}{\partial w_{ij}}\dfrac{\partial E}{\partial z_j} = y_i \sum_k (g'(z_j) w_{jk} g'(z_k)\dfrac{\partial E}{\partial y_k})$