Homework 2 Report Problem Set

Professor Pei-Yuan Wu
EE5184 - Machine Learning

B05902074　資工三 魏佑珊

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。

	Kaggle public	Kaggle private
Logistic	0.82040	0.82160
Generative	0.82120	0.82120

我覺得generative表現不如logistic的原因是generative model本身就比logistic model多了一個假設，更重視「資料是如何分布的」。所以generative的精神是想要找出一種最佳的分布型態去fit資料。但logistic更像是在gradient descent的過程中想辦法把資料「分開」，不那麼注重資料本身的分布型態。因此，generative假設性較強，也就容易和資料之間產生比較大的偏差。

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。

(以下，將gender, education, marriage, pay 做one hot)

	Kaggle public	Kaggle private
Logistic Naïve	0.52880	0.52459
Logistic + one hot	0.78170	0.78060
Generative Naïve	0.81200	0.80600
Generative + one hot	0.82000	0.82160

由上發現，做了one hot之後，不論是用generative或是logistic model，準確率都會提升。我覺得是因為one hot可以更準確地描述特定的資料，像是婚姻狀態、教育程度等等。因為在這些欄位中，數字只是用來分類，而非真正去描述一個值；像是在婚姻狀態中可能有1和2，而1和2代表的並不是真正準確的數值，而只是用來區分不同的婚姻狀態。因此，使用one hot可以透過增加feature，把婚姻狀態單純地用0/1描述。

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）。

DROP WHICH FEATURE	AVERAGE VALIDATION ACCURACY
Drop LIMIT_BAL	0.82407
Drop SEX	0.8236
Drop EDUCATION	0.8236
Drop MARRIAGE	0.8241
Drop AGE	0.82375
Drop PAY_0	0.8059
Drop PAY_2	0.8234
Drop PAY_3	0.8249
Drop PAY_4	0.82175
Drop PAY_5	0.82305
Drop PAY_6	0.82305
Drop BILL_AMT1	0.82325
Drop BILL_AMT2	0.82355
Drop BILL_AMT3	0.82305
Drop BILL_AMT4	0.8231
Drop BILL_AMT5	0.82325
Drop BILL_AMT6	0.82325
Drop PAY_AMT1	0.82358
Drop PAY_AMT2	0.82355
Drop PAY_AMT3	0.8231
Drop PAY_AMT4	0.8232
Drop PAY_AMT5	0.82327
Drop PAY_AMT6	0.82325

我分別把各種feature去掉之後，進行training process 2 次，每次都使用training data中約3000筆資料做validation，計算validation data的正確率，取2次平均。

由上表觀察來看，正確率最低的是Drop PAY_0 和 PAY_4，因此應該是PAY的feature對model正確率影響最大。

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。

	Kaggle public	Kaggle private
Logistic Naïve	0.52880	0.52459
Logistic + feature scaling	0.78060	0.78120
Generative Naïve	0.81200	0.80600
Generative + feature	0.81200	0.80600

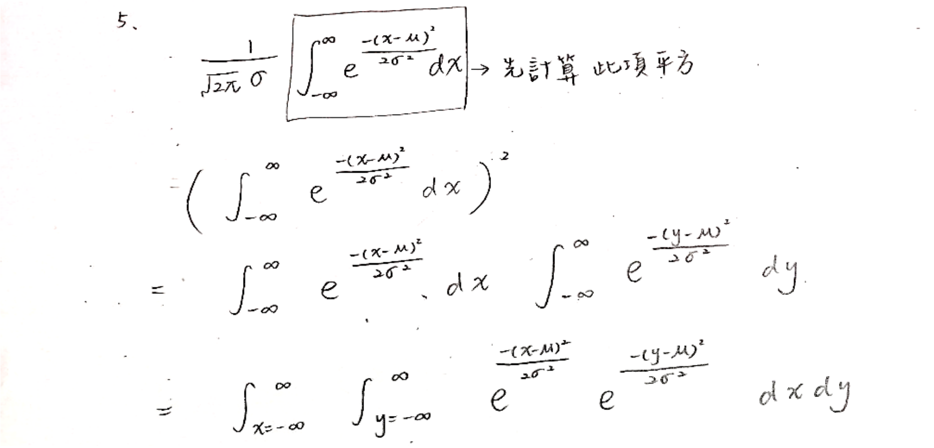
scaling		

我發覺feature normalization對logistic model的影響較大，而對generative model幾乎沒甚麼影響。我認為主要是因為feature normalization是縮小各個feature的範圍差距用以改進gradient descent，但generative model並沒有用到gradient descent的地方，反之，對logistic model就有很大的提升效果。當然另一個好處是做過normalize之後，取exp比較不容易overflow。

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

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

please show that such integral over $(-\infty, \infty)$ is equal to 1.



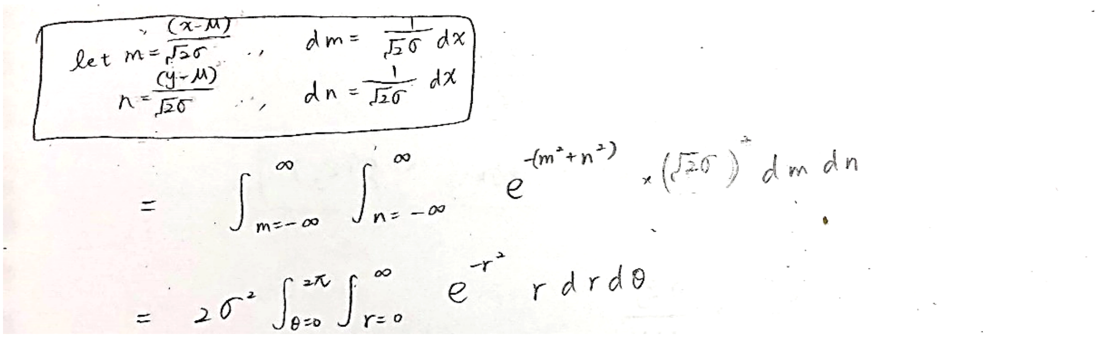
5、

$$\frac{1}{\sqrt{2\pi}\,\sigma}\boxed{\int_{-\infty}^{\infty} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx} \rightarrow \text{先計算 此項平方}$$

$$\left(\int_{-\infty}^{\infty} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx\right)^2$$

$$= \int_{-\infty}^{\infty} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx \int_{-\infty}^{\infty} e^{\frac{-(y-\mu)^2}{2\sigma^2}}\,dy$$

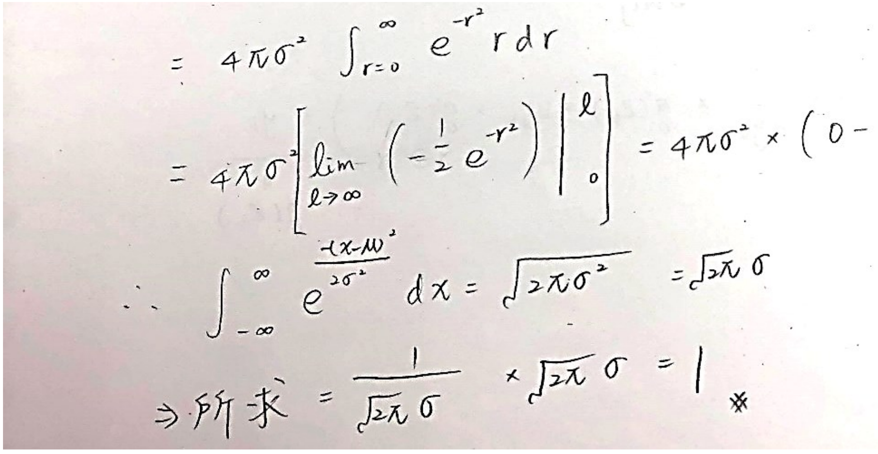
$$= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\, e^{\frac{-(y-\mu)^2}{2\sigma^2}}\,dx\,dy$$



$$\boxed{\begin{array}{ll} \text{let } m = \frac{(x-\mu)}{\sqrt{2}\sigma}, & dm = \frac{1}{\sqrt{2}\sigma}\,dx \\ n = \frac{(y-\mu)}{\sqrt{2}\sigma}, & dn = \frac{1}{\sqrt{2}\sigma}\,dx \end{array}}$$

$$= \int_{m=-\infty}^{\infty} \int_{n=-\infty}^{\infty} e^{-(m^2+n^2)} \times (\sqrt{2}\sigma)^2\,dm\,dn$$

$$= 2\sigma^2 \int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} e^{-r^2}\, r\,dr\,d\theta$$



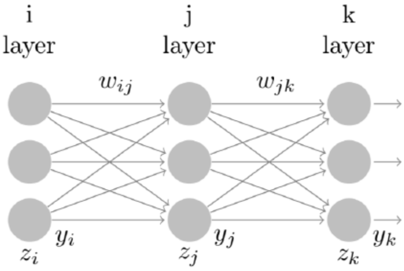
$$= 4\pi\sigma^2 \int_{r=0}^{\infty} e^{-r^2}\, r\, dr$$

$$= 4\pi\sigma^2 \left[\lim_{l\to\infty} \left(-\frac{1}{2} e^{-r^2}\right) \Big|_{0}^{l} \right] = 4\pi\sigma^2 \times \left(0 - \left(-\frac{1}{2}\right)\right) = 2\pi\sigma^2$$

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

$\Rightarrow$ 所求 $= \dfrac{1}{\sqrt{2\pi}\,\sigma} \times \sqrt{2\pi}\,\sigma = 1$ ※

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.



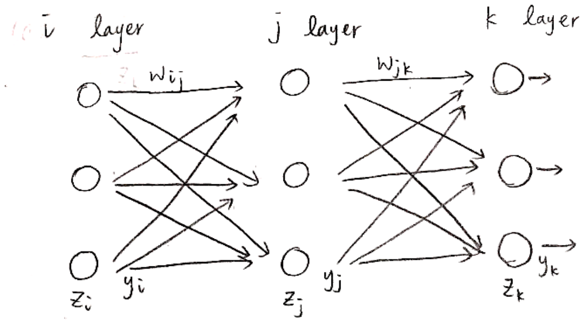
For convenience, we may consider only one training example and ignore the bias term.Forward propagation of the input z_i is done as follows. Where $g(z)$ is some differentiable function (e.g. the logistic function).

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

Derive the general expressions for the following partial derivatives of an error function E, also sime differentiable function, in the feed-forward neural network depicted. In other words, you should derive these partial derivatives into ”computable derivative” (e.g. $\frac{\partial E}{\partial y_k}$ or $\frac{\partial z_k}{\partial w_{jk}}$).

$$(a)\frac{\partial E}{\partial z_k} \quad (b)\frac{\partial E}{\partial z_j} \quad (c)\frac{\partial E}{\partial w_{ij}}$$

6.



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

(b) $$\frac{\partial E}{\partial z_j} = \sum_k \boxed{\frac{\partial E}{\partial z_k}} \times \boxed{\frac{\partial z_k}{\partial y_j}} \times \boxed{\frac{\partial y_j}{\partial z_j}}$$

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

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

$$= \left(\Sigma \frac{\partial E}{\partial y_k} \times g'(z_k) \times w_{jk} \times g'(z_j) \right) \times \underset{\overset{\|}{g(z_i)}}{\underline{y_i}}$$