學號：R07725019 系級： 資管碩二 姓名：鄒雅雯

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

Generative Model：

Kaggle Public score：0.84398 / Kaggle Private score：0.84338

Logistic Regreesion：

Kaggle Public score：0.85528 / Kaggle Private score：0.85014

無論是在 Kaggle 的 Public 還是 Private score 上，Logistic Regression 的結果皆較佳。

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

以下皆以 Logistic Regression 進行實作：

Without normalization：

Kaggle Public score：0.84398 / Kaggle Private score：0.84338

With normalization

Kaggle Public score：0.85528 / Kaggle Private score：0.85014

從結果中看出有對特徵進行標準化，大幅改善的模型準確率。

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

我使用的 best model 是 Random Forest，資料前處理的部分一樣對特徵進行標準化，而後用 Random Forest 進行訓練。結果如下：

Kaggle Public score：0.79852 / Kaggle Private score：0.79412

Random Forest 在 Training data 上訓練得到將近 100%的準確率，但在 Testing data 上的表現較不佳，得出使用 Random Forest 很容易 Overfitting。

4. (3%) Refer to math problem

https://hackmd.io/0fDimqO7RaSCPpD_minSGQ?both

*1

$$P(C_k|x) = \frac{P(C_k, x)}{P(x)} = \frac{P(x|C_k)P(C_k)}{P(x)}$$

Likelihood function: $C_{i,k}$ = x_i 所屬的類別

$$P(t) = P(C_{1,k}|x_1)\,P(C_{2,k}|x_2)\cdots P(C_{N,k}|x_N)$$

$$= \frac{P(x_1|C_{1,k})P(C_{1,k})}{P(x_1)}\,\frac{P(x_2|C_{2,k})P(C_{2,k})}{P(x_2)}\cdots\frac{P(x_N|C_{N,k})P(C_{N,k})}{P(x_N)}$$

∵ $P(x_1) \sim P(x_N)$ 是固定的，不影響 $P(t)$

$$\rightarrow P(t) = P(x_1|C_{1,k})P(C_{1,k})\,P(x_2|C_{2,k})P(C_{2,k})\cdots P(x_N|C_{N,k})P(C_{N,k})$$

$P(x_i|C_{i,k}) = N(x_i|\mu_{i,k}, \Sigma) \rightarrow N$ is Gaussian distribution

mean = $\mu_{i,k}$: x_i 所屬類別 C_k 的 μ

Σ : 所有類別都一樣

$$\rightarrow P(t) = \pi_{1,k}N(x_1|\mu_{1,k},\Sigma)\,\pi_{2,k}N(x_2|\mu_{2,k},\Sigma)\cdots\pi_{N,k}N(x_N|\mu_{N,k},\Sigma)$$

$$= \prod_{n=1}^{N}\left[(\pi_k N(x_n|\mu_k,\Sigma))^{t_n}\prod_{\substack{i=1\\i\neq k}}^{K}(\pi_i N(x_n|\mu_i,\Sigma))^{1-t_n}\right]$$ k: x_n 所屬類別

$\max P(t) = \max \ln P(t)$

$$\rightarrow L = \ln P(t) = \sum_{n=1}^{N}\left[t_n \ln \pi_k N(x_n|\mu_k,\Sigma) + \sum_{\substack{i=1\\i\neq k}}^{K}(1-t_n)\ln \pi_i N(x_n|\mu_i,\Sigma)\right]$$

$$= \sum_{n=1}^{N} t_n \ln \pi_k + \sum_{n=1}^{N} t_n \ln N(x_n|\mu_k,\Sigma) + \sum_{n=1}^{N}\sum_{\substack{i=1\\i\neq k}}^{K}(1-t_n)\ln \pi_i + \sum_{n=1}^{N}\sum_{\substack{i=1\\i\neq k}}^{K}(1-t_n)\ln N(x_n|\mu_i,\Sigma)$$

$$\frac{\partial L}{\partial \pi_k} = \frac{\partial\left(\sum_{n=1}^{N} t_n \ln \pi_k + \sum_{n=1}^{N}\sum_{\substack{i=1\\i\neq k}}^{K}(1-t_n)\ln \pi_i\right)}{\partial \pi}$$ (其它跟 π 無關)

$$\because \sum_{i=1}^{K}\pi_i = 1 \rightarrow \sum_{\substack{i=1\\i\neq k}}^{K}\pi_i = 1-\pi_k \rightarrow \sum_{\substack{i=1\\i\neq k}}^{K}\ln \pi_i = \ln(1-\pi_k)$$

$$\rightarrow \frac{\partial L}{\partial \pi_k} = \frac{\partial\left(\sum_{n=1}^{N} t_n \ln \pi_k + \sum_{n=1}^{N}(1-t_n)\ln(1-\pi_k)\right)}{\partial \pi_k} = \sum_{n=1}^{N}\left[\frac{t_n}{\pi_k} - \frac{(1-t_n)}{1-\pi_k}\right] = 0$$

$$\rightarrow \sum_{n=1}^{N}\left[t_n(1-\pi_k) - (1-t_n)\pi_k\right] = \sum_{n=1}^{N}\left[t_n - t_n\pi_k - \pi_k + t_n\pi_k\right] = \sum_{n=1}^{N}\left[t_n - \pi_k\right] = \sum_{n=1}^{N} t_n - N\pi_k = 0$$

$$\rightarrow N\pi_k = \sum_{n=1}^{N} t_n \Rightarrow \pi_k = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_k}{N}$$

*2

$$\frac{\partial \log(\det \Sigma)}{\partial \sigma_{ij}} = \frac{\partial \log(\det \Sigma)}{\partial \det \Sigma} \frac{\partial \det \Sigma}{\partial \sigma_{ij}}$$

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

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

*3

$$L(\mu_k, \Sigma_k) = f(x_1)^{t_{1k}} f(x_2)^{t_{2k}} \cdots f(x_N)^{t_{Nk}} = \prod_{n=1}^{N} f(x_n)^{t_{nk}}$$

$$= \prod_{n=1}^{N} \left[\frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}(x_n-\mu_k)^T \Sigma_k^{-1} (x_n-\mu_k)\right)\right]^{t_{nk}}$$

$$\max L(\mu_k, \Sigma_k) = \max \log L(\mu_k, \Sigma_k)$$

$$\log L(\mu_k, \Sigma_k) = \sum_{n=1}^{N} t_{nk} \log\left(\frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}(x_n-\mu_k)^T \Sigma_k^{-1} (x_n-\mu_k)\right)\right)$$

$$= \sum_{n=1}^{N} \frac{-D}{2} t_{nk} \log(2\pi) + \sum_{n=1}^{N} \frac{-1}{2} t_{nk} \log|\Sigma_k| + \sum_{n=1}^{N} t_{nk} \left[\frac{-1}{2}(x_n-\mu_k)^T \Sigma_k^{-1} (x_n-\mu_k)\right]$$

$$\frac{\partial \log L(\mu_k, \Sigma_k)}{\partial \mu_k} = \sum_{n=1}^{N} t_{nk} \frac{-1}{2} \left[-\Sigma_k^{-1}(x_n-\mu_k) - (x_n-\mu_k)^T \Sigma_k^{-1} \right]$$

$$= \sum_{n=1}^{N} t_{nk} \frac{-1}{2} \left(-2 \Sigma_k^{-1} (x_n-\mu_k) \right) = \sum_{n=1}^{N} t_{nk} \Sigma_k^{-1} (x_n-\mu_k) = 0$$

$$\rightarrow \sum_{n=1}^{N} t_{nk}(x_n-\mu_k) = 0 \rightarrow \sum_{n=1}^{N} t_{nk} x_n - \sum_{n=1}^{N} t_{nk}\mu_k = 0$$

$$\rightarrow N_k \mu_k = \sum_{n=1}^{N} t_{nk} x_n \rightarrow \underline{\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} x_n} \ \#$$

$$\log L(\mu_k, \Sigma_k) = \sum_{n=1}^{N} \frac{-D}{2} t_{nk} \log(2\pi) + \sum_{n=1}^{N} \frac{-1}{2} t_{nk} \log|\Sigma_k| + \sum_{n=1}^{N} t_{nk} \frac{-1}{2} \operatorname{tr}\left[(x_n-\mu_k)(x_n-\mu_k)^T \Sigma_k^{-1}\right] \text{(by hint)}$$

$$\frac{\partial \log L(\mu_k, \Sigma_k)}{\partial \Sigma_k^{-1}} = \sum_{n=1}^{N} \frac{t_{nk}}{2} \Sigma - \frac{1}{2} \sum_{n=1}^{N} t_{nk} (x_n-\mu_k)(x_n-\mu_k)^T = 0$$

$$\rightarrow \sum_{n=1}^{N} t_{nk} \Sigma - \sum_{n=1}^{N} t_{nk} (x_n-\mu_k)(x_n-\mu_k)^T = 0$$

$$\rightarrow N_k \Sigma_k = \sum_{n=1}^{N} t_{nk} (x_n-\mu_k)(x_n-\mu_k)^T$$

$$\rightarrow \Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} (x_n-\mu_k)(x_n-\mu_k)^T = S_k$$

By Problem 1 (Mixture)

$$p(x|\theta, \pi) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma)$$

$$\rightarrow \Sigma = \sum_{k=1}^{K} \pi_k S_k = \sum_{k=1}^{K} \frac{N_k}{N} S_k$$

where $S_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} (x_n-\mu_k)(x_n-\mu_k)^T$