Gaussian Mixture Model

Damaris Stephanie, Lionel Komen, and Omer Elhussien

April 10, 2023

Abstract

Gaussian Mixture Model is an unsupervised probabilistic model category that states all the generated data points are derived from a mixture of finite Gaussian distributions with unknown parameters. It's usually helpful to model more complex distributions; In this report, we present the ideas behind a Gaussian mixture model, the assumptions made when, and we implement EM algorithm from the initialization step (by using different initialization methods) to the maximization step.

Introduction

Revealing hidden structures is one of the main goals of the clustering scheme. Two main categories are known in clustering: one that relies on similarity and dissimilarity distances, and the other classifies data into groups. The latter relies on a model-based approach assuming that the current population is a mixture of several sub-populations. The Gaussian mixture model is one member of the latter class.

Its ability to reflect a massive class of sample distributions, mixed with smooth approximations to arbitrarily shaped densities, could be the main factor for the wide use of a Gaussian mixture model. It is a parametric density function expressed as a sum of several Gaussian densities with given weights [1][2][3].

In the following sections, we first state the assumptions made when introducing Gaussian Mixtures Models, after which we talk about the EM algorithm and how to implement it.

Model assumptions

Let us suppose we are given a training set $\{x^{(1)}, \ldots, x^{(n)}\}$, which are *i.i.d*. We are interested in fitting the data using the below model,

$$\mathbb{P}(x^{(i)}, z^{(i)}) = \mathbb{P}(x^{(i)}|z^{(i)})\mathbb{P}(z^{(i)}), \tag{1}$$

where:

- $z^{(i)} \sim multinomial(\phi)$, $\phi_j \geq 0$, $\sum_{j=1}^{k} \phi_j = 1$, and $\phi_j = \mathbb{P}(z^{(i)} = j)$.
- $x^{(i)}|z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$.

In the above model, we assume that each $x^{(i)}$ was generated by randomly choosing $z^{(i)} \in \{1, \ldots, k\}$, then $x^{(i)}$ was drawn from one of k Gaussian depending on $z^{(i)}$. The above model is called Gaussian Mixture Model. Further, $z^{(i)}$'s are called hidden (latent) variables [4].

The parameters of the above model are $\phi, \mu,$ and Σ. The likelihood of the data is given as,

$$\mathcal{L}(\phi, \mu, \Sigma) = \prod_{i=1}^{n} \mathbb{P}(x^{(i)}; \phi, \mu, \Sigma) \iff$$

$$\log\left[\mathcal{L}(\phi, \mu, \Sigma)\right] = \sum_{i=1}^{n} \log\left[\mathbb{P}(x^{(i)}; \phi, \mu, \Sigma)\right]$$

$$\log \mathcal{L}(\phi, \mu, \Sigma) = \sum_{i=1}^{n} log \sum_{z^{(i)}=1}^{k} \mathbb{P}(x^{(i)}|z^{(i)}; \mu, \Sigma)\ \mathbb{P}(z^{(i)}; \phi). \tag{2}$$

The above likelihood on equation 2 could be easily solved with $z^{(i)}$'s were known. The likelihood will be,

$$\log\left[\mathcal{L}\right] = \sum_{i=1}^{n} log\mathbb{P}(x^{(i)}|z^{(i)}; \phi, \mu, \Sigma) + log\mathbb{P}(z^{(i)}; \phi). \tag{3}$$

our estimated parameters are:

- $\phi_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{z^{(i)} = j\}$,
- $\mu_j = \frac{\sum_{i=1}^{n} \mathbb{I}\{z^{(i)}=j\}x^{(i)}}{\sum_{i=1}^{n} \mathbb{I}\{z^{(i)}=j\}}$,
- $\Sigma_j = \frac{\sum_{i=1}^{n} \mathbb{I}\{z^{(i)}=j\}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_{i=1}^{n} \mathbb{I}\{z^{(i)}=j\}}$.

Expectation Maximization Algorithm

In the current problem, our $z^{(i)}$'s are unknown. We need to introduce the EM algorithm to solve the problem at hand.

EM is an iterative algorithm that has two main steps [5]:

- E-step: It tries to guess the values of $z^{(i)}$'s.
- M-step: It updates the parameters of the model based on our guesses.

In the E-step, we calculate
$\mathbb{P}(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma) =$

$$\frac{\mathbb{P}(x^{(i)}|z^{(i)} = j; \mu, \Sigma)\ \mathbb{P}(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} \mathbb{P}(x^{(i)}|z^{(i)} = l; \mu, \Sigma)\ \mathbb{P}(z^{(i)} = l; \phi)}.$$

Algorithm 1: EM algorithm

repeat until convergence:

{

(E-step) for each $i, j,$ set:

$$w_j^{(i)} = \mathbb{P}(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma).$$

(M-step) update the parameters:

$$\phi_j = \frac{\sum_{i=1}^{n} w_j^{(i)}}{n},$$

$$\mu_j = \frac{\sum_{i=1}^{n} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{n} w_j^{(i)}},$$

$$\Sigma_j = \frac{\sum_{i=1}^{n} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} w_j^{(i)}}.$$

}

Initialization Methods

There are several initialization methods which are [6] :

1- Random starting values: In this framework, each sample is assigned randomly to one of the k clusters while considering the same number of samples in each cluster. Then, the mean vectors and covariance matrices are estimated from these groups to start the first E-step of the EM algorithm. Several permutations should be implemented to ensure a complete examination of the likelihood function.

2- Iteratively Constrained EM: This approach assumes that the quickest parameters increase the likelihood will provide a better solution than those that do not. It starts by running several iteratively constrained EM. Then, the best parameters will be used as starting values for EM until convergence. The current approach is challenged by stating the number of iterations in the initial-stage EM since the given value could work with some likelihood functions but not all of them. Furthermore, the number of initial stages needs to be determined.

3- K-means clustering: The clusters with the lowest sum square error are used as starting points for the M-step. These clusters are reached after hundred times of computations and comparisons. This approach is challenged by the assumptions that K-means pose on the shape of the clusters. Furthermore, highly heterogeneous and nonspherical clusters will be poorly fitted. Finally, it is susceptible to locally optimal solutions.

4- Agglomerative hierarchical clustering: Hierarchical clustering is implemented on the data. Then, the result is partitioned into several clusters. These partitions are used to estimate the parameters of the first E-step. This approach will work with hierarchically ordered data. However, it will offer a single solution.

5- Sum scores: Each example's sum score is found. Second, these scores are used to order the data. Third, the ordered data are allocated into k groups. These groups are used to estimate the parameters of the first E-step. Several challenges are linked with the sum score: the single solution option, and it relies on the data.

Jensen's Inequality

Let f be a convex function, and let X be a random variable. Then,

$$\mathbb{E}\left[f(X)\right] \geq f\left[\mathbb{E}X\right].$$

Moreover, if f is strictly convex, then

$$\mathbb{E}\left[f(X)\right] = f\left[\mathbb{E}X\right] \quad iff \; X = \mathbb{E}(X) \; with \; probability \; 1.$$

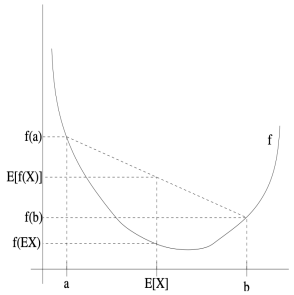

Figure 1: *The intuition behind Jensen's Inequality*

The evidence lower bound

Let $\{x^{(1)}, \ldots, x^{(n)}\}$ be a training set which are *i.id*. Let $\mathbb{P}(x, z; \theta)$ be our model with z being the latent variable. Then,

$$\mathbb{P}(x; \theta) = \sum_{z} \mathbb{P}(x, z; \theta). \tag{4}$$

The likelihood is given by,

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \mathbb{P}(x^{(i)}; \theta) \iff log\left[\mathcal{L}\right] = \sum_{i=1}^{n} log\left[\mathbb{P}(x^{(i)}; \theta)\right]$$

$$log\left[\mathcal{L}\right] = \sum_{i=1}^{n} log\left[\sum_{z^{(i)}} \mathbb{P}(x^{(i)}, z^{(i)}; \theta)\right].$$

The EM-algorithm maximizes $\mathcal{L}$ using two steps:

E-step: construct a lower bound on $\mathcal{L}$.
M-step: Optimize that lower-bound.

We will consider our optimization using a single data point, then after we bring the summation back.

$$\log \mathbb{P}(x;\theta) = log\left[\sum_z \mathbb{P}(x,z;\theta)\right]. \tag{5}$$

Let the possible values of z follow a distribution $\mathbb{Q}$, $\sum_z Q(z) = 1$, and $Q(z) \geq 0$.

$$\begin{aligned} log\mathbb{P}(x;\theta) = \\ &= log\left[\sum_z \mathbb{P}(x,z;\theta)\right] \\ &= log\left[\sum_z \mathbb{Q}(z)\frac{\mathbb{P}(x,z;\theta)}{\mathbb{Q}(z)}\right]. \end{aligned}$$

Using Jensen's inequality, we get

$$\geq \sum_z \mathbb{Q}(z) log\left[\frac{\mathbb{P}(x,z;\theta)}{\mathbb{Q}(z)}\right].$$

So for any distribution $\mathbb{Q}$, the above formula gives a lower-bound on $log\mathbb{P}(x;\theta)$.

So, we know that

$$\mathbb{Q}(z) = \frac{\mathbb{P}(x,z;\theta)}{\sum_z \mathbb{P}(x,z;\theta)} = \frac{\mathbb{P}(x,z;\theta)}{\mathbb{P}(x;\theta)} = \mathbb{P}(z|x;\theta). \tag{6}$$

With the value of $\mathbb{Q}(z)$ in the equation 6 we get,

$$\sum_z \mathbb{Q}(z) log\left[\frac{\mathbb{P}(x,z;\theta)}{\mathbb{Q}(z)}\right] =$$

$$\sum_z \mathbb{P}(z|x;\theta) log\frac{\mathbb{P}(z|x;\theta)\mathbb{P}(x;\theta)}{\mathbb{P}(z|x;\theta)} = log\left[\mathbb{P}(x;\theta)\right].$$

The evidence lower bound is given by:

$$ELBO(x;\mathbb{Q},\theta) = \sum_z \mathbb{Q}(z) log\left[\frac{\mathbb{P}(x,z;\theta)}{\mathbb{Q}(z)}\right]. \tag{7}$$

So,

$$\forall \mathbb{Q}, \theta, x, \quad log\mathbb{P}(x;\theta) \geq ELBO(x;\mathbb{Q},\theta).$$

EM algorithm updates $\mathbb{Q}$ and θ by:

a. setting $\mathbb{Q}(z) = \mathbb{P}(z|x;\theta)$.
b. maximizing $ELBO(x;\mathbb{Q},\theta)$ w.r.t θ while fixing the choice of $\mathbb{Q}$.

Other forms of ELBO

The ELBO given by equation 7 has many others forms. The first one is showing as follows:

$$\begin{aligned} ELBO(x;Q,\theta) &= E_{z\sim Q}[\log p(x,z;\theta)] - E_{z\sim Q}[\log Q(z)] \\ &= E_{z\sim Q}[\log p(x|z;\theta)] - D_{KL}(Q||p_z), \end{aligned}$$

where:

- D_{KL} is the KL divergence given by:

$$D_{KL}(Q||p_z) = \sum_z Q(z) \log \frac{Q(z)}{p(z)}.$$

The second form of ELBO look likes:,

$$ELBO(x;Q,\theta) = \log p(x) - D_{KL}(Q||p_{z|x}).$$

Now, writing the log-likelihood for our Gaussian Mixture model, we will have:

$log\left[\mathcal{L}(\phi,\mu,\Sigma)\right] =$
$\sum_{i=1}^n \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\phi,\Sigma,\mu)}{Q_i(z^{(i)})} =$

$$\sum_{i=1}^n \sum_{j=1}^k Q_i(z^{(i)}=j) \log \frac{p(x^{(i)}|z^{(i)}=j;\mu,\Sigma)p(z^{(i)}=j;\phi)}{Q_i(z^{(i)}=j)} =$$

$$\sum_{i=1}^n \sum_{j=1}^k w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{(d/2)}|\Sigma_j|^{(1/2)}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{(-1)}(x^{(i)}-\mu_j)\right).\phi_j}{w_j^{(i)}}.$$

EM algorithm convergence

Now, we are interesting on addressing the question, will the EM algorithm converge?[4]

Let $\theta^{(t)}$ and $\theta^{(t+1)}$ are the parameters from two successive iterations of EM. We are interested in showing that EM always monotonically improves the log-likelihood.

$$log\left[\mathcal{L}(\theta^{(t)})\right] = \sum_{i=1}^n ELBO(x^{(i)}, \mathbb{Q}_i^{(t)}, \theta^{(t)})$$

$$\begin{aligned} log\left[\mathcal{L}(\theta^{(t+1)})\right] &\geq \sum_{i=1}^n ELBO(x^{(i)}; \mathbb{Q}_i^{(t)}, \theta^{(t+1)}) \\ &\geq \sum_{i=1}^n ELBO(x^{(i)}; \mathbb{Q}_i^{(t)}, \theta^{(t)}) \\ &\geq log\left[\mathcal{L}(\theta^{(t)})\right]. \end{aligned}$$

Related materials For GitHub implementation please check:

`https://github.com/ndams55/Gaussian_Mixture_Model_AMMI_2023`.

For mathematical proves, check:

`https://drive.google.com/file/d/1cL_g0XttfOaaBb2dV1IrwrkVxdOzQb1M/view`.

References

[1] Cathy Maugis, Gilles Celeux, and Marie-Laure Martin-Magniette. "Variable selection for clustering with Gaussian mixture models". In: *Biometrics* 65.3 (2009), pp. 701–709.

[2] Douglas A Reynolds et al. "Gaussian mixture models." In: *Encyclopedia of biometrics* 741.659-663 (2009).

[3] Geoffrey J McLachlan and Suren Rathnayake. "On the number of components in a Gaussian mixture model". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.5 (2014), pp. 341–355.

[4] Andrew Ng. *CS229 Lecture Notes*. 2022. URL: https://cs229.stanford.edu/main_notes.pdf.

[5] B.S. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. A Hodder Arnold Publication. Wiley, 2001. ISBN: 9780340761199. URL: https://books.google.sn/books?id=htZzDGlCnQYC.

[6] Emilie Shireman, Douglas Steinley, and Michael J Brusco. "Examining the effect of initialization strategies on the performance of Gaussian mixture modeling". In: *Behavior research methods* 49 (2017), pp. 282–293.