Assignment 2

Ashmitha Jaysi Sivakumar, CE21B024

q1)i)

As the data consists of 1s and 0s, it is a safe assumption to try out a Bernoulli Distribution. The equations and steps followed are given below.

Step 1: Initialise the variables:

π_k : Proportion of the mixture which follows the kth distribution

Θ : Params to define the distribution

γ : lambda

Step 2:

Find the probability distribution function (bernoulli) for each data point according to the initialized params

$$P(x \mid \{\theta_k\}, \{\pi_k\}) = \sum_{k=1}^{K} \pi_k \cdot \theta_k^x (1-\theta_k)^{1-x}$$

$$\gamma(z_{nk}) = \frac{\pi_k \prod_{i=1}^{5} P(x_{ni}|\theta_{ki})}{\sum_{j=1}^{K} \pi_j \prod_{i=1}^{5} P(x_{ni}|\theta_{ji})}$$

Step 3:

$$\theta_{ki} = \frac{\sum_{n=1}^{10} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{10} \gamma(z_{nk})}$$

$$\pi_k = \frac{1}{10} \sum_{n=1}^{10} \gamma(z_{nk})$$

Calculate the new theta and pi for the new iteration

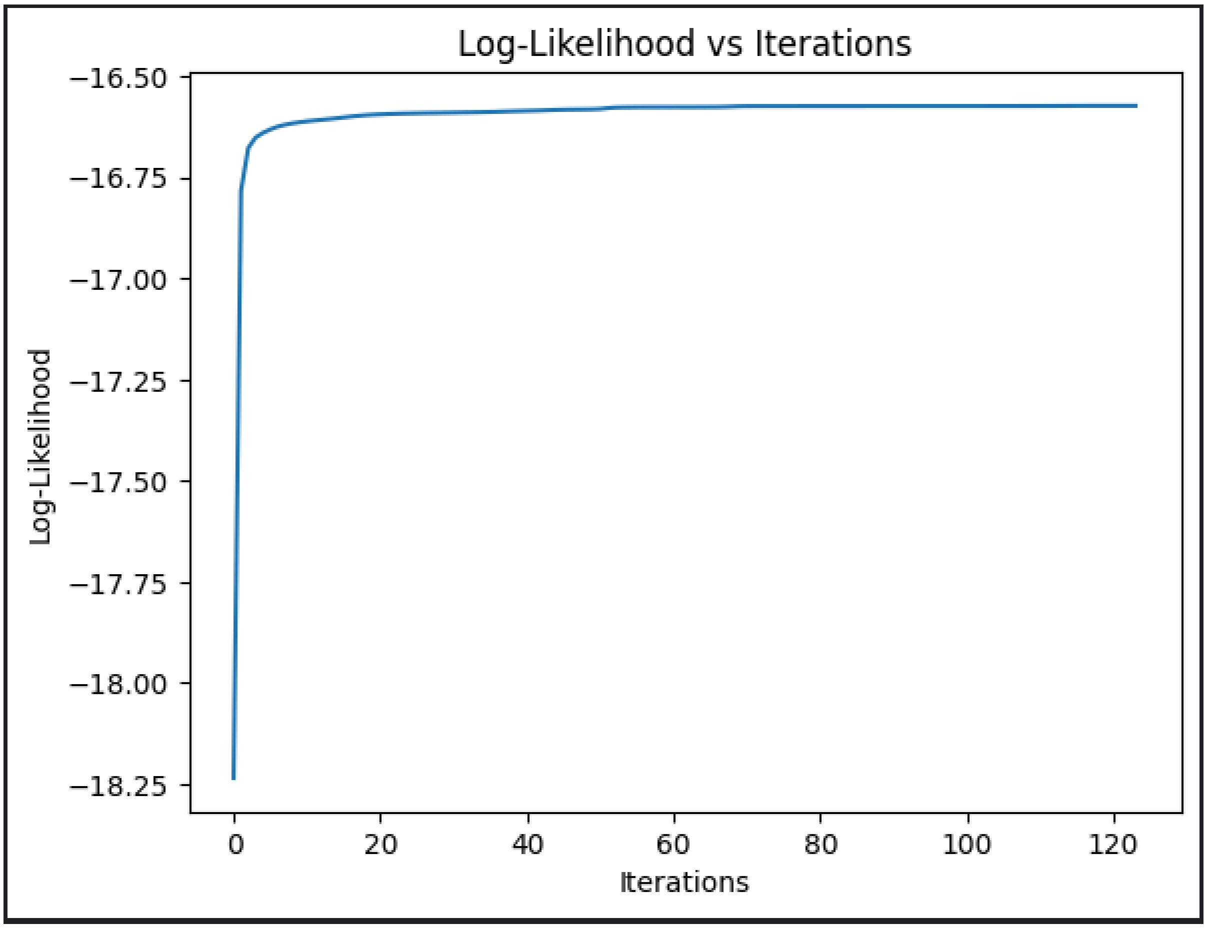
Step 4:

FInd the likelihoods and the log likelihood

$$\text{likelihoods}[i] = \sum_{k=1}^{K} \pi_k \cdot P(x_i \mid \theta_k)$$

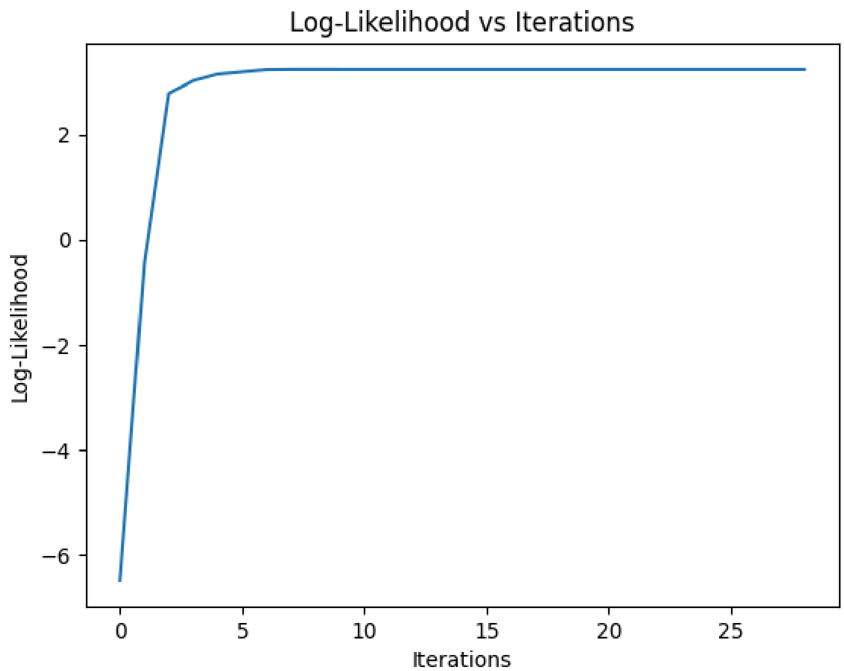
$$\text{log\_likelihood} = \frac{1}{N} \sum_{i=1}^{N} \log\left(\text{likelihoods}[i]\right)$$

After some iterations, the below graph is obtained



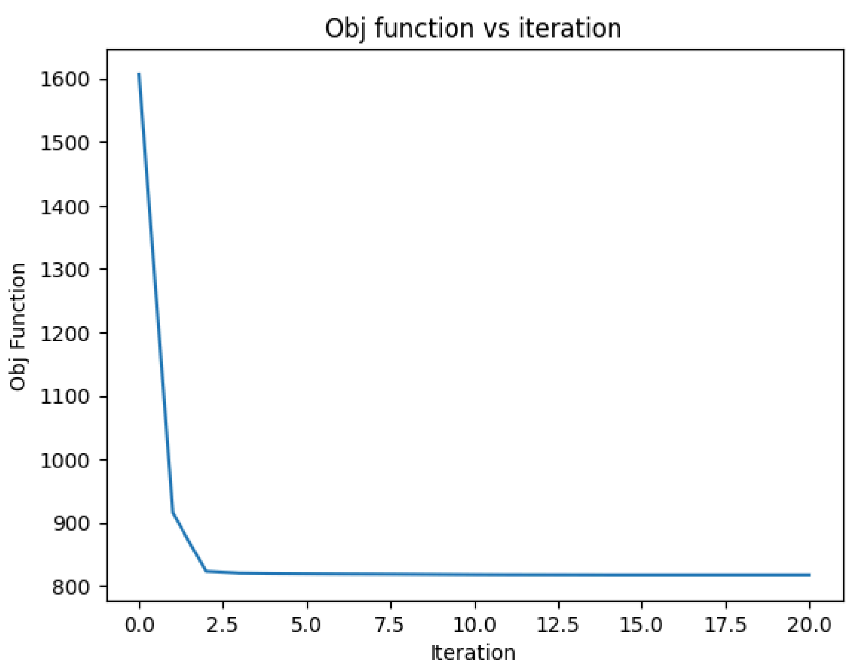
ii) Follow the same steps above but with Gaussian distribution:

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$



The Gaussian Distribution seems to have a better fit than the Bernoulli Distribution because we have obtained a positive log likelihood (~3) in the second qn compared to a negative (~-16.5) log likelihood.

iii) Is with K-Means Algorithm



iv) The first approach took much more iterations and has a negative log likelihood. So that can be considered as not the best algorithm. Between K- means and GMM, GMM can model much more complex data than K-means hence we can choose GMM for this data.

Q2)

i) Least Squared Solution is obtained as:

$$w^* = (XX^T)^{-1} Xy$$

ii) With gradient descent:

Step 1: initialize w

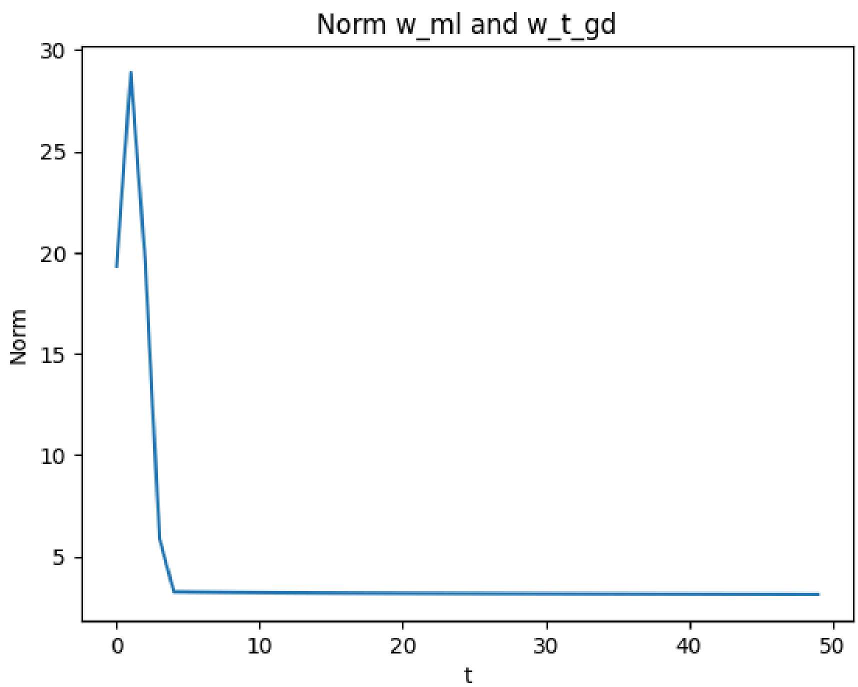
Step 2: use step size as 1/t (t -> iteration number)

Step 3: Iterate with

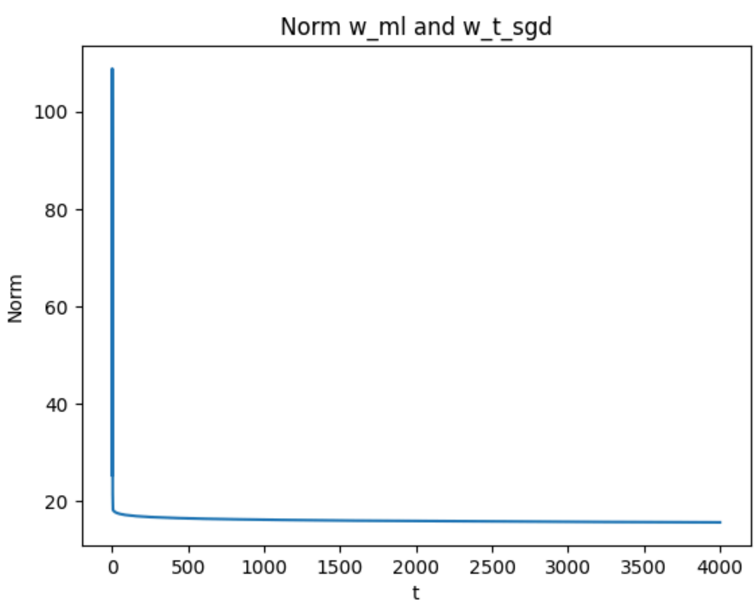
$$\nabla F(w) = \boxed{2(XX^T)w - 2Xy}$$

$$w^{t+1} = w^t - \eta^t \nabla F(w^t)$$

The w values obtained through this approach converges to to that obtained analytically after reaching a peak.

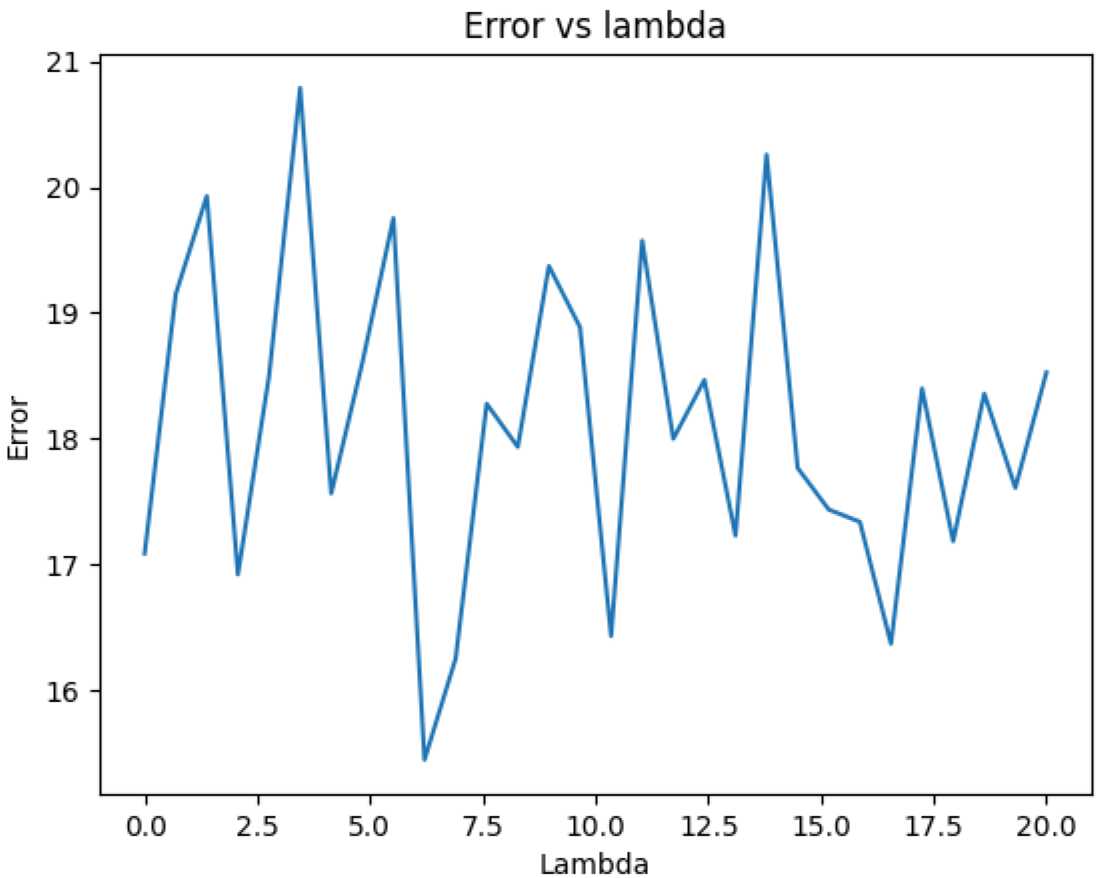


iii)



There is a smooth decrease in the norm and reaching a constant minima.

iv)




```
error for W_ml is 19.92145623761605б
error for W_r 41.30573639172386
```

The analytically obtained coefficients is better. Probably the params such as learning rate and initialisation affects the error.