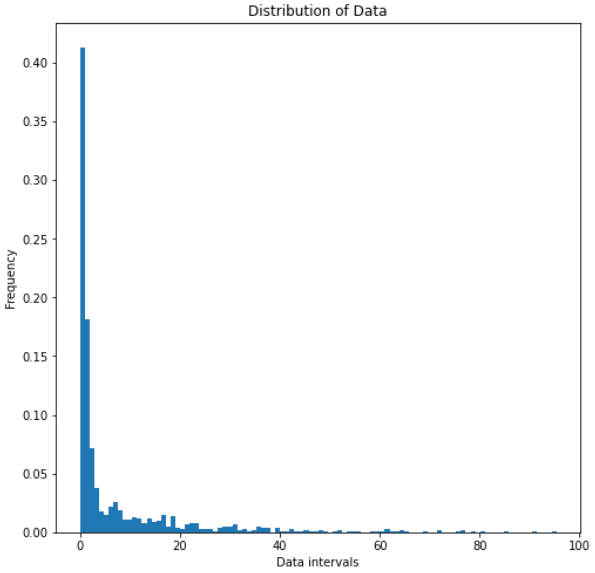
CS5691: Pattern Recognition and Machine Learning
Assignment 2

Name: Chetan Reddy N
Roll Number: ME19B093

Question 1 (i) - Mixture of Exponentials

- The probabilistic mixture that could have generated this data is a **mixture of exponentials**. Justifications for choosing it: (a) The data is positive (b) The distribution is closer to an exponential distribution.



- The EM algorithm for a mixture of exponentials has been derived below:

Exponential Distribution:

$$f_{exp}(x) = \begin{cases} \frac{1}{\mu} e^{-\frac{x}{\mu}} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Mixture of Exponentials:

$$f_x(x) = \sum_{k=1}^{K} \pi_k \frac{1}{\mu_k} e^{-\frac{x}{\mu_k}}$$

K - number of components

μ_k - mean of the component k

π_k - probability that component k is chosen

$$\text{log\_likelihood} = \sum_{i}^{n} \left(\log \left(\sum \pi_k \frac{1}{\mu_k} e^{-x/\mu_k} \right) \right)$$

$$\text{modified\_log\_likelihood} = \sum_{i} \sum_{k} \lambda_k^i \log \left(\frac{\pi_k}{\lambda_k^i} \frac{1}{\mu_k} e^{-\frac{x}{\mu_k}} \right)$$

Fixing λ, maximising on $\theta = \{\pi_1, \pi_2 \dots \pi_K, \mu_1, \mu_2 \dots \mu_K\}$

$$\hat{\mu}_{k_{MML}} = \frac{\sum_i \lambda_k^i x_i}{\sum_i \lambda_k^i}$$

$$\hat{\pi}_{k_{MML}} = \frac{\sum_i \lambda_k^i}{n}$$

PROOF

$$\frac{\partial}{\partial \mu_k} \text{modified\_logL} = 0 \Rightarrow \frac{\partial}{\partial \mu_k} \sum_i \lambda_k^i \left(\log \mu_k + \frac{x_i}{\mu_k} \right) = 0$$

$$\Rightarrow \sum \lambda_k^i \left(\frac{1}{\mu_k} - \frac{x_i}{\mu_k^2} \right) = 0 \Rightarrow \mu_k = \frac{\sum \lambda_k x_i}{\sum \lambda_k}$$

PROOF

Can be proved by using lagrange multiplier similar to mixture of gaussians

Fixing θ, maximising of λ

$$\hat{\lambda}_k^i = \frac{f\left(x_i/\mu_k\right)\pi_k}{\sum_{k=1}^{K} f\left(x_i/\mu_k\right)\pi_l} = \frac{\frac{\pi_k}{\mu_k} e^{-\frac{x_i}{\mu_k}}}{\sum_{l=1}^{K} \frac{\pi_l}{\mu_l} e^{-x_i/\mu_l}}$$

PROOF

$$\text{modified\_log\_likelihood} = \sum_i \sum_k \lambda_k^i \log \left(\frac{\frac{\pi_k}{\mu_k} e^{-\frac{x_i}{\mu_k}}}{\lambda_k^i} \right)$$

$$= \sum_i \sum_k \lambda_k^i \log \left(\frac{a_{ik}}{\lambda_k^i} \right) \quad \text{(} a_{ik} \text{: constant)}$$

$$= \sum_i \sum_k \left[\lambda_k^i \log(a_{ik}) - \lambda_k^i \log \lambda_k^i \right]$$

$\lambda_k^i > 0$, $\lambda_k^i < 1$, $\sum_k \lambda_k^i = 1$

The maximisation solved by method of lagrange multiplier

EM Algorithm

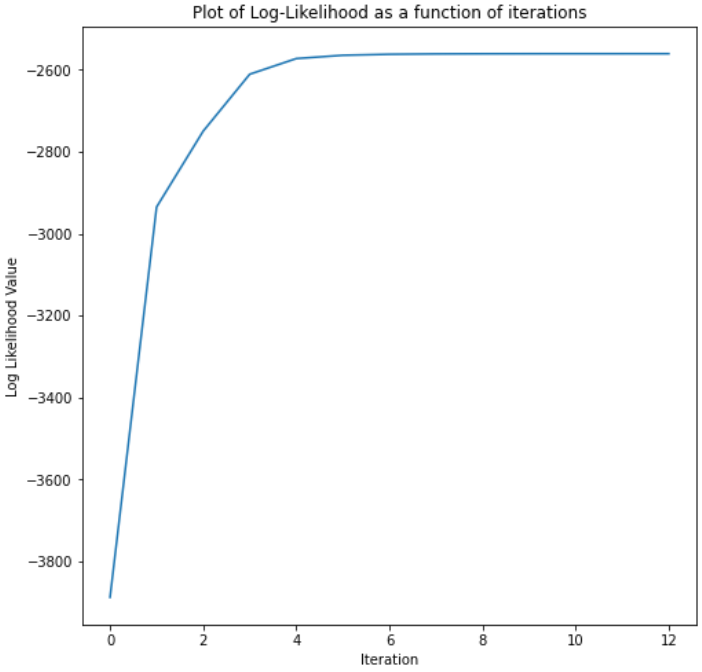
$\theta^0 = \{\pi_1^0, \pi_2^0 \dots \pi_K^0, \mu_1^0, \mu_2^0 \dots \mu_K^0\}$ # theta is initialized

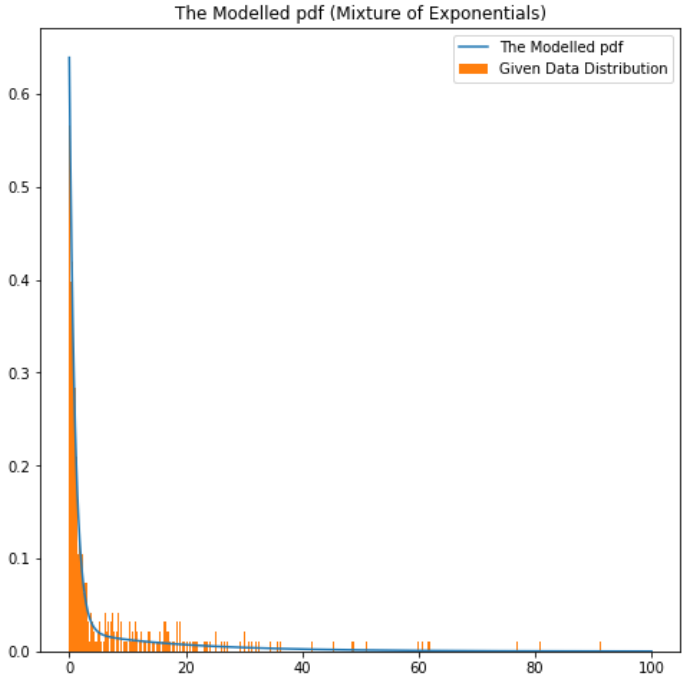
while $\|\theta_\bullet^{t+1} - \theta^t\| < \epsilon$:

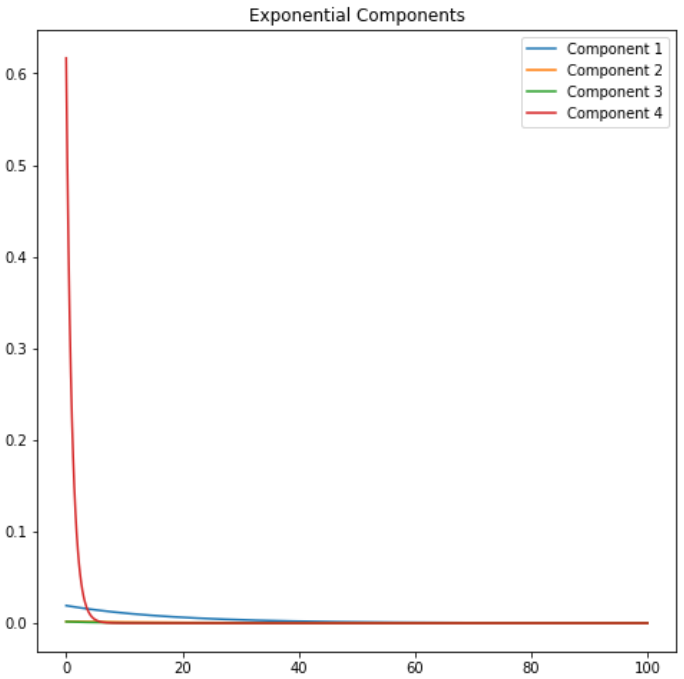
$\lambda^{t+1} = \underset{\lambda}{\text{argmax}}$ modified_logL (θ^t, λ) # E-step

$\theta^{t+1} = \underset{\theta}{\text{argmax}}$ modified_logL (θ, λ^{t+1}) # M-step

Results:





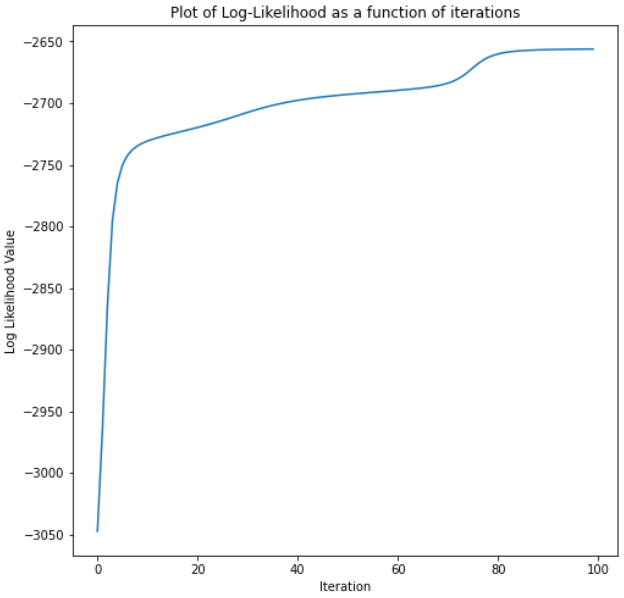


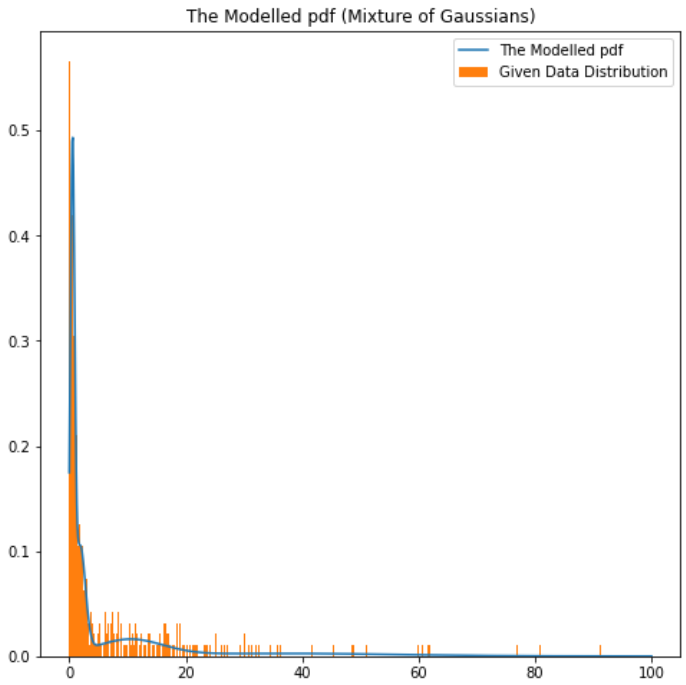
Question 1 (ii) - Mixture of Gaussians

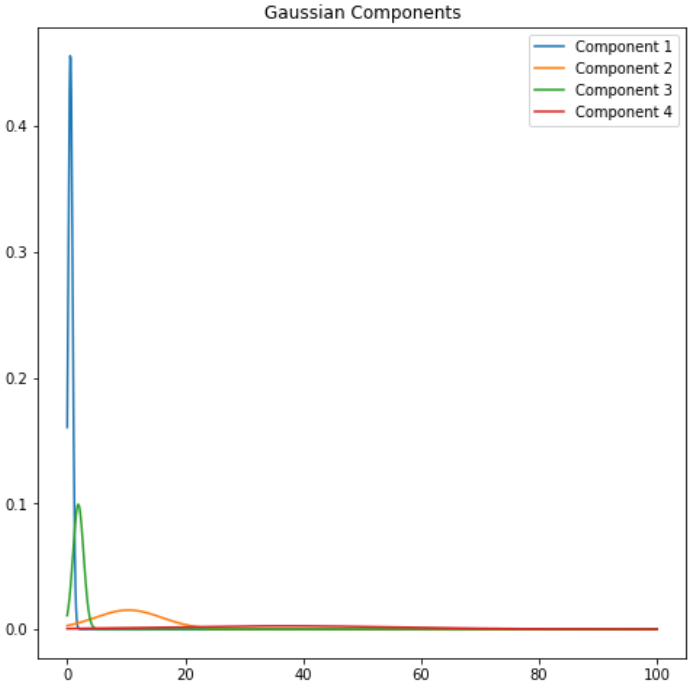
- The EM algorithm is implemented for a mixture of Gaussians. The gaussian mixture is a very versatile model capable of fitting any kind of data. We can see that it fits even skewed data (like the one given) very well.

- To initialise the parameter space (theta), Lloyds algorithm has been used during the implementation

Results:



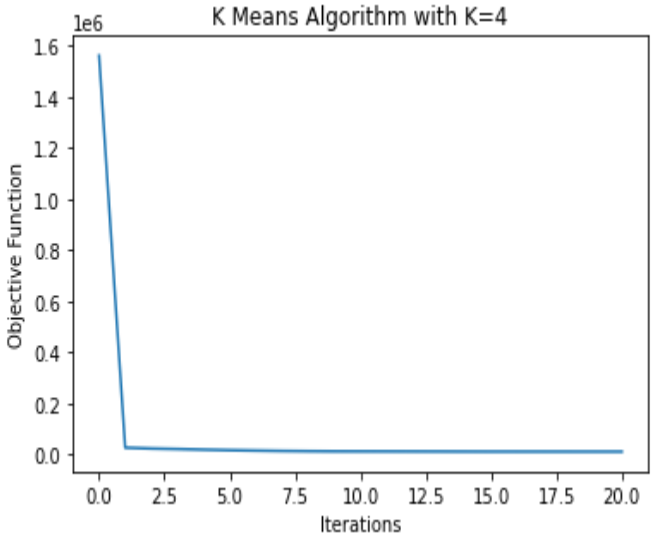


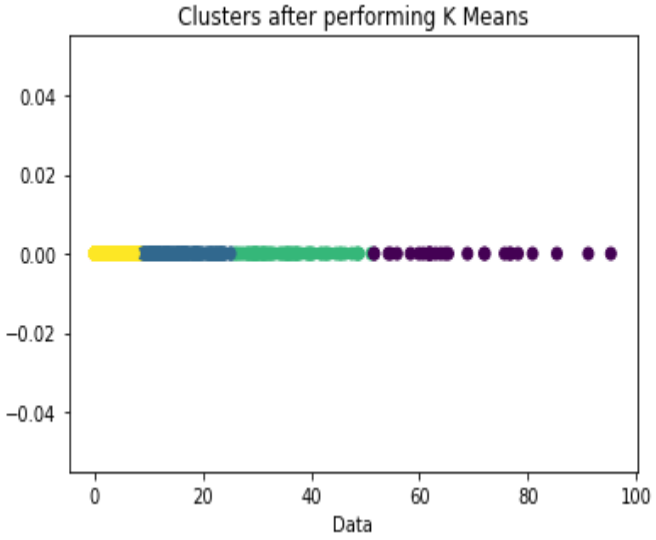


Question 1 (iii) - K Means Algorithm

- K means the algorithm is run on the 1D data with K=4. As expected, it clumps the high frequency data together

Results





Question 1 (iv)

- For the given dataset, the **mixture of exponentials** would be the best model. As mentioned above, the reasons include: (a) Positive Data (b) Distribution looks Exponential (c) The Likelihood value can be seen to be higher than that of gaussian
- While Gaussian Mixture generates a good fit too, it unnecessarily assigns a non-zero probability to negative values. Further, it involves the estimation of more number of parameters compared to exponential.
- K means is an algorithm used to cluster data and therefore it cannot be used to model the data.

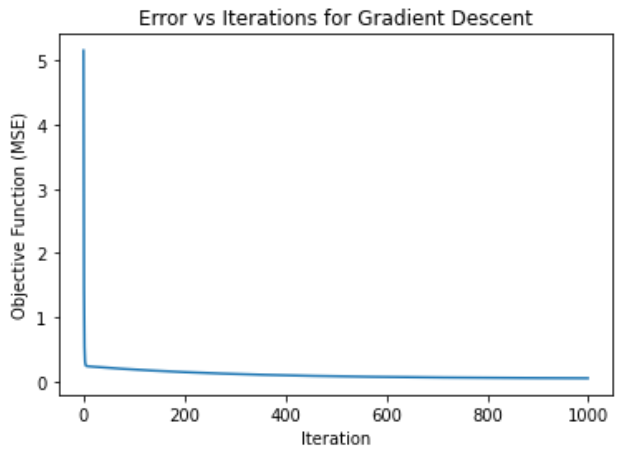
Question 2 (i) - Analytical Solution

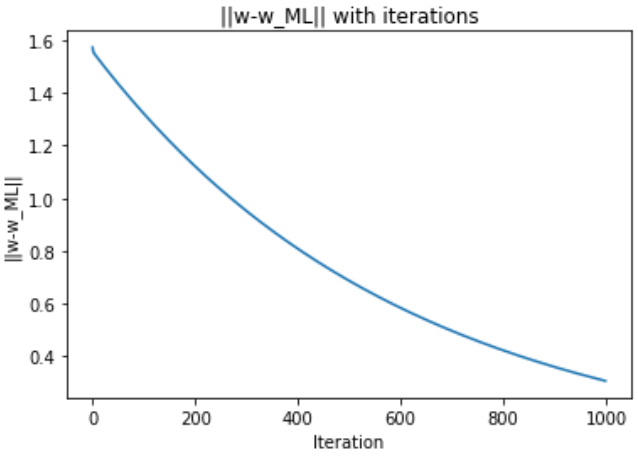
- The analytical solution is obtained and the results are shown below
- Training Set MSE (Mean Squared Error): 0.03969
- Test Set MSE: 0.37073

Question 2 (ii) - Gradient Descent

- Update Step:
  $\mathbf{w^{t+1} = w^t - \eta^t [2X(X^T w^t - y)]}$

- A learning rate (η) of 0.01 is used and the results are as follows:
- Training Set MSE: 0.04
- Test Set MSE: 0.31
-





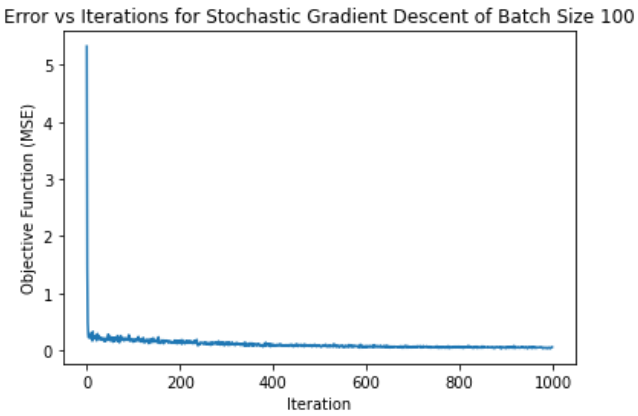
- This shows that the $\mathbf{w^t}$ becomes closer to the analytical solution (w_{ML}) with the number of iterations.
- As the number of iterations approaches infinity, w^t becomes exactly the analytical solution

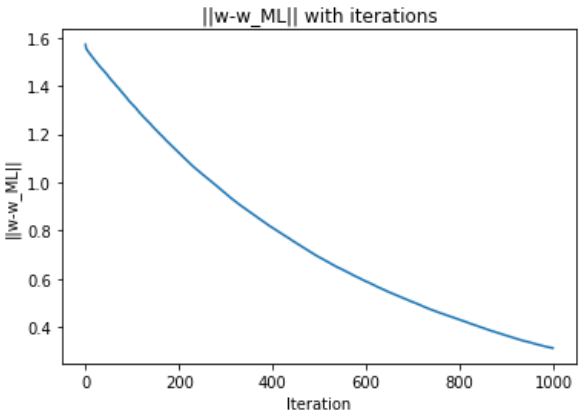
---

Question 2 (iii) - Stochastic Gradient Descent with Batch Size = 100

- The update step
  $\mathbf{w^{t+1} = w^t - \eta^t [2X'(X'^T w^t - y')]}$
  X' and y' are sampled from X and y with batch size of 100

- The results are similar to that of gradient descent





- The objective function is not as smooth as a gradient descent algorithm. This is expected since we are taking a sample of the entire dataset to calculate the gradient which creates noise.
- However, the objective function eventually reaches the optimal value with the number of iterations
- As in the case of gradient descent, w^t becomes closer to w_{ML} as the number of iterations increases.

Question 2 (iv) - Ridge Regression

- Ridge regression includes an additional term in the objective function in addition to mse.
- K-fold cross-validation has been performed to choose the best value of λ
- The best results are obtained for λ=0.05:
  Training Set MSE: 0.09977
  Test Set MSE: 0.23833

w_R - Ridge Regression with λ=0.05	**w_{ML} - Analytical Solution**
Training Set MSE: 0.09977 Test Set MSE: 0.23833	Training Set MSE: 0.03969 Test Set MSE: 0.37073
The ridge regression model performs better on the test dataset which is evident from the test set MSE.	Higher Training set MSE shows that the model is overfitted and thus performs poorly on the test set compared to the ridge regression model

- In conclusion, the ridge regression model is better than an analytical solution

---