HW2 Report

學號：B06705024 系級：資管四 姓名：郭宇軒

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
Logistic Model	0.86093	0.85640
Generative Model	0.86117	0.85468

我們可以發現， generative model 在 kaggle public score 上有稍高一點點的表現，但在 kaggle private score 上，則由 logistic model 表現較高，而相對上 logistic model 在 private score 的表現更優於 generative model 在 public score 的分數。

但因為 logistic regression 屬於找出一組參數讓預測值與實際值誤差越小越好，且也有更多的調整空間，在 testing data上會有更好的結果。因此認為 logistic regression 為較佳的訓練方式。

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

	Public Score	Private Score
Logistic Model with Normalization	0.86093	0.85640
Logistic Model without Normalization	0.77481	0.77877
Generative Model with Normalization	0.86117	0.85468
Generative Model without Normalization	0.77493	0.76808

可以看到經過feature normalization後，會讓 Logistic Model 和 Generative Model的準確率都大幅度的提升。

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

我使用 sklearn 中的 GradientBoostingClassifier。並且在 data preprocess 中加上了所有連續變數的平方項，其中 age 和 capital_gain 加入三次和四次方項，並再加入所有連續變數的 log，完成預處理。

經過了參數的調整後，最後確定 n_estimators=550, learning_rate=0.2, random_state=42, min_samples_split=1550, min_samples_leaf=15, max_depth=4, max_features=‘sqrt'，得到了 public score：0.87616 和 private score ：0.87397 的成績，相比於 logistic 和 generative 都是顯著的提升。

4. (3%) Refer to math problem

1.

Likelihood Function:

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

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

If $\ln(L(\theta))$ existed maximum, subject to $\sum_{k=1}^{K}\pi_k = 1$, we can find $L(\theta)$ maximum value.

$$L(\pi,\lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{n,k}\left[\ln\left(P(x_n|C_n)\right) + \ln \pi_k\right] + \underbrace{\lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)}_{\text{lagrange multiplier}}$$

$$\frac{\partial L}{\partial \pi_k} = 0 , \quad \frac{1}{\pi_k}\sum_{n=1}^{N} t_{n,k} + \lambda = 0$$

$$\frac{1}{\pi_k} N_k = -\lambda \Rightarrow \pi_k = \frac{-N_k}{\lambda} \quad \text{——} \ ①$$

$$\frac{\partial L}{\partial \lambda} = 0 , \quad \sum_{k=1}^{K}\pi_{k-1} - 1 = 0 \Rightarrow \sum_{k=1}^{K}\pi_k = 1 \quad \text{——} \ ②$$

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

$$\Rightarrow \pi_k = \frac{-N_k}{N} \ \#$$

2. $$\frac{\partial \log(\det \Sigma)}{\partial \sigma_{ij}} = \frac{\partial \det(\Sigma)}{\partial \sigma_{ij}} \cdot \frac{1}{\det(\Sigma)}$$

$$= \frac{1}{\det(\Sigma)} \cdot \frac{\partial \sum (-1)^{i+j} \sigma_{ij} M_{ij}}{\partial \sigma_{ij}}$$

$$= \frac{1}{\det(\Sigma)} \cdot (-1)^{i+j} M_{ij} \quad ①$$

$$e_j \Sigma^{-1} e_i^T = e_j \frac{1}{\det(E)} \tilde{\Sigma} \cdot e_i^T$$

$$= \frac{1}{\det(E)} (-1)^{i+j} M_j e_i^T$$

$$= \frac{1}{\det(E)} (-1)^{i+j} M_{ij}. \quad ②$$

Since ① and ② are equal

$\Rightarrow$ proved #.

3. $p(x|C_k) = \mathcal{N}(x|\mu_k, \Sigma)$ (Gaussian)

$$L(\theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{ \frac{-1}{2} (x^{(i)} - \mu_k)^T \Sigma^{-1} (x^{(i)} - \mu_k) \right\}$$

$$\ln(L(\theta)) = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{-n}{2} \ln(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} \left[(x^{(i)} - \mu_k)^T \Sigma^{-1} (x^{(i)} - \mu_k) \right]$$

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} \ln \mathcal{N}(x|\mu_k, \Sigma)$$

$$= \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{n,k} (x_n - \mu_k)^T \Sigma^{-1} (x_n - \mu_k) + \lambda$$

$$\frac{\partial L}{\partial \mu_k} = 0, \quad \frac{\partial \sum_{n=1}^{N} t_{n,k} (x_n - \mu_k)^T \Sigma^{-1} (x - \mu_k)}{\partial \mu} = 0$$

$$\Rightarrow \sum_{n=1}^{N} t_{n,k} \Sigma^{-1} (x_n - \mu_k) = 0 \quad \text{since } \Sigma \text{ is positive definite}$$

$$\Rightarrow N_k \mu_k - \sum_{n=1}^{N} t_{n,k} x_n = 0$$

$$\mu_k = \frac{\sum_{n=1}^{N} t_{n,k} x_n}{N_k}$$

$$L(\mu, \Sigma | x^n) = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{-n}{2} \ln(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} \left[(x^n - \mu_k)^T \Sigma^{-1} (x^n - \mu_k) \right]$$

$$= \frac{-N}{2} \log|\Sigma^{-1}| - \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \left[t_{n,k} (x^n - \mu_k)^T \Sigma^{-1} (x^n - \mu_k) \right] + \lambda$$

$$\frac{\partial L}{\partial \Sigma^{-1}} = 0, \quad \frac{-1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{n,k} (x_n - \mu_k)(x_n - \mu_k)^T + \frac{N}{2} \Sigma = 0$$

$$\Rightarrow \Sigma = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} t_{n,k} (x_n - \mu_k)(x_n - \mu_k)^T}{N}$$

$$= \frac{N_k}{N} \times \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} t_{n,k} (x_n - \mu_k)(x_n - \mu_k)^T}{N_k}$$

$$= \frac{N_k}{N} \cdot S_k \quad *$$