學號：R10922111 系級： 資工所碩一 姓名：黃冠瑋

1. (1%) 請比較說明generative model、logistic regression兩者的異同為何？再分別列出本次使用的資料中五個分得正確/不正確的sample, 並說明為什麼如此？

generative model 是根據數學理論推出可能值，logistic regression則是一直調整直到尋找到適合的參數。

2. (1%) 請實作兩種feature scaling的方法 (feature normalization, feature standardization), 並說明哪種方法適合用在本次作業？

	normalization	standardization
logistic regression	1.69399	1.68803

normalization 較適合，因為他會把資料轉換成0~1，且因為搜集資料的feature只有前面幾個需要做調整，後面都屬於binary feature。因此轉換後，這時所有的feature都會在同一個範圍內，在進行梯度下降時便可以減少迭代次數，增加精準度。

3. (1%) 請說明你實作的best model及其背後「原理」為何？你覺得這次作業的dataset比較適合哪個model？為什麼？

我使用decision tree 來做分類。他運行的方式就像一棵樹，從root往下到leaf，每個節點都是一個feature，根據feature再細分成不同的節點。

我覺得這次作業適合decision tree這個模型，因為我們的feature有很大一部分都屬於離散的，很容易做劃分。

4. (3%) Refer to math problem

https://hackmd.io/@GfOkB4kgS66YhhM7j6TJew/BJ-wGv8HY

1.

$$P(x_n, t_n) = \prod_{k=1}^{K} [P(x_n|C_k)\,P(C_k)]^{t_{nk}} = \prod_{k=1}^{K} (P(x_n|C_k)\,\pi_k)^{t_{nk}}$$

⇒ Max likelihood function $L(\theta) = \prod_{n=1}^{N}\prod_{k=1}^{K} (P(x_n|C_k)\,\pi_k)^{t_{nk}}$

$\log L(\theta) = \ell(\theta) = \sum\sum t_{nk}[\log P(x_n|C_k) + \log \pi_k]$

Since $\sum_{k=1}^{K} \pi_k = 1$. $\therefore \ell(\theta,\lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}[\log P(x_n|C_k) + \log \pi_k] + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$

$$\frac{\partial \ell(\theta)}{\partial \pi_k} = \frac{1}{\pi_k}\sum_{n=1}^{N} t_{nk} + \lambda = 0 \quad \rightarrow \quad \pi_k = -\frac{\sum_{n} t_{nk}}{\lambda} = \frac{-N_k}{\lambda}$$

↑ To maximize $\ell(\theta)$, $\frac{\partial \ell(\theta,\lambda)}{\partial \pi_k} = 0 = \frac{\partial \ell(\theta,\lambda)}{\partial \lambda}$

$$\frac{\partial \ell(\theta,\lambda)}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 = 0 \quad \Rightarrow \quad \sum_{k=1}^{K}\pi_k = 1 = \sum_{k=1}^{K}\frac{-N_k}{\lambda} = \frac{-N}{\lambda}$$

$$\rightarrow \lambda = -N$$

So, $\pi_k = \frac{-N_k}{-N} = \frac{N_k}{N}$

2.

Assume $A \in \mathbb{R}^{m\times m}$, and it's non-singular. $|A| = \sum_{j=1}^{m}(-1)^{i+j} a_{ij}|A_{ij}| = \sum_{i=1}^{m}(-1)^{i+j} a_{ij}|A_{ij}|$

then $\frac{\partial |A|}{\partial a_{ij}} = (-1)^{i+j}|A_{ij}|$

And since A is non-singular, $\therefore \exists\, X \in \mathbb{R}^{m\times 1}$, s.t. $AX = e_i^T \Rightarrow X = A^{-1}e_i^T$

$\rightarrow X^{(j)} = e_j A^{-1} e_i^T$

By Cramer's rule, $X^{(j)} = \frac{(-1)^{i+j}|A_{ji}|}{|A|} = e_j A^{-1} e_i^T$

By chain rule, $\frac{\partial \log|A|}{\partial a_{ij}} = \frac{1}{|A|}\frac{\partial |A|}{\partial a_{ij}} = \frac{(-1)^{i+j}}{|A|}|A_{ij}| = e_j A^{-1} e_i^T$

let $A = \Sigma$, we get $\frac{\partial \log|\Sigma|}{\partial \sigma_{ij}} = e_j \Sigma^{-1} e_i^T$

3.

By (1), maximum likelihood function : $\prod_{n=1}^{N} (z_1 N(x_n|\mu_1,\Sigma))^{t_{n1}} (z_2 N(x_n|\mu_2,\Sigma))^{t_{n2}} \cdots (z_k N(x_n|\mu_k,\Sigma))^{t_{nk}}$

$\log L = \sum_{n=1}^{N} t_{n1} \log N(x_n|\mu_1,\Sigma) + t_{n2} \times \log N(x_n|\mu_2,\Sigma) + \cdots + t_{nk} \times \log N(x_n|\mu_k,\Sigma) + t_{n1}\log z_1 + t_{n2}\log z_2 + \cdots$

Then, to maximize likelihood for mean, $\frac{\partial \log L}{\partial \mu_k} = 0$

$$\frac{\partial \log L}{\partial \mu_k} = \sum_{n=1}^{N} t_{nk} \frac{\partial}{\partial \mu_k} \left(-\frac{k}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right)$$

$$= \sum_{n=1}^{N} t_{nk} \frac{\partial}{\partial(x-\mu_k)} \left(-\frac{k}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) \right) \cdot \frac{\partial(x-\mu_k)}{\partial \mu_k}$$

$\because \frac{\partial N^T A N}{\partial N} = 2AN$ if N is independent from A and $A^T = A$

$$= \sum_{n=1}^{N} t_{nk} \left(-\frac{1}{2} \cdot 2\Sigma^{-1}(x-\mu_k) \right) \cdot (-1)$$

$$= \sum_{n} t_{nk} \Sigma^{-1}(x-\mu_k) = 0$$

$$\Rightarrow 0 = \sum_{n} t_{nk} x_n - \sum_{n} t_{nk} \mu_k$$

$$\Rightarrow \sum_{n} t_{nk} x_n = N_k \mu_k \quad \Rightarrow \quad \mu_k = \frac{\sum_{n=1}^{N} t_{nk} x_n}{N_k}$$

$$\log L(\mu, \Sigma | x_n) = \sum_{n=1}^{N} C_n + t_{n1} \log N(x_n|\mu_1,\Sigma) + \cdots + t_{nk}\log N(x_n|\mu_k,\Sigma)$$

$$\frac{\partial \ell(\mu,\Sigma|x_n)}{\partial \Sigma^{-1}} = \frac{\partial}{\partial \Sigma^{-1}} \left(\sum_{n=1}^{N} C_n + t_{n1}\left(C_1' + \frac{1}{2}\log|\Sigma^{-1}| - \frac{1}{2} \cdot \mathrm{tr}\left[(x_n-\mu_1)(x_n-\mu_1)^T \Sigma^{-1} \right] \right) + \cdots + t_{nk}\left(C_k' + \frac{1}{2}\log|\Sigma^{-1}| - \frac{1}{2} \cdot \mathrm{tr}\left[(x_n-\mu_k)(x_n-\mu_k)^T \Sigma^{-1} \right] \right) \right)$$

$$= \sum_{n=1}^{N} \left(t_{n1}\left(\frac{1}{2}\Sigma - \frac{1}{2}(x_n-\mu_1)(x_n-\mu_1)^T \right) + \cdots + t_{nk}\left(\frac{1}{2}\Sigma - \frac{1}{2}(x_n-\mu_k)(x_n-\mu_k)^T \right) \right)$$

$$= 0$$

$$\Rightarrow \sum_{n=1}^{N} \frac{1}{2}\Sigma \underbrace{(t_{n1} + \cdots + t_{nk})}_{t_{n1}+\cdots+t_{nk}=1} = \frac{1}{2}\sum_{n=1}^{N} t_{n1}(x_n-\mu_1)(x_n-\mu_1)^T + \cdots + t_{nk}(x_n-\mu_k)(x_n-\mu_k)^T$$

$$\sum_{k=1}^{k} \frac{N_k}{N} S_k \quad \#$$

$$\therefore N \cdot \Sigma = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk}(x_n-\mu_k)(x_n-\mu_k)^T$$

$$\Rightarrow \Sigma = \frac{1}{N} \cdot \sum_{k=1}^{K} \sum_{n=1}^{N} t_{nk}(x_n-\mu_k)(x_n-\mu_k)^T = \sum_{k=1}^{k} \frac{N_k}{N} \cdot \frac{1}{N_k} \underbrace{\sum_{n=1}^{N} t_{nk}(x_n-\mu_k)(x_n-\mu_k)^T}_{S_k}$$