Gaussian Mixture Models: The EM Algorithm

March 12, 2020

1 Introduction

The purpose of this write up is to provide intuition on the theory behind the EM algorithm, specifically with respect to Gaussian Mixture Models. The EM algorithm is a general purpose algorithm which allows one to iteratively compute maximum likelihood estimators for statistical distributions. In our case, we are concerned with the maximum likelihood estimators of a Gaussian Mixture Model. We begin by describing the pdf of the Gaussian Mixture Model, for which we intend to estimate parameters. We then proceed to sketch out the proof of the EM algorithm (e.g. why it finds the MLE).

2 Gaussian Mixture Model PDF

The Gaussian Mixture Model's pdf looks like this:

$$p(x_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \tag{1}$$

The equation above describes a mixture of K Gaussians, where the k-th Gaussian has weight π_k. μ_k and Σ_k denote the mean and covariance matrix respectively of the k-th Guassian. The EM algorithm will find the MLE for π_k, μ_k and Σ_k.

3 Log Likelihood of the GMM

Recall from Probability Theory that the likelihood function for one data point just the pdf of the distribution we are trying to fit. Let θ be the tuple $(\pi_{z_i}, \mu_{z_i}, \Sigma_{z_i})$. The log-likelihood for n data points is:

$$l(\theta|x_1, \ldots, x_n) = \log \prod_{i=1}^{n} \left(\sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(x_i|\mu_{z_i}, \Sigma_{z_i}) \right) \tag{2}$$

Note that we've swapped z_i with k. The meaning remains the same, but the notation is more convenient, since z_i can be identified as the component for

the ith data point. Furthermore, z_i will be treated as a variable from this point onwards. Recall that the MLE process attempts to fit the parameters of a statistical distribution to a given data set. Since the GMM case contains multiple gaussians, solving for the MLE requires us to "assign" data points to specific Gaussians (components). We do not know of this assignment in advance; otherwise we would not require this undertaking. The use of z_i will allow us to capture the existence of this unobserved variable.

Now the main goal is to find parameters θ, such that the function above is maximized. Since maximizing the log likelihood directly (due to the nested summation), we make use of the EM algorithm to maximize the function.

4 Overview of the EM Algorithm

The EM algorithm consists of an Expectation Step and a Maximization Step. We begin with an initial guess of θ. Then, we will iterate between two steps until convergence (to a local optimum; possibly global). First, we will compute the expected value of the log likelihood function fixing the existing set of parameters. In the process, we will compute a maximal posterior for component assignment (implying that the assignment is "soft", since it is based in probabilities). During the maximization step, we will fix the posterior of the component assignments and pick a new set of parameters θ, such that the expected value is maximized. Intuitively, we maximize on the posterior of the component assignment, then on the parameters θ and repeat this process until we reach a local optimum. The next few sections are focused on the mathematics of this process.

5 Expectation Step

To complete the Expectation step, we first compute the expected value of the log likelihood function. As noted above, we hold θ fixed and find some posterior component assignment for each data point such that the expectation is maximized. We denote the posterior of the component assignment as $q(z_i)$, which is the probability of some component assignment for the ith data point. This is also sometimes call the responsibility of the component.

Note that our original log likelihood function (2) can be rewritten as follows:

$$\begin{aligned} l(\theta|x_1,\ldots,x_n) &= \log \prod_{i=1}^{n} \left(\sum_{z=1}^{K} \pi_z \mathcal{N}(x_i|\mu_z,\Sigma_z) \right) \\ &= \log \prod_{i=1}^{n} \left(\sum_{z_i=1}^{K} p(x_i|\mu_{z_i}\Sigma_{z_i}) p(z_i|\pi_{z_i}) \right) \\ &= \log \prod_{i=1}^{n} \left(\sum_{z_i=1}^{K} p(x_i,z_i|\theta) \right) \end{aligned}$$

Now, we can take the expectation over $q(z)$ and find the optimal $q(z)$. We will find $q(z)$, not by using conventional calculus techniques, but by using other methods. The expectation is:

$$f(\theta) = \mathbb{E}_{q(z_1,\dots z_n)}\left[\log \prod_{i=1}^{n} p(x_i, z_i|\theta)\right] \tag{3}$$

In order to maximize $f(\theta)$, we can show that $f(\theta)$ is always the lower bound of the log likelihood. Before doing so, we can fix x_i, to avoid cumbersome product notation (the results will still hold for what follows). Our log likelihood becomes

$$l(\theta|x) = \log\left(\sum_{z=1}^{K} p(x, z|\theta)\right) \tag{4}$$

And our expectation becomes

$$f(\theta) = \mathbb{E}_{q(z)}[\log p(x, z|\theta)] \tag{5}$$

Now, we can rewrite the log likelihood as follows

$$\begin{aligned}\log\left(\sum_{z=1}^{K} p(x, z|\theta)\right) &= \log\left(\sum_{z=1}^{K} q(z)\frac{p(x, z|\theta)}{q(z)}\right) \\ &\geq \sum_{z=1}^{K} q(z)\log\frac{p(x, z|\theta)}{q(z)}\end{aligned}$$

The inequality holds due to Jensen's Inequality and the fact that $\log\sum_z$ is concave. It will not be proven here, though you may look up a proof if you are interested. We can continue to simplify the equation.

$$\begin{aligned}\sum_{z=1}^{K} q(z)\log\frac{p(x, z|\theta)}{q(z)} &= \sum_{z=1}^{K} q(z)\log p(x, z|\theta) - \sum_{z=1}^{K} q(z)\log q(z) \\ &= \mathbb{E}_{q(z)}[\log p(x, z|\theta)] + C\end{aligned}$$

Note that $C \geq 0$ above, which proves that the expectation is a lower bound on the log likelihood. In particular, we attain equality only if $q(z) = p(z|x, \theta)$,

which is the posterior of the Gaussian components.

$$\begin{aligned}
\sum_{z=1}^{K} q(z) \log \frac{p(x, z|\theta)}{q(z)} &= \sum_{z=1}^{K} p(z|x, \theta) \log \frac{p(x, z|\theta)}{p(z|x, \theta)} \\
&= \sum_{z=1}^{K} p(z|x, \theta) \log \frac{p(z|x, \theta) p(x|\theta)}{p(z|x, \theta)} \\
&= \sum_{z=1}^{K} p(z|x, \theta) \log p(x|\theta) \\
&= \log p(x|\theta) \sum_{z=1}^{K} p(z|x, \theta) \\
&= \log p(x|\theta) \\
&= \log \sum_{z=1}^{K} p(x, z|\theta)
\end{aligned}$$

The last step reintroduces z by taking advantage of the fact that the marginal of x is equivalent to the summation of x and z over all values of z.

Clearly, since $l(\theta|x) = \log \Big(\sum_{z=1}^{K} p(x, z|\theta) \Big)$, we have shown that the expectation is maximized and attains equality when $q(z) = p(z|x, \theta)$.

We can now compute an explicit form for $q(z_1, \dots z_n)$. We have that

$$q(z_1, \dots z_n) = \prod_{i=1}^{n} p(z_i|x_i, \theta) \tag{6}$$

Expanding $p(z_i|x_i, \theta)$, we have

$$\begin{aligned}
\tau_{ik} = p(z_i = k|x_i) &= \frac{p(z_i = k, x_i)}{\sum_{k' \in K} p(z_i = k', x_i} \\
&= \frac{\pi_k \mathcal{N}(x_i|\mu_k \Sigma_k)}{\sum_{k' \in K} \pi_{k'} \mathcal{N}(x_i|\mu_{k'} \Sigma_{k'})}
\end{aligned}$$

Now, we can use the above to compute the expectation, which we can call $f(\theta)$:

$$\begin{aligned} f(\theta) &= \mathbb{E}_{q(z)}\left[\log \prod_{i=1}^{m} p(x_i, z_i|\theta)\right] \\ &= \sum_{i=1}^{n} \mathbb{E}_{p(z_i|x_i,\theta)}[\log p(x_i, z_i|\theta)] \\ &= \sum_{i=1}^{n} \mathbb{E}_{p(z_i|x_i,\theta)}[\log \pi_{z_i} \mathcal{N}(x_i|\mu_{z_i}\Sigma_{z_i})] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}[\log \pi_{z_i} \mathcal{N}(x_i|\mu_{z_i}\Sigma_{z_i})] \end{aligned}$$

This completes the Expectation Step. The expectation ($f(\theta)$) derived above is the maximum possible function given the parameters, θ and the posterior of the component assignment (τ_{ik}).

6 Maximization Step

To complete the Maximization Step, we fix the posterior of the component assignments (τ_{ik}), and then recompute the parameters which maximize the Expectation derived above. In order to do so, we use conventional calculus techniques and take the partial derivatives with respect to each parameter. As of now the derivation of only one parameter is shown. The rest are given without proof.

We compute the value of π_k using the Lagrangian

$$\begin{aligned} L &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}[\log \pi_k + C] + \lambda\left(1 - \sum_{k'=1}^{K} \pi_{k'}\right) \\ \frac{\partial L}{\partial \pi_k} = 0 &= \sum \frac{\tau_{ik}}{\pi_k} - \lambda \\ &\rightarrow \pi_k = \frac{1}{\lambda}\sum_{i=1}^{n} \tau_{ik} \\ &\rightarrow \lambda = n \\ &\rightarrow \pi_k = \frac{1}{n}\sum_{i=1}^{n} \tau_{ik} \end{aligned}$$

Likewise, forming a Lagrangian and solving for μ_k and Σ_k, we get

$$\mu_k = \frac{\sum_{i=1}^{n} \tau_{ik} x_i}{\sum_{i=1}^{n} \tau_{ik}}$$

$$\Sigma_k = \frac{\sum_{i=1}^{n} \tau_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} \tau_{ik}}$$

This completes the maximization step as well.

7 Concluding Remarks

To conclude, Expectation Maximization starts with an initial (uneducated) guess and repeatedly computes the posterior of each component and the parameters of the model such that the expected value of the log likelihood is maximized on each step. This enables the algorithm to eventually converge to a local maxima, though repeated random initializations may be necessary in order to find a global (or at least a better local) optimum.