2) (a) (5pts) Write down the log-likelihood implied by this model and find the maximum likelihood estimator (MLE) for the priors $p(\mathcal{C}_k) = \pi_k$ and the class means $\boldsymbol{\mu}_k$, for $k = 1, ..., K$. Note that you do not need to derive the MLE for the covariance matrix.

(2.1) $$p(\mathbf{x}|\mathcal{C}_k) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}).$$

We know that the posterior $p(\mathcal{C}_k|\mathbf{x})$ can be written in terms of the softmax function

(2.2) $$p(\mathcal{C}_k|\mathbf{x}) = \frac{\exp\{a_k\}}{\sum_j \exp\{a_j\}} \quad \text{where} \quad a_k = \mathbf{w}_k^T\mathbf{x} + w_{k0}.$$

Here, we also know that

(2.3) $$\mathbf{w}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \quad \text{and} \quad w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log(p(\mathcal{C}_k)).$$

$$P(x|C_k) \propto p(C_k|x)\,P(x)$$

Then we have $\prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k P(x_n|C_k)^{t_{nk}}$

Then log likelihood of the model is:

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

① MLE for $P(C_k) = \pi_k$:

introduce lagrangian multiplier to conserve constraint $\sum_{k=1}^{K} \pi_{nk} = 1$

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

Derive WRT π_k and let it equal to 0

$$\frac{\partial}{\partial \pi_k}\mathcal{L}(\pi_k, \lambda) = \sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda$$

$$\rightarrow \sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda = 0$$

$$\pi_k = -\frac{1}{\lambda}\sum_{n=1}^{N} t_{nk}$$

Sum both sides over k:

$$\sum_{k=1}^{K} \pi_k = -\frac{1}{\lambda}\sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk} = 1$$ due to constraint $\sum_{k=1}^{K} \pi_k = 1$

Then $-\frac{1}{\lambda}\sum_{k=1}^{K} N_k = 1$ where N_k denotes # of data with class label k

Then $\lambda = -N$

Hence $$\hat{\pi}_k = \frac{\sum_{n=1}^{N} t_{nk}}{N} = \frac{N_k}{N}$$

② MLE for μ_k :

Since,
$$p(x_n | C_k) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x_n-\mu_k)^T \Sigma^{-1}(x_n-\mu_k)\right\}$$

The log likelihood of the model is :
$$\mathcal{L}(\pi_k, \mu_k) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left[-\frac{1}{2}(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k) - \ln(|2\pi\Sigma|^{1/2}) + \ln \pi_k\right]$$

Deriving WRT μ_k :
$$\frac{\partial}{\partial \mu_k} \mathcal{L}(\pi_k, \mu_k) = \sum_{n=1}^{N} t_{nk}\Sigma^{-1}(x_n-\mu_k) = 0$$

Then,
$$\sum_{n=1}^{N} t_{nk}\Sigma^{-1}x_n = \sum_{n=1}^{N} t_{nk}\Sigma^{-1}\mu_k$$
$$\Sigma^{-1}\sum_{n=1}^{N} t_{nk}x_n = \Sigma^{-1}\mu_k \sum_{n=1}^{N} t_{nk}$$
$$\sum_{n=1}^{N} t_{nk}x_n = \mu_k \sum_{n=1}^{N} t_{nk}$$

Then
$$\mu_k = \frac{\sum_{n=1}^{N} t_{nk}x_n}{\sum_{n=1}^{N} t_{nk}}$$

Simplifying
$$\hat{\mu}_k = \frac{\sum_{n=1}^{N} t_{nk}x_n}{N_k}$$
where N_k denotes # of data with class label k

c) Training accuracy : 89.39%

Test accuracy : 82.5 %

\# Samples used : 10,000

d)

Comparing the training accuracy, logistic regression has 89.43% and GDA has 89.39%. They are very neck and neck in terms of which model has better accuracy in the training set.
However, comparing the test accuracy, logistic regression has 88% and GDA has 82.5%. We can see that generally speaking, logistic regression has far better accuracy in the training set. Although, we have to keep in mind that logistic model used 60,000 samples for training and GDA only used 10,000 for training as memory couldn't handle more. I do have to comment that I think logistic model is far better, as it is able to compute much faster and has a slightly higher accuracy.