Notes

Bob Effinger, Miru Park, Sushil Kumar

July 28, 2019

1 Abstract

Gaussian Mixture Models (GMM) are a common approach used for unsupervised learning tasks. We will introduce the concepts that are required to use and understand a GMM for simple clustering tasks. GMMs can be used for many different fields, but in recent years they have been used for machine learning tasks such as feature extraction for speech data. In addition GMM improves on some areas which our previously learned method, K-means, does not perform well.

Learning Objectives:

(1) Be able to understand and use basic probability and Gaussian Distribution.

(2) Be able to pick out scenarios in which GMM outperform K-means.

(3) Be able to implement GMM using the EM algorithm and analyze results.

2 Data, Probability, Gaussian Distribution

2.1 Data and Probability

Definition 1: Samples or data are observations from a trial or an experiment. Sample value is the value associated with the sample. This is very much analogous to the term instances and samples we have been seeing throughout the course. For example, in our previous example, a sample was a single gummy bear and our data was a collection of those 10 gummy bears.

Definition 2: Supposed we have a set of samples or data X = $\{x_1, x_2, x_3, ..., x_N\}$Sample mean is the average value of our samples, which we denote as μ and μ = 1/N $\times \sum_{i=1}^{N} x_i$, where N is the number of our samples and x_i is the value of ith sample.

Definition 3: The expectation of some unobserved sample or an instance X is E[X] = $\sum_{i \in Range(X)} P(X = x_i) \times x_i$, where Range(X) is the set of all possible values our sample can take.

Remark: For now, let's restrict our definition to only the sample values we have observed. In general, to compute the expectation value of some unobserved sample X or in general our data, we need to know its distribution. We will introduce what it means for X to follow some probability distribution D, denoted as X $\sim$ D in the subsequent sections.

Example: Let's consider some data we obtained from surveying some portion of population; we recorded how old random freshmen students are. Suppose we interviewed five freshmen students. Let X denote the age of a random freshman and suppose we have the following observations: $x_1 = 17$, $x_2 = 17$, $x_3 = 18$, $x_4 = 18$, and $x_5 = 19$. Then, we have E[X] $= 0.2 \times 17 + 0.5 \times 18 + 0.3 \times 19 = 18.1$. From this data, we can infer that a random freshman is 18.1 years old.

Remark: It is important to distinguish between Definition 2 and Definition 3. The former gives us information about what we observed and the latter tells us what to expect given some observations.
We end the first section with two more important definitions.

Definition 4: Sample variance S^2 measures the squared distance of samples from the sample mean. S^2 = 1/(n-1) $\times \sum_{i=1}^{n} (x_i - \mu)^2$, where n is the number of samples.

Definition 5: Variance σ^2 measures expected squared distance from the expected value E[X] and is denoted as σ^2 = E[$(x - E[x])^2$].

2.2 Introduction to Gaussian Distribution

A crucial question one might ask is: what are the likelihoods of samples of our data or how are samples in our data distributed? In order to answer questions of this nature, we need to know how what probability distribution our data follows. Once we have some information about how our data is distributed, we denote X $\sim$ D(Ω) and say X follows some probability distribution D, where Ω is the set of parameters that characterize the distribution.

Definition 1: Probability density function (PDF), Pr(x) is the probability that an unobserved sample is near x. Pr(x) = $\frac{1}{\sigma\sqrt{2\pi}}$exp($\frac{-(x-\mu)^2}{2\sigma^2}$) and μ = E[X]. Often, we denote $Pr(x|\mu,\sigma^2)$ and say probability of x is parameterized by μ and σ.

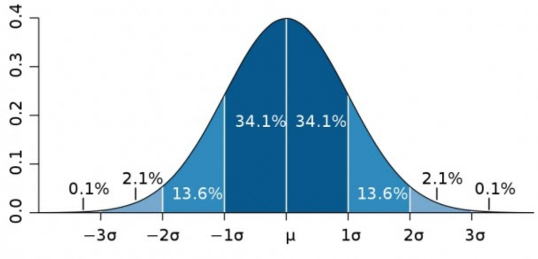
Last but not least, under our new framework of continuous data and distribution, we need to revise our definitions of expected value E[X] and variance σ^2.

Definition 2: If our data X follows a Gaussian distribution, we denote it as X $\sim$ N(μ, σ^2) and say that X is normally distributed.

Definition 3: E[X] = $\int_{-\infty}^{\infty} xPr(x)dx = \mu$.

Definition 4: σ^2 = E[(x - E[X])^2] = $\int_{-\infty}^{\infty}(x - E[X])^2 Pr(x)dx$.

Following figure illustrates the PDF of our Gaussian Distribution.

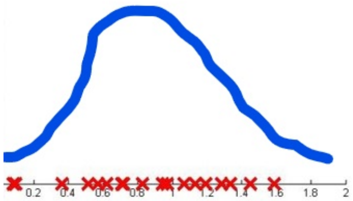


2.3 Data, Distribution, Parameter Estimation, and Maximum Likelihood Estimation

As was shown in the last section, in order for us to compute the PDF, we need to know the probability distribution. Most of the time, when we have data, we face two challenges. First challenge is that more often than not, we do not what probability distribution our data follows i.e X $\sim$ D and D is unknown. Second challenge is that even if we know or can guess D, we actually do not know its parameters. For example, if we have a good reason to believe that our data X, X $\sim$ N(μ, σ^2), we still do not know the true μ and σ^2.

Thus part of our task is to first guess what kind of distribution our data follows and estimate its parameters. The Details of parameter estimation is beyond the scope of this class. But here is the intuition.

Take a look at this data, X = $\{x_i, ..., x_n\}$ for $x_i \in \mathbb{R}$.
We can see most of the data points are clustered in the middle - around the mean - and other further out towards the boundary of the line. Then it is reasonable to assume this data is normally distributed (illustrated as the blue curve representing density of data accumulation).



Now, our second task is to guess their parameters, mean and variance. To do this parameter estimation, we use a technique from statistics called Maximum Likelihood Estimation, otherwise known as MLE. Derivation and understanding of this concept requires some background knowledge in statistics and probability theory.

The big picture is to find the maximum likelihood parameters μ and σ that will minimize the dissimilarity between our true distribution $Pr(X|\mu, \sigma^2)$(our training data's distribution) and newly estimated distribution $Pr(X|\hat{\mu}, \hat{\sigma^2})$.

It turns out that the following parameters are the maximum likelihood parameters for our data that follows the Gaussian distribution.
(i) $\hat{\mu}$ = 1/N × $\sum_{i=1}^{N} x_i$
(ii) $\hat{\sigma^2}$ = 1/N × $\sum_{i=1}^{N} (x_i - \hat{\mu})^2$.

From here and on, any families of probability distribution we are fitting our data to are assumed to be Gaussian.

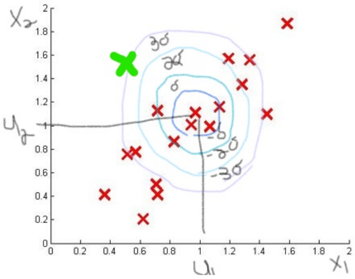
2.4 Multivariate Gaussian Distribution

So far we have considered only univariate Gaussian distribution. In other words, given some data X = $\{x_1, ..., x_n\}$, we assumed each sample $x_i \in \mathbb{R}$. However, as we have seen throughout the course, that is hardly ever the case; each instance usually had multiple features. Thus now let us assume $\underline{x_i} \in \mathbb{R}^{\ltimes}$. For simplicity, you can imagine that n = 2. However, What is about to be presented holds for all finite $n > 2$. Let's see an example where we might want to resort to multivariate Gaussian to model our data.

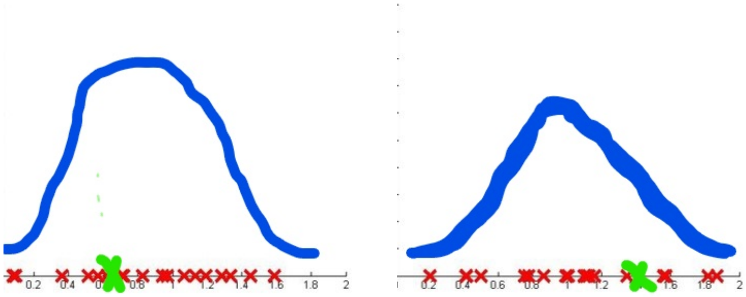
Suppose that we want to observe CPU load and memory use of some machine for anomlay detection. Let x_1^i be i^{th} machine's cpu load and let x_2^i be i^{th} machine's memory use.

Consider the following figure (scatter graph) that represents relationship between the two features.

Horizontal axis represents x_1 and vertical axis represents x_2.



We can see it has some linear relationship. Now consider this new point I just drew in Green (call it $\underline{y}$). You would like to think that specific machine with the corresponding features might be acting in an anomalous manner, but lets see what happens if we decide to project that co-ordinate onto 1-d space for x_1's density and for x_2's density.



We can see that for both projections, the machine's behaviour isn't too "anomalous"; both projections lie reasonably within 1.5 standard deviations away from the densely populated region.

Then for cpu load, we have $Pr(x_1|\mu_1, \sigma_1^2)$ and for memory use we have $Pr(x_2; \mu_2, \sigma_2^2)$.

If we computed $Pr(y_1|\mu_1, \sigma_1^2)$ and $Pr(y_2|\mu_2, \sigma_2^2)$, densities of each individual feature of y will not be high enough to be detected for anomaly. So instead of modeling and inspecting $Pr(x_1^i)$ and $Pr(x_2^i)$ individually, we model and inspect $Pr(\underline{x^i})$ where $\underline{x^i} = \begin{bmatrix} x_1^i & x_2^i \end{bmatrix}^T$. Now our sample lives in a two dimensional space, not one. Thus we now introduce a multivariate Gaussian distribution and move away from univariate case.

We now introduce our new pdf for multivaraite Gaussian.

Definition 1: Let x $\in \mathbb{R}^n$ and let $\underline{\mu} \in \mathbb{R}^n$. Then probability density function is $Pr(x|\underline{\mu}, \underline{\Sigma}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\underline{\Sigma}|}} \exp(\frac{1}{2}(\underline{x} - \underline{\mu})^T \underline{\Sigma}^{-1}(\underline{x} - \underline{\mu}))$, where $\underline{\Sigma}$ (n x n) is the covariance matrix and $|\underline{\Sigma}|$ is the determinant of the covariance matrix.

Intuitively, covariance measures the variability of two features x_1 and x_2. Thus for example, let us measure age(a) and height(h) for children (what are the features and instance here ? activity). It is highly possible results will be strongly correlated; older children will be taller and heavier. So, lets assume for a second we actually have all the parameters: μ_a, μ_h, μ_w, σ_a, σ_h, and σ_w. Then the covariance of age and height - cov(a,h) - measures the connection of age and height.In our case, covariance of these two feautres is computed as: E[(age - mean age)(height - mean height)].

Thus, for Bivaraite Guassian case, we will have a 2 X 2 covariance matrix.

We don't introduce the definition nor the construction of covariance matrix to keep things simple. You just need to know what it is and what role it plays, moving forward into the section.
Now let's take a look at some examples of how our Gaussian density function changes with different covariance matrices.

The left most pictures are the contours, then in the middle we have the surface of the Guassian density, and the right most pictures are the parameters.

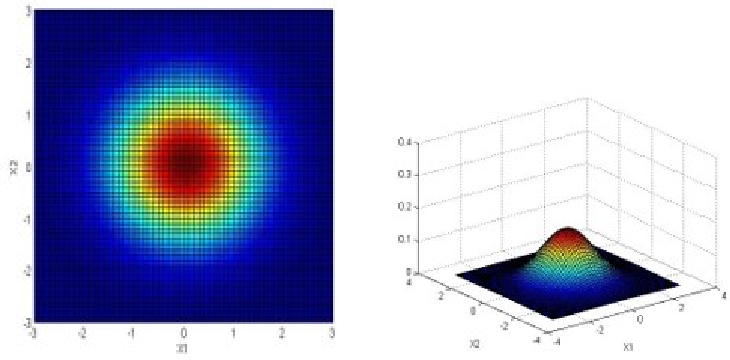
We now end the section with showing how to compute maximum likelihood parameters for multivariate Gaussian case.

In the multivariate case, we compute the parameters for $\underline{\mu}$ and$\underline{\Sigma}$. Formulae are very similar to the univariate case.

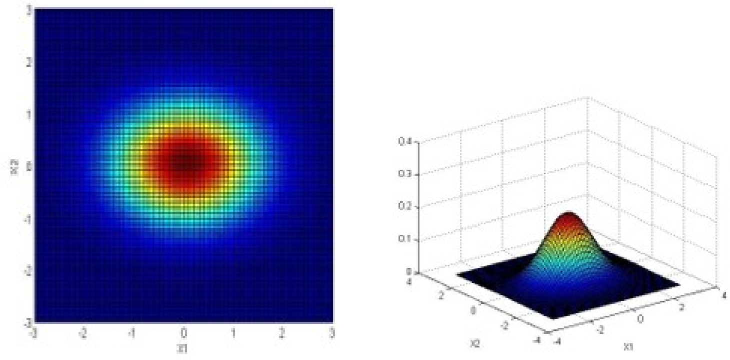
Let X = $\{\underline{x^i}, ..., \underline{x^n}\}$ be our training data set i.e. we have n many samples. Then the following are the MLE parameters.

(i) $\underline{\hat{\mu}}$ = 1/n $\times \sum_{i=1}^{n} \underline{x^i}$.

(ii) $\underline{\hat{\Sigma}}$ = 1/n $\times \sum_{i=1}^{n}(\underline{x^i} - \underline{\hat{\mu}})(\underline{x^i} - \underline{\hat{\mu}})^T$



$$\mu = \begin{bmatrix}0\\0\end{bmatrix} \Sigma = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$$



$$\mu = \begin{bmatrix}0\\0\end{bmatrix} \Sigma = \begin{bmatrix}1 & 0\\0 & 0.6\end{bmatrix}$$

3 Overview of GMM

GMMs come from taking the idea of unsupervised methods and combining it with the statistical methodology involving Gaussian Distribution. It is the idea that instead of viewing a data set as a single population, we want to view the data as multiple separate populations. And by doing so we need to determine the mean and the standard deviation of all the separate gaussian distributions that make up the population. By determining these separate distributions we can estimate the probability that future data belongs to the individual curves.

This Approach was first seen in the “Estimating the Components of a Mixture of Normal Distributions”. As we have seen with K-means algorithm, given N many observations(data points), the goal is to group these N many observations(data points) into K number of clusters. Although K-means has many advantages, unfortunately, it does not perform very well when the geometry of the clusters is non-circular for a 2 dimensional case, or in general a non-isometric case. In addition K-means algorithm fails to generalize well in cases where clusters overlap.

GMMs are also different from K-means in the way that they classify each data point. The GMM model gives each feature vector a probability of belonging to the different components of the model. This can help for tasks in which knowing the likelihood of a given data point belonging to a specific cluster is more important than a simple grouping.

Given data $\underline{\text{X}}$, like K-means, we begin with the assumption that $\underline{\text{X}}$ has K Clusters, or K labels. In the case of GMM we believe these K labels each belong to a different Gaussian. In order to approximate the overall distribution we thusly need to determine the K Gaussians that make up the overall distributions. Or to say it in another way we are looking for the parameters to fit

$Pr(\underline{X}|\underline{\mu}, \underline{\Sigma}, \pi) = \sum_{k=1}^{K} \pi_k Pr(\underline{X}|\underline{\mu}_k, \underline{\Sigma}_k)$.

The main Point here is:

$Pr(\underline{x_i} \in C_j) = \pi_j \times \frac{\alpha}{\beta}$, where $\alpha = Pr(\underline{x_i}|\underline{\mu_j}, \underline{\Sigma_j})$, $\beta = \sum_{k=1}^{K} \pi_k Pr(\underline{x_k}|\underline{\mu}_k, \underline{\Sigma}_k)$ and $Pr(\underline{x_i} \in C_j)$ is the probability that $\underline{x_i}$ is in cluster j.

It is important to note the differences between hard and soft classification analysis. Hard classification analysis, simply tells that a data point belongs to a particular cluster (this is what K-Means algorithm provides). But, soft classification provides probability of a data point belonging to any cluster. The Gaussian Mixture Model, presented in this document, provides the tools to perform both hard and soft classification analysis.

We still need to determine the value that should be used for π_k and the approach we will be using will be the EM approach.

4 Expectation Maximization Algorithm

Expectation Maximization (EM) algorithm is an iterative optimization technique and contains two steps: Expectation and Maximization. Given a gaussian mixture model, the goal is to maximize the log likelihood function, given by:

$$\ln\left\{\Pr(\underline{X}|\underline{\mu}, \underline{\Sigma}, \pi)\right\} = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \Pr(\underline{x_n}|\underline{\mu_k}, \underline{\Sigma_k})\right\}$$

with respect to the parameters comprising the means, covariances, and mixing coefficients of each distribution. $\underline{X}$ corresponds to the data set, $\underline{\mu}$ is the mean, $\underline{\Sigma}$ is the variance, π is the weight, N is the total number of data, K is the total number of clusters (aka, the total number of gaussian distributions), $\underline{x}_n$ is the n-th data in data set, π_k is the mixing coefficient associated with distribution k, $\underline{\mu_k}$ is the mean associated with distribution k, and $\underline{\Sigma_k}$ is the covariance associated with distribution k. It is important to note that the mixing coefficient, π_k, is constrained by

$$0 \leq k \leq 1$$

and

$$\sum_{k=1}^{K} \pi_k = 1$$

.

The first step is to initialize the means ($\underline{\mu_k}$), covariances ($\underline{\Sigma_k}$), and mixing coefficients π_k for each gaussian distribution k. The second step is to evaluate the responsibilities (defined by latent variable γ_k) using the current set of parameter values:

$$\gamma_k(\underline{X}) = \frac{\pi_k \Pr(\underline{X}|\underline{\mu_k}, \underline{\Sigma_k})}{\sum_{j=1}^{K} \pi_j \Pr(\underline{X}|\underline{\mu_j}, \underline{\Sigma_j})}$$

The third step is to evaluate the log likelihood of the given function for the data set. The fourth step involves re-estimating the parameters for each distribution using the current responsibilities:

$$\underline{\mu_k} = \frac{\sum_{n=1}^{N} \gamma_k(\underline{x_n})\underline{x_n}}{\sum_{n=1}^{N} \gamma_k(\underline{x_n})}$$

$$\underline{\Sigma_k} = \frac{\sum_{n=1}^{N} \gamma_k(\underline{x_n})(\underline{x_n} - \underline{\mu_k})(\underline{x_n} - \underline{\mu_k})^T}{\sum_{n=1}^{N} \gamma_k(\underline{x_n})}$$

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_k(\underline{x_n})$$

The final step involves checking if the log likelihood function has satisfied some convergence criteria (which checks if there is negligible variation in the log likelihood value between iterations). If it does not converge, then repeat steps two, three, and four until convergence is achieved.

For a univariate distribution (or data set with one feature), the probability distribution function is simply:

$$\Pr(x_n|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x_n-\mu_k)^2}{2\sigma_k^2}}$$

where, σ_k is the standard deviation of the k-th gaussian distribution. It is worth noting that the data point x_n, distribution mean μ_k, and distribution variance σ_k simplify to a scalar variable for the univariate case. They are computed during each iteration using the following formulas:

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma_k(x_n)x_n}{\sum_{n=1}^{N} \gamma_k(x_n)}$$

$$\sigma_k = \left(\frac{\sum_{n=1}^{N} \gamma_k(x_n)(x_n - \mu_k)^2}{\sum_{n=1}^{N} \gamma_k(x_n)} \right)^{-1/2}$$

For a multivariate distribution (or data set with multiple features), the probability distribution function is:

$$\Pr(\underline{x_n}|\underline{\mu_k}, \underline{\Sigma_k}) = \frac{1}{\sqrt{2\pi|\underline{\Sigma_k}|}} e^{-\frac{(\underline{x_n}-\underline{\mu_k})^T \underline{\Sigma}^{-1} (\underline{x_n}-\underline{\mu_k})}{2}}$$

5 Example of Expectation Maximization Algorithm

Suppose, we have a randomly generated data, as shown in Figure 1, with approximately three clusters. The contours correspond to the areas where probability distribution function is constant for the respective cluster. Figure 1 shows the initialization of 3 different gaussian distributions at an arbitrary point. We see that in Figure 2, after two iterations of EM algorithm, the gaussian distributions are slowly moving into place and finding appropriate cluster centers. After five iterations of EM algorithm, in Figure 3, we see that the gaussian distributions have more or less found the cluster centers with the variance still being adjusted. After 20 iterations, in Figure 4, it is visually clear the algorithm has converged with the mean and variance of each distribution appropriately capturing the space of each cluster. This is corroborated by the convergence of log likelihood function, as seen in Figure 5.

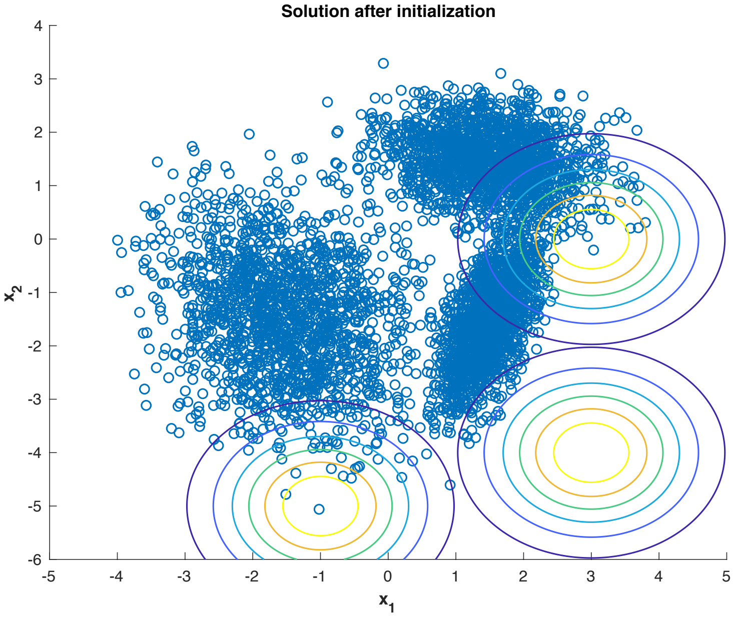


Figure 1: Solution after initialization of EM algorithm

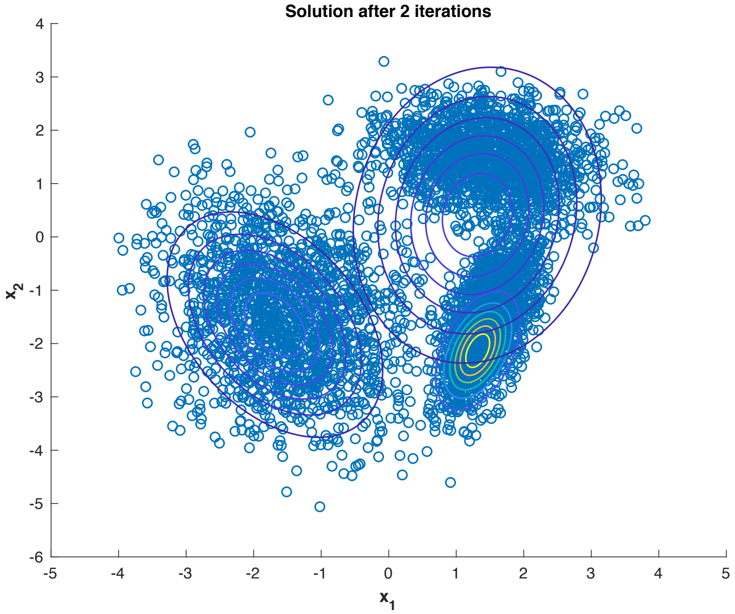


Figure 2: Solution of expectation maximization algorithm after two iterations

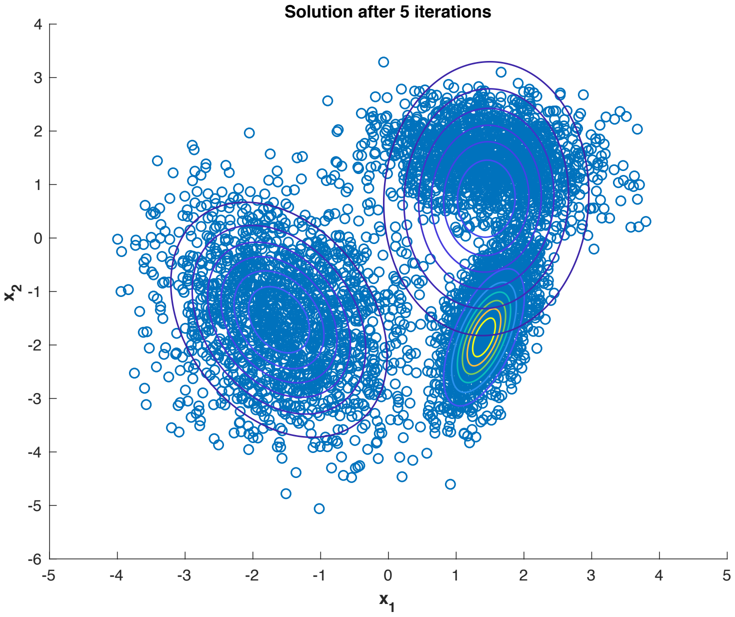


Figure 3: Solution of expectation maximization algorithm after five iterations

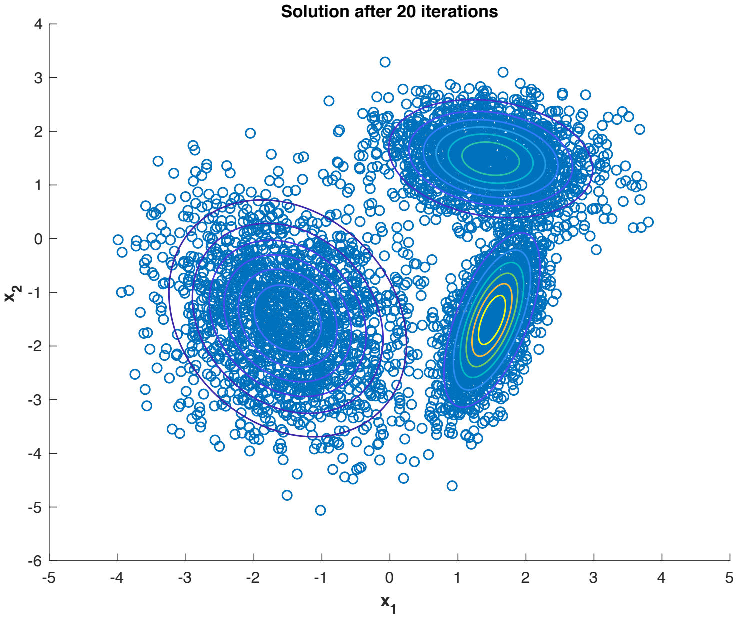


Figure 4: Solution of expectation maximization algorithm after 20 iterations

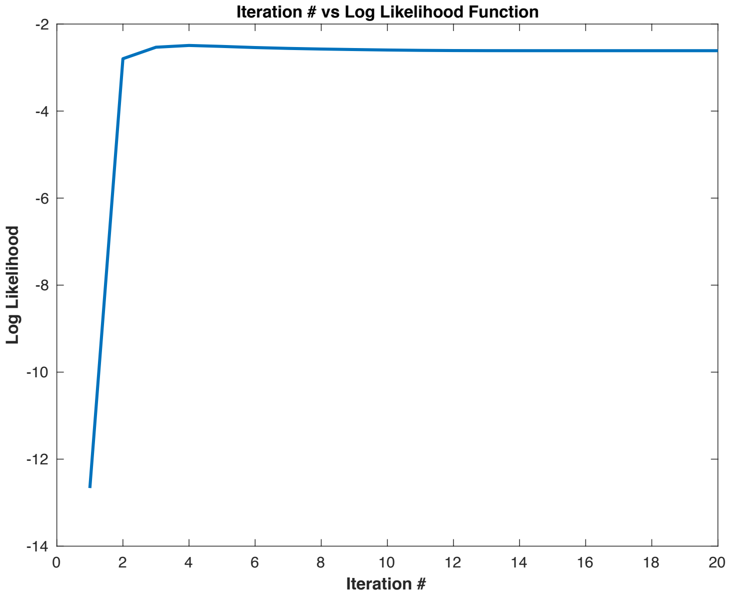


Figure 5: Log-Likelihood value as a function of Iteration #