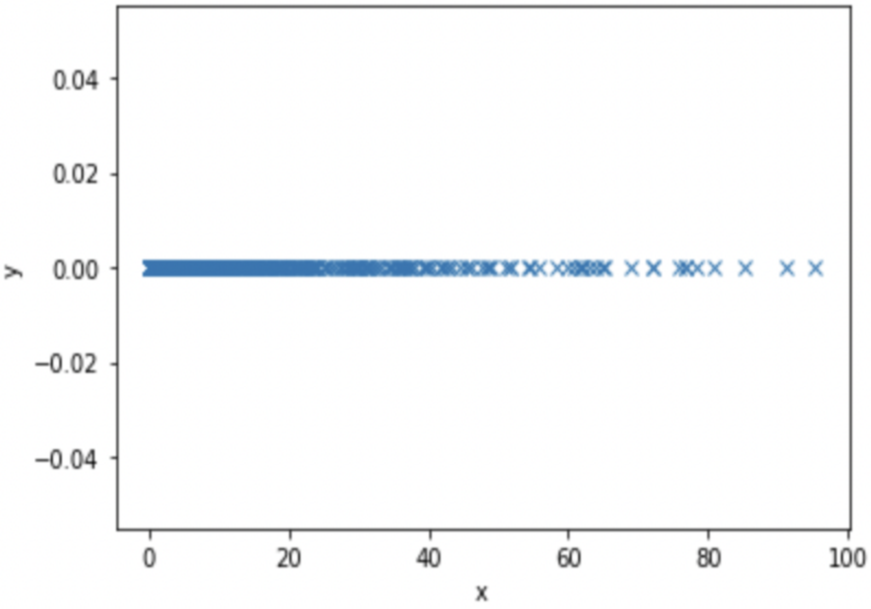
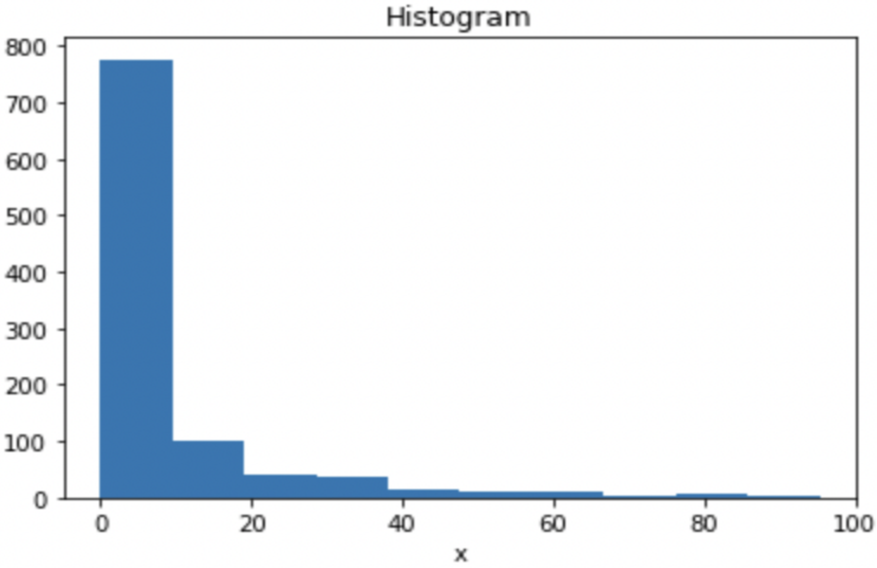
Question 1

(i)

The following image displays all the 1000 data points. It is clear from the image that most of the data points lie within (0, 20).



The histogram also suggests that most of the data points lie within the interval (0, 20).



If we sample a large number of datapoints from an exponential distribution, it is very likely that we will see a similar pattern. Hence I am assuming that the 1000 data points are generated from a mixture of exponentials.

Mixture of exponentials :-

Step1: Pick which mixture a datapoint comes from

Step2: Generate datapoint from that mixture.

parameters to estimate?

$\lambda_1, \lambda_2, \dots \lambda_K$ for the 'K' exponentials.

$\pi_1, \pi_2, \pi_3 \dots \pi_K$

Maximum Likelihood:-

$$L(\lambda_1, \lambda_2 \dots, \lambda_K, \pi_1, \pi_2, \dots \pi_K) = P(x_1, x_2 \dots x_n ; \lambda_1, \lambda_2 \dots, \lambda_K, \pi_1, \pi_2, \dots \pi_K)$$

$$= \prod_{i=1}^{n} P(x_i ; \lambda_1, \lambda_2, \dots \lambda_K, \pi_1, \pi_2, \dots \pi_K)$$

$$= \prod_{i=1}^{n} f_{mix}(x_i ; \lambda_1, \lambda_2, \dots \lambda_K, \pi_1, \pi_2, \dots \pi_K)$$

$$= \prod_{i=1}^{n} \left[\sum_{K=1}^{K} \pi_K \underbrace{f(x_i ; \lambda_K)}_{\text{exponential distribution.}} \right]$$

$$\underbrace{L(\theta)}_{\text{all parameters}} = \prod_{i=1}^{n} \left[\sum_{K=1}^{K} \pi_K \lambda_K e^{-\lambda_K x_i} \right]$$

$$\log L(\theta) = \sum_{i=1}^{n} \log \left(\sum_{K=1}^{K} \pi_K \lambda_K e^{-\lambda_K x_i} \right)$$

For every data point i, we have parameters $\{\gamma_1^i, \gamma_2^i, \dots, \gamma_K^i\}$ such that $\forall i \sum_{K=1}^{K} \gamma_K^i = 1$, $0 \leq \gamma_K^i \leq 1$ $\forall i, K$.

$$\log L(\theta) = \sum_{i=1}^{n} \log \left\{ \sum_{K=1}^{K} \gamma_K^i \left(\frac{\pi_K \lambda_K e^{-\lambda_K x_i}}{\gamma_K^i} \right) \right\}$$

By Jensen's inequality,

$$\log L(\theta) \geq \text{modified\_log} L(\theta, \gamma)$$

$$= \sum_{i=1}^{n} \sum_{K=1}^{K} \gamma_K^i \log \left(\frac{\pi_K \lambda_K e^{-\lambda_K x_i}}{\gamma_K^i} \right)$$

Fix γ and maximize over θ

$$\max_{\theta} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_K^i \log\left(\frac{\pi_K \lambda_K e^{-\lambda_K x_i}}{\gamma_K^i}\right)$$

$$\{\lambda_1, \lambda_2, \ldots \lambda_K, \; \pi_1, \pi_2, \ldots \pi_K\} \quad \sum_{i=1}^{n} \sum_{K=1}^{K} \gamma_K^i \left\{\log\left(\pi_K \lambda_K e^{-\lambda_K x_i}\right) - \log \gamma_K^i\right\}$$

$$\sum_{i=1}^{n} \sum_{K=1}^{K} \gamma_K^i \left\{\log \pi_K + \log \lambda_K + \log e^{-\lambda_K x_i} - \log \gamma_K^i\right\}$$

$$\sum_{i=1}^{n} \sum_{K=1}^{K} \left(\gamma_K^i \log \pi_K + \gamma_K^i \log \lambda_K - \gamma_K^i \lambda_K x_i - \underbrace{\gamma_K^i \log \gamma_K^i}_{\text{constant}}\right)$$

Taking derivative w.r.t λ, we get

$$\boxed{\frac{1}{\lambda_K} = \frac{\sum_{i=1}^{n} \gamma_K^i x_i}{\sum_{i=1}^{n} \gamma_K^i}}$$

similarly,

$$\boxed{\pi_K = \frac{\sum_{i=1}^{n} \gamma_K^i}{n}}$$

Thus when γ is fixed, we get

$$\frac{1}{\hat{\lambda}_K^{MML}} = \frac{\sum_{i=1}^{n} \gamma_K^i x_i}{\sum_{i=1}^{n} \gamma_K^i} \quad \text{—— (i)}$$

$$\hat{\pi}_K^{MML} = \frac{\sum_{i=1}^{n} \gamma_K^i}{n} \quad \text{—— (ii)}$$

Fix θ and maximize over γ

$$\sum_{i=1}^{n}\sum_{K=1}^{K} \gamma_K^i \log\left(\frac{\pi_K \lambda_K e^{-\lambda_K x_i}}{\gamma_K^i}\right)$$

$$= \sum_{i=1}^{n}\sum_{K=1}^{K} \gamma_K^i \log\left(\frac{a_{iK}}{\gamma_K^i}\right) \longrightarrow \text{constant equal to } \pi_K \lambda_K e^{-\lambda_K x_i}$$

$$= \sum_{i=1}^{n}\sum_{K=1}^{K} \left[\gamma_K^i \log a_{iK} - \gamma_K^i \log \gamma_K^i\right]$$

There is a set of γ_K^i's for every i, so can maximize separately for every i

For a fixed i, we want

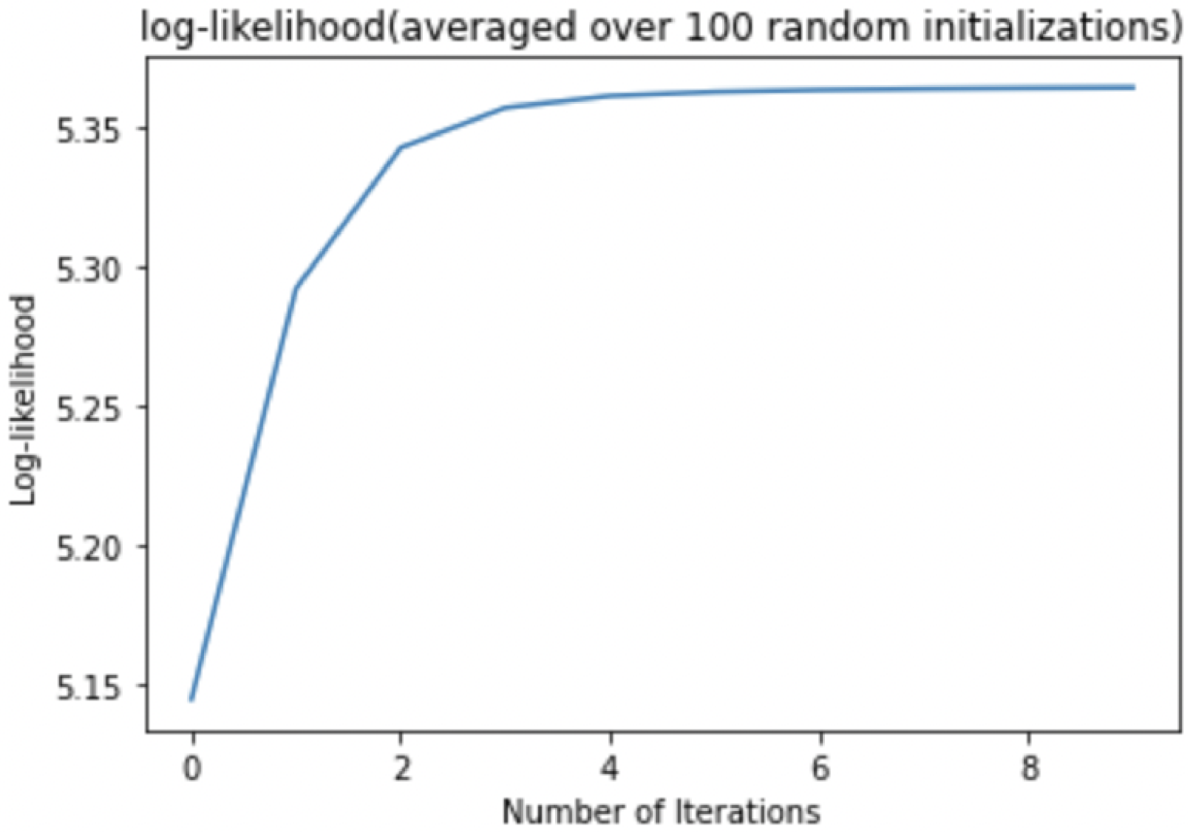
$$\max_{\gamma_1^i, \gamma_2^i, \dots \gamma_K^i} \sum_{K=1}^{K} \gamma_K^i \log(a_{iK}) - \gamma_K^i \log \gamma_K^i \quad \text{such that } \sum_{K=1}^{K} \gamma_K^i = 1$$

$$0 \leq \gamma_K^i \leq 1.$$

this has an analytical solution that can be obtained using method of Lagrange multipliers.

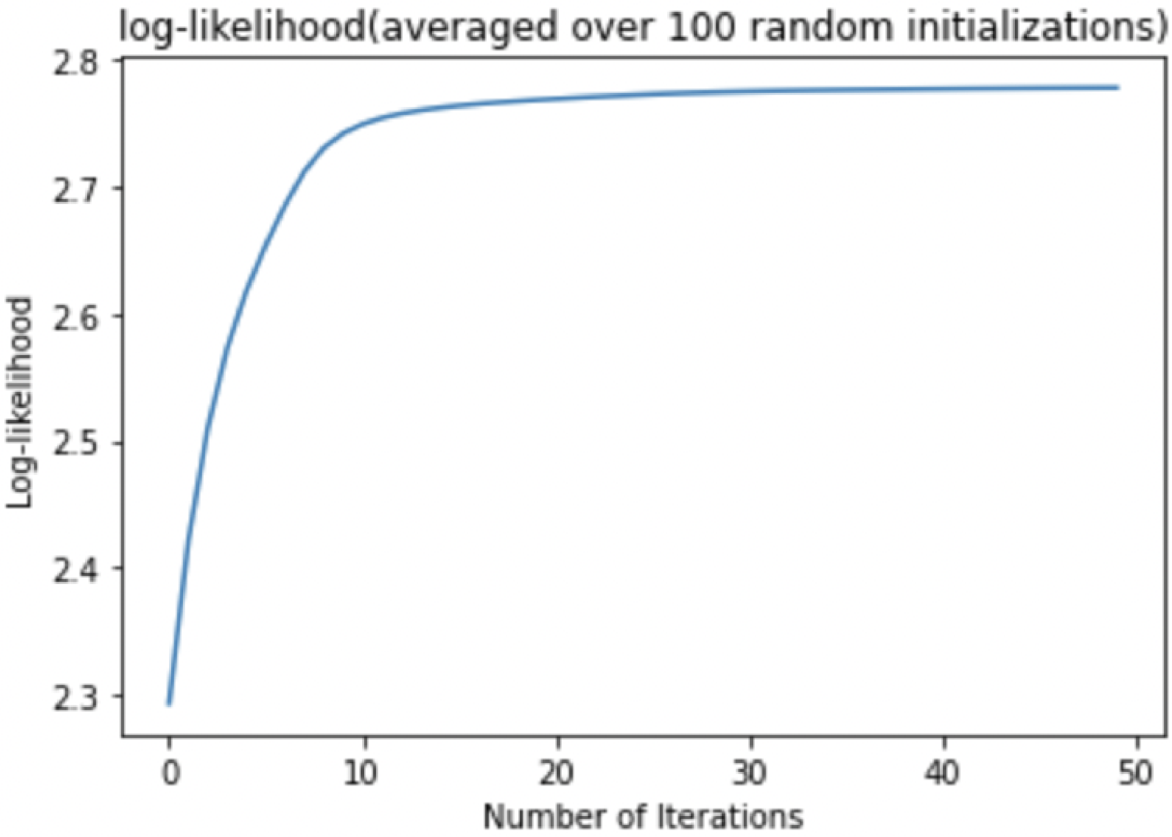
$$\hat{\gamma}_K^{i:\,MML} = \frac{\lambda_K e^{-\lambda_K x_i} \cdot \pi_K}{\sum_{l=1}^{K} \left(\lambda_l e^{-\lambda_l x_i} \cdot \pi_l\right)} \qquad \text{—— (III)}$$

Implemented the EM algorithm for mixture of exponentials by setting the number of mixtures K = 4. Please refer to the “Mixture of Exponentials.ipynb” file for implementation.



(ii)

Implemented the EM algorithm for mixture of gaussians by setting the number of mixtures K = 4. Please refer to the “Mixture of Gaussians.ipynb” file for implementation.



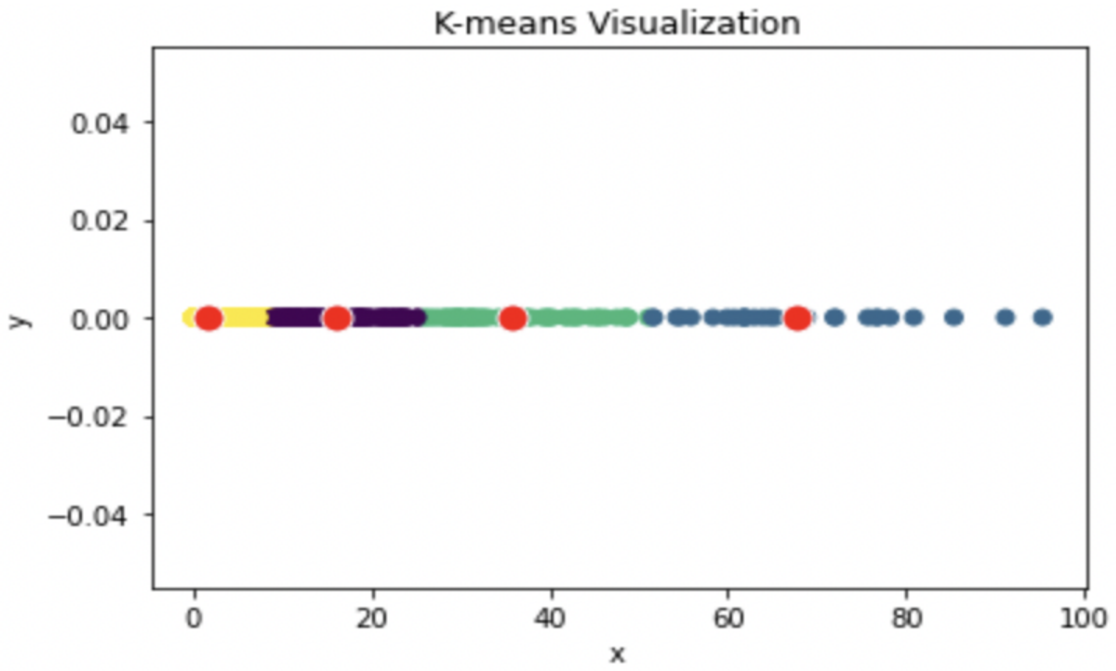
In part (i) I assumed that the data points were generated from a mixture of exponentials. I implemented the EM algorithm for a mixture of exponentials in order to estimate the parameters of the model. Our goal is to find the set of parameter values that maximize the log likelihood function. As we know after every iteration of the EM algorithm the value of the log likelihood function increases until we converge to a local maxima. The log-likelihood vs number of iterations plot for the part (i) suggests the same. For a mixture of exponentials the EM algorithm converged after 9 or 10 iterations. The average log likelihood value obtained was approximately 5.36.

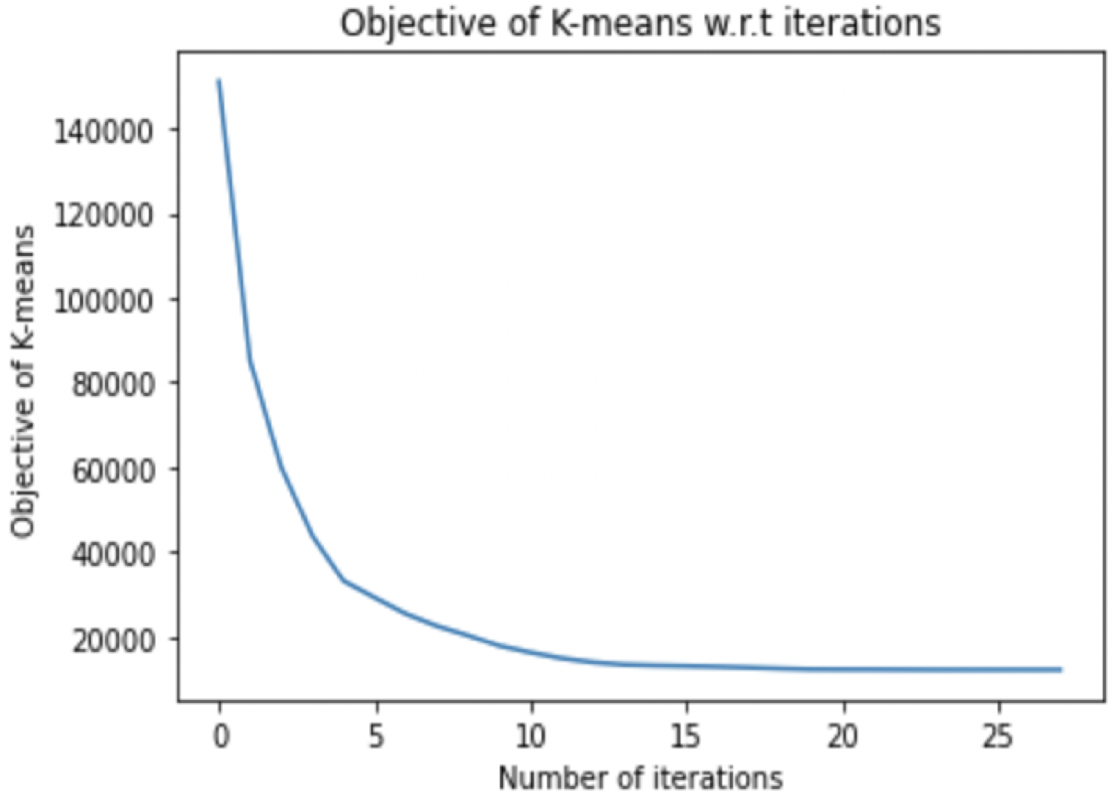
In part (ii) I implemented the EM algorithm for mixture of gaussians with the same goal of estimating the parameters. In this case also the log likelihood value increased after every iteration. The log likelihood vs number of iterations plot for part(ii) suggests the same as well. For a mixture of gaussians the EM algorithm took 50 iterations to converge. The average log likelihood value obtained was approximately 2.78.

The log-likelihood value of a probabilistic model is a way to measure the goodness of the model. The higher the value of the log-likelihood, the better a model describes the data. In this case if we assume the data points were generated using a mixture of exponentials then we have an average log-likelihood value of 5.36. On the other hand if we assume a mixture of gaussians as the underlying probabilistic model, then we have an average log-likelihood value of 2.78. So a mixture of exponentials is a better probabilistic model for describing the observed data points.

(iii)

Implemented the K-means algorithm with K = 4. Please refer to the “K Means.ipynb” file for implementation.





(iv)

In part (i) I assumed that the data points were generated from a mixture of four exponentials. I estimated the parameters of the model using the EM algorithm.

In part (ii) I considered a mixture of four gaussians as the data generation story and estimated the parameters using the EM algorithm.

And in part (iii) I implemented the K means algorithm with K = 4.

There are some fundamental differences between what the EM algorithm does and what the K means algorithm does. For every datapoint K means performs hard clustering in the sense that it assigns a datapoint to a particular cluster with certainty. On the other hand the EM algorithm produces a softer version of clustering. EM gives us a probability distribution for every datapoint, which allows us to know what is the probability that a particular data point belongs to a particular cluster.

K means only considers the means in the sense that a mean describes a cluster. In EM mean and variance describes a cluster/mixture. When the clusters are mixed(which is the case in this problem), we can use the variance information to separate one cluster from another. EM has more parameters than K means, hence EM allows us to understand the data better. EM has more directional information to separate out the clusters as opposed to K means.

For the above reasons we could say that EM is a better choice for this particular problem.

We implemented the EM algorithm for two mixture models - Exponential Mixture and Gaussian Mixture. The log-likelihood value of a probabilistic model is a way to measure the goodness of the model. The higher the value of the log-likelihood, the better a model describes the data. In this case if we assume the data points were generated using a mixture of exponentials then we have an average log-likelihood value of 5.36. On the other hand if we assume a mixture of gaussians as the underlying probabilistic model, then we have an average log-likelihood value of 2.78. So a mixture of exponentials is a better probabilistic model for describing the observed data points.

Hence for this dataset, I would choose a mixture of exponentials as the underlying probabilistic model and I will run the EM algorithm in order to estimate the parameters of the model.