學號：R06522709　系級：機械碩三　姓名：鄭呈毅

1. (0.5%) 請比較你實作的 **generative model**、**logistic regression** 的準確率，何者較佳?

	Public score	Private score
Generative model	0.84344	0.84399
Logistic regression	0.85540	0.84989

根據助教手把手 code，Logistic regression 的準確率較佳.

2. (0.5%) 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

	Before normalization	After normalization
Generative model	0.84416	0.84416
Logistic regression	0.78846	0.85055

Normalization 的部分同助教手把手 code，針對型態為 continuous 的所有資料。從我做出來的結果中可以發現 Generative model 的準確率不會受到 normalization 的影響；而 logistic regression 的變化除了反映在準確率之外，在訓練過程當中的 training loss 也會忽大忽小，這應該就是因為特徵標準化之前，continuous 的資料與 one hot encoding 的資料在 gradient 上相差較大，但是卻共用同一 learning rate，導致模型不易收斂，我試過將 epoch 調高，得到的結果依然與 simple baseline 相去甚遠，這說明了特徵標準化對於 logistic regression 是否能找到最佳解有很大的影響力。

3. (1%) 請說明你實作的 **best model**，其訓練方式和準確率為何?

我一開始是用 Keras 搭一個 MLP，一樣先針對 continuous 的資料做特徵標準化。Hidden layer 有三層 fully connected layer，在第一個 hidden layer 會 dropout 50%的 nodes，損失函數使用 binary crossentropy。這樣訓練下來可以拿勉強通過 public strong baseline，但是還離 private 有一段距離。最後打聽到 sklearn 的 Gradient Boosting Classifier 好像很好用，一用之下不得了，兩三行 code 就帶我飛過 private strongbaseline，然後標準化對這個 classifier 好像也沒有影響。另外我在訓練的時候發現 fnlwgt 這個特徵似乎對準確率沒甚麼幫助，在兩個 model 當中我都嘗試刪掉這個特徵，訓練出來的結果反而還變好一點點；後來上網查這個特徵是某種用於該人口普查的比重，可以合理推測它跟年收入沒有關係，因此與其把它放在資料裡混淆視聽不如刪掉它。

4. (3%) Refer to math problem
https://hackmd.io/0fDimqO7RaSCPpD_minSGQ?both

1. Prior Probabilities $P(C_k) = \pi_k$

General class-conditional densities $P(x|C_k)$ $k = 1, \ldots, K$.

Ans: The probablity of one data point is

$$P(x,t) = P(x|t)\,P(t) = \prod_{k=1}^{K} \left(P(x|C_k)\pi_k\right)^{t_k}$$

The probablity of entire data set

$$L(\theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left(P(x_n|C_k)\pi_k\right)^{t_{n,k}}$$

Take log on both side ($L(\theta)$ remain the same)

$$L(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} t_{n,k}\left[\log P(x_n|C_k) + \log \pi_k\right]$$

Subject $L(\theta)$ to the constraint that $\sum_{k=1}^{K} \pi_k = 1$.

$$L(\pi,\lambda) = \sum_{i=1}^{N} \sum_{k=1}^{K} t_{n,k}\left[\log P(x_n|C_k) + \log \pi_k\right] + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

Take derivative with respect to π_k and set it to 0

$$\frac{\partial L(\pi,\lambda)}{\partial \pi_k} = \frac{1}{\pi_k} \sum_{i=1}^{N} t_{n,k} + \lambda = 0 \Rightarrow \pi_k = -\frac{1}{\lambda} \sum_{i=1}^{N} t_{n,k} = -\frac{N_k}{\lambda}$$

N_k is number of data labeled with k

$$\frac{\partial L(\pi,\lambda)}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 \Rightarrow \sum_{k=1}^{K} \pi_k = 1$$

$$\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} -\frac{N_k}{\lambda} = 1 \Rightarrow \lambda = -N.$$

$$\pi_k = -\frac{N_k}{\lambda} = \frac{N_k}{N}$$

2. Show $\frac{\partial}{\partial \sigma_{ij}} \log|\Sigma| = e_j \Sigma^{-1} e_i^T = \Sigma^{-1}_{ji}$ is equal to

show $\frac{\partial}{\partial \Sigma} \log|\Sigma| = \Sigma^{-T}$.

known $\frac{\partial}{\partial \Sigma} \log|\Sigma| = \frac{1}{|\Sigma|} \frac{\partial}{\partial \Sigma} |\Sigma|$

known $\Sigma^{-1} = \frac{1}{|\Sigma|} \tilde{\Sigma}$, where $\tilde{\Sigma}$ is matrix of cofactors

known $|\Sigma| = \sum_j (-1)^{i+j} \sigma_{ij} M_{ij}$, where i is arbitrary.

$$\Rightarrow \frac{\partial}{\partial \sigma_{ij}} |\Sigma| = (-1)^{i+j} M_{ij}.$$

$$\Rightarrow \frac{\partial}{\partial \Sigma} |\Sigma| = \tilde{\Sigma}^T.$$

$$\frac{\partial}{\partial \Sigma} \log|\Sigma| = \frac{1}{|\Sigma|} \frac{\partial}{\partial \Sigma} |\Sigma| = \frac{1}{|\Sigma|} \tilde{\Sigma}^T = \Sigma^{-T}.$$

$$\Rightarrow \frac{\partial}{\partial \sigma_{ij}} \log|\Sigma| = e_j \Sigma^{-1} e_i^T.$$

\#

3. Let $X^{(1)}, X^{(2)}, \dots X^{(N)}$ be vectors and each of them represents a data point with P variables.

The product of individual density is:

$$\prod_{i=1}^{N} N(\mu_k, \Sigma)$$

Taking the logarithm gives the log-likelihood function:

$$L(\mu, \Sigma | x^{(i)}) = \log \prod_{i=1}^{N} \frac{1}{(2\pi)^{P/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)\right)$$

$$= \sum_{i=1}^{N} \left(-\frac{P}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)\right)$$

$$= -\frac{NP}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)$$

derive μ_k.

Take derivative with respect to μ and equate to 0.

$$\frac{\partial}{\partial \mu} L(\mu, \Sigma | x^{(i)}) = \sum_{i=1}^{N} \Sigma^{-1} (\mu - x^{(i)}) = 0$$

since Σ is positive definite.

$$0 = N\mu - \sum_{i=1}^{N} x^{(i)}$$

$$\Rightarrow \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} = \bar{X}$$

Derive Σ

$$\begin{cases} tr[ACB] = tr[CAB] = tr[BCA] \\ X^t A X = tr[X^T A X] = tr[X^t X A] \\ \frac{\partial}{\partial A} tr[AB] = B^T \\ \frac{\partial}{\partial A} \log|A| = A^{-T} \end{cases}$$

These properties allow us to calculate

$$\frac{\partial}{\partial A} X^t A X = \frac{\partial}{\partial A} tr[X^T X A] = [X X^t]^T = X^{TT} X^T = X^T .$$

compute the derivative with regard to Σ^{-1}

$$L(\mu, \Sigma | x^{(i)}) = C - \frac{N}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)$$

$$= C + \frac{N}{2} \log|\Sigma^{-1}| - \frac{1}{2} \sum_{i=1}^{N} tr[(x^{(i)} - \mu)(x^{(i)} - \mu)^T \Sigma^{-1}].$$

$$\Rightarrow \frac{\partial}{\partial \Sigma^{-1}} L(\mu, \Sigma | x^{(i)}) = \frac{N}{2} \Sigma - \frac{1}{2} \sum_{i=1} (x^{(i)} - \mu)(x^{(i)} - \mu)^T, \quad \text{since } \Sigma^T = \Sigma.$$

Equating to 0

$$0 = N\Sigma - \sum_{i=1}^{N} (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^T.$$