EM Algorithm

Cyril Equilbec

February 24, 2019

1 Introduction

Probability reminder

$$p(a) = \sum_b p(a, b)$$
$$p(a, b) = p(a \mid b)p(b)$$
$$p(b \mid a) = \frac{p(a, b)}{p(a)} = \frac{p(a \mid b)p(b}{p(a)}$$

Problem

The goal of this algorithm is to maximize $p(x \mid \theta)$ over some parameters θ, this problem is equivalent to maximizing $\log p(x \mid \theta)$ over the same parameters.

Find : $\theta^* = \arg\max_\theta \;\; p(x \mid \theta) = \arg\max_\theta \;\; \log p(x \mid \theta)$

To solve this problem, we'll introduce a dummy variable z and it's instrumental distribution $q(z) > 0, \sum q(z) = 1$ such that :

$$\begin{aligned}
\log p(x \mid \theta) &= \log \sum_z p(x, z \mid \theta) \\
&= \log \sum_z \frac{p(x, z \mid \theta)}{q(z)} q(z) \\
&= \log E \left[\frac{p(x, z \mid \theta)}{q(z)} \right]_{z \sim q(z)}
\end{aligned} \tag{1}$$

Jensen Inequality

If X is a random variable and ϕ is a convex function then :

$$\phi(E(X)) \leq E(\phi(X))$$

In our case, the log function is concave, therefore we can apply the Jensen inequality to $-\log$:

$$-\log E(X) \leq E(-\log(X))$$
$$E(\log(X)) \leq \log E(X)$$

Hence, the lower bound :

$$E\left[\log \frac{p(x, z \mid \theta)}{q(z)}\right]_{z \sim q(z)} \leq \quad \log E\left[\frac{p(x, z \mid \theta)}{q(z)}\right]_{z \sim q(z)} \tag{2}$$

Instead of directly maximizing $\log p(x \mid \theta)$, one can maximize the lower bound $E\left[\log \frac{p(x,z|\theta)}{q(z)}\right]_{z \sim q(z)}$ under the constraint $\sum q(z) = 1$

This is done in two steps : first find $q(z)$ that maximize this lower bound for any θ then find θ that maximize this lower bound given the optimal distribution $q(z)$

$$\begin{aligned} E\left[\log \frac{p(x, z \mid \theta)}{q(z)}\right]_{z \sim q(z)} &= E\left[\log p(x, z \mid \theta) - \log q(z)\right]_{z \sim q(z)} \\ &= E\left[\log p(x, z \mid \theta)\right] - E\left[\log q(z)\right] \\ &= \sum_z q(z) \log p(x, z \mid \theta) - \sum_z q(z) \log q(z) \end{aligned} \tag{3}$$

To find the optimal $q(z)$, we form the Lagrangian

$$\Lambda(q) = \sum_z q(z) \log p(x, z \mid \theta) - \sum_z q(z) \log q(z) + \lambda(1 - \sum_z q(z))$$

Functional derivative

We recall that :

$$\frac{\partial \sum q \log p}{\partial q} = \log p$$
$$\frac{\partial \sum q \log q}{\partial q} = \log q + q\frac{1}{q} = \log q + 1$$
$$\frac{\partial \lambda(1 - \sum q)}{\partial q} = -\lambda$$

Hence, taking the derivative of Λ with respect to q gives :

$$\frac{\partial \Lambda(q)}{\partial q} = \log p(x, z \mid \theta) - \log q(z) - 1 - \lambda \tag{4}$$

Setting this derivative to 0 leads to :

$$\begin{aligned} \frac{\partial \Lambda(q)}{\partial q} = 0 &\Leftrightarrow \log p(x, z \mid \theta) - \log q(z) - 1 - \lambda = 0 \\ &\Leftrightarrow q(z) = p(x, z \mid \theta)e^{-1-\lambda} \end{aligned} \tag{5}$$

We recall that $\sum_z q(z) = 1$ and $\sum_z p(x, z \mid \theta) = p(x \mid \theta)$, hence summing both sides of the equality gives us :

$$\begin{aligned} \sum_z q(z) &= \sum_z p(x, z \mid \theta)e^{-1-\lambda} \\ e^{-1-\lambda} &= \frac{1}{p(x \mid \theta)} \end{aligned} \tag{6}$$

Given any θ, the best lower bound is reached when

$$\boxed{q(z) = \frac{p(x, z \mid \theta)}{p(x \mid \theta)} = p(z \mid x, \theta)}$$

Using this result with θ^t, we now have :

$$\begin{aligned} E\left[\log \frac{p(x, z \mid \theta)}{q(z)}\right]_{z \sim q(z)} &= E\left[\log \frac{p(x, z \mid \theta)}{p(z \mid x, \theta^t)}\right]_{z \sim p(z|x,\theta^t)} \\ &= E\Big[\log p(x, z \mid \theta)\Big] - E\Big[\log p(z \mid x, \theta^t)\Big] \end{aligned} \tag{7}$$

As we found the optimal $q(z)$, the goal is to maximize the lower bound given by (7) over θ. The second term in (7) only depends on θ^t and not θ, we can ignore it in the maximization step.

Our problem can be therefore rewritten as :

$$\boxed{\underset{\theta}{max}\ E\Big[\log p(x, z \mid \theta)\Big]_{z\sim p(z|x,\theta^t)}}$$

EM Algorithm

Goal : find $\theta^* = \arg\max_{\theta} p(x \mid \theta)$

Loop (after random initialization) :

- E-step : $\mathcal{L}(\theta)_t = E\Big[\log p(x, z \mid \theta)\Big]_{z\sim p(z|x,\theta^t)}$
- $\mathcal{M}$-step : $\theta^{t+1} = \arg\max_{\theta} \mathcal{L}(\theta)_t$

2 EM Algorithm for Gaussian Mixture Model

2.1 Data Generation

Let us consider a data set $\{x_1...x_N\}, x_n \in R^m$.

Each data point is assumed to be drawn from a Gaussian Mixture Model of K Gaussian such that :

$$p(x_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)$$

Where μ_k denotes the mean vector of the kth multivariate Gaussian, it's a coordinate in m-dimensional space, which represents the location where samples are most likely to be generated. Σ_k is the covariance matrix.

One can generate those points with the introduction of a new random variable z_k such that :

$$\begin{cases} p(z_k = i) = \pi_i \\ p(x_k \mid z_k = i) = \mathcal{N}(x_k; \mu_i, \Sigma_i) \end{cases}$$

For $K = 3$ and $m = 2$, the data set looks like the figure below.

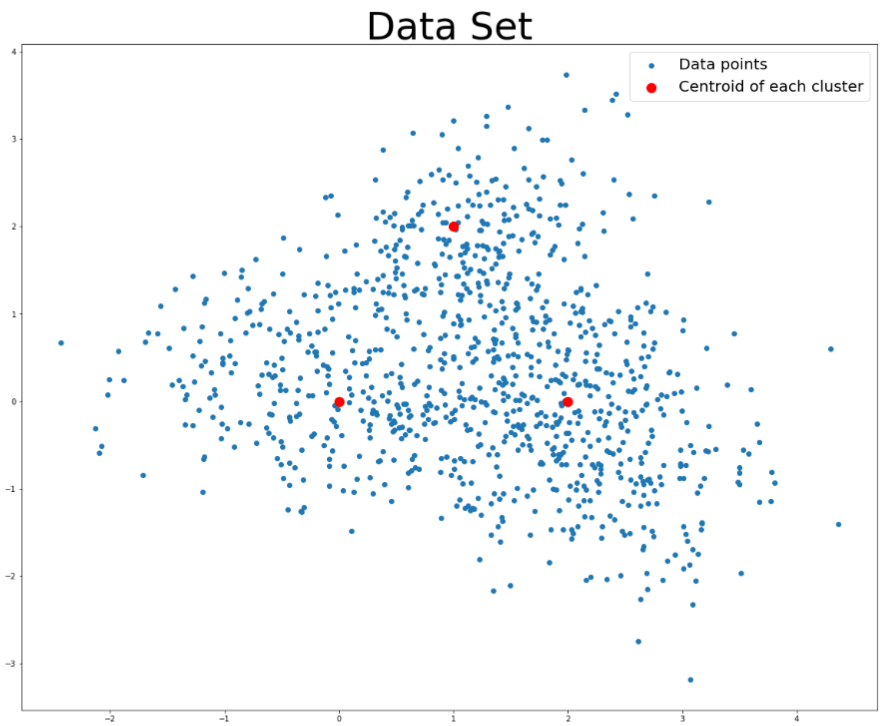

Figure 1: Data set generated by 3 Gaussian

2.2 Expectation step

TO DOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO

2.3 Maximization step

Let's recall that at time step t, we have :

$$\gamma_i^t(x) = \frac{\pi_i^t \mathcal{N}(x, \mu_i^t, \Sigma_i^t)}{\sum_{j=1}^{K} \pi_j^t \mathcal{N}(x, \mu_j^t, \Sigma_j^t)}$$

And our Expectation is :

$$L_t(\theta_k) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k^t(x_n) \mathcal{N}(x_n, \mu_k, \Sigma_k) + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k^t(x_n) \log \pi_k$$

Where $\theta_k = (\pi_k, \mu_k, \Sigma_k)$

We see that the first part only depends on μ_k, Σ_k while the second part only depends on π_k. **We can therefore maximize those terms separately.**

2.3.0.1 π_k^{t+1}

Let's maximize $L_t(\theta)$ with respect to π_k under the constraint that $\sum\limits_{k=1}^{K} \pi_k = 1$

By definition, $\pi_k^{t+1} = \underset{\pi_k}{\operatorname{argmax}}\ L_t(\theta)$ with $\sum\limits_{k=1}^{K} \pi_k = 1$

To solve this problem, we form the Lagrangian

$$\Lambda = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k^t(x_n) \log \pi_k + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

Taking the derivative of the Lagrangian with respect to π_k leads to:

$$\frac{\partial \Lambda}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\partial}{\partial \pi_k} \sum_{k=1}^{K} \gamma_k^t(x_n) \log \pi_k + \lambda \frac{\partial}{\partial \pi_k}(\sum_{k=1}^{K} \pi_k - 1)$$

Those partial derivatives are non-null only when reaching index k inside the sum, hence this result :

$$\frac{\partial \Lambda}{\partial \pi_k} = \sum_{n=1}^{N} \frac{1}{\pi_k} \gamma_k^t(x_n) + \lambda$$

Setting this to 0 gives $\pi_k^* = \pi_k^{t+1}$,

$$\sum_{n=1}^{N} \gamma_k^t(x_n) = -\lambda \pi_k^* \ \ and\ \ \pi_k^* = -\frac{1}{\lambda} \sum_{n=1}^{N} \gamma_k^t(x_n)$$

To find the value of λ one can sum for $k = 1$ to $k = K$:

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k^t(x_n) = -\lambda \sum_{k=1}^{K} \pi_k^*$$

We recall that $\gamma_k^t(x_n)$ and π_k^* are probabilities therefore their sum equal 1. Hence $\lambda = -N$

And our final result :

$$\boxed{\pi_k^{t+1} = \frac{1}{N} \sum_{n=1}^{N} \gamma_k^t(x_n)} \tag{8}$$

2.3.0.2 μ_k^{t+1}

Let's maximize $L_t(\theta)$ with respect to μ_k

By definition, $\mu_k^{t+1} = \underset{\mu_k}{\operatorname{argmax}}\ L_t(\theta)$

First, let's re-use a previous result and introduce another intermediate result:

$$\log \mathcal{N}(x, \mu_i, \Sigma_i) = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log det(\Sigma_i) - \frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)$$

Expending this expression gives :

$$\log \mathcal{N}(x, \mu_i, \Sigma_i) = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log det(\Sigma_i) - \frac{1}{2}(x_i^T \Sigma_i^{-1} x_i - x_i^T \Sigma_i^{-1} \mu_i - \mu_i^T \Sigma_i^{-1} x_i + \mu_i^T \Sigma_i^{-1} \mu_i) \tag{9}$$

Taking the derivative of this expression with respect to μ_i leads to:

$$\frac{\partial}{\partial \mu_i} \log \mathcal{N}(\S, \mu_\rangle, \pm_\rangle) = -\frac{1}{2}(-x_i^T \Sigma_i^{-1} - \Sigma_i^{-1} x_i + (\Sigma_i^{-1} + (\Sigma_i^{-1})^T) \mu_i)$$

Since $-x_i^T \Sigma_i^{-1} = -\Sigma_i^{-1} x_i$ and Σ_i is a symmetric positive definite matrix (covariance matrix) then Σ_i^{-1} is also symmetric and we have $\Sigma_i^{-1} + (\Sigma_i^{-1})^T = 2\Sigma_i^{-1}$, we can rewrite our derivative as :

$$\frac{\partial}{\partial \mu_i} \log \mathcal{N}(x, \mu_i, \Sigma_i) = -\frac{1}{2}(-2\Sigma_i^{-1} x_i + 2\Sigma_i^{-1} \mu_i) = \Sigma_i^{-1}(x_i - \mu_i)$$

Now, taking the derivative of $L_t(\theta)$ with respect to μ_k leads to:

$$\begin{aligned} \frac{\partial L_t(\theta)}{\partial \mu_k} &= \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k^t(x_n) \log \mathcal{N}(\S_\backslash, \mu_\|, \pm_\|) \\ &= \sum_{n=1}^{N} \gamma_k^t(x_n) \Sigma_k^{-1}(x_n - \mu_k) \end{aligned}$$

Setting this to 0 gives $\mu_k^* = \mu_k^{t+1}$:

$$\sum_{n=1}^{N} \gamma_k^t(x_n) \Sigma_k^{-1}(x_n - \mu_k^*) = 0$$

And our final result :

$$\boxed{\mu_k^{t+1} = \frac{\sum\limits_{n=1}^{N} \gamma_k^t(x_n) x_n}{\sum\limits_{n=1}^{N} \gamma_k^t(x_n)}} \tag{10}$$

2.3.0.3 Σ_k^{t+1}

Let's maximize $L_t(\theta)$ with respect to Σ_k

By definition, $\Sigma_k^{t+1} = \underset{\Sigma_k}{\operatorname{argmax}} \; L_t(\theta)$

For this update, few intermediate results are required.

- $\frac{\partial det(A)}{\partial A} = det(A)A^{-1}$ and $det(A^{-1}) = \frac{1}{det(A)}$

- $\frac{\partial \log(f)}{\partial \Sigma_i} = \frac{1}{f}\frac{\partial f}{\partial \Sigma_i}$ (because $\partial \log(f) = \frac{\partial f}{f}$ or by applying the chain-rule)

Since Σ_k^{-1} appears in the expression of $L_t(\theta)$, we'll differentiate with respect to Σ_k^{-1} instead of Σ_k.

We therefore have :

$$\frac{\partial}{\partial \Sigma_i^{-1}} \log \mathcal{N}(x, \mu_i, \Sigma_i) = \frac{\partial}{\partial \Sigma_i^{-1}}(-\frac{N}{2}\log 2\pi - \frac{1}{2}\log det(\Sigma_i) - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))$$

$$= \frac{1}{2}\frac{\partial}{\partial \Sigma_i^{-1}}(\log det(\Sigma_i^{-1}) - (x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))$$

$$= \frac{1}{2}(\frac{1}{det(\Sigma_i^{-1})}\frac{\partial det(\Sigma_i^{-1})}{\partial \Sigma_i^{-1}} - \frac{\partial}{\partial \Sigma_i^{-1}}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))$$

$$= \frac{1}{2}(\Sigma_i - (x-\mu_i)(x-\mu_i)^T)$$

Plugging this result into $\frac{\partial L_t(\theta)}{\partial \Sigma_k}$ gives us :

$$\frac{\partial L_t(\theta)}{\partial \Sigma_k} = \frac{1}{2}\sum_{n=1}^{N} \gamma_k^t(x_n)(\Sigma_k - (x_n-\mu_k)(x_n-\mu_k)^T)$$

Setting this derivative to 0 gives $\Sigma_k^* = \Sigma_k^{t+1}$

$$\frac{1}{2}\sum_{n=1}^{N} \gamma_k^t(x_n)(\Sigma_k^* - (x_n-\mu_k)(x_n-\mu_k)^T) = 0$$

And our final result :

$$\boxed{\Sigma_k^{t+1} = \frac{\sum_{n=1}^{N} \gamma_k^t(x_n)(x_n-\mu_k)(x_n-\mu_k)^T}{\sum_{n=1}^{N} \gamma_k^t(x_n)}} \tag{11}$$

2.4 EM Algorithm for GMM

Result: μ, Σ, π
Data: $\{x_1...x_N\}$, $x_i \in R^m$
Initialization: μ, Σ, π must be randomly chosen, multiples initializations may be required to avoid local optima
while *not converged* **do**

Expectation step:

$$\gamma_k^t(x_n) = \frac{\pi_k^t \mathcal{N}(x_n, \mu_k^t, \Sigma_k^t)}{\sum\limits_{j=1}^{K} \pi_j^t \mathcal{N}(x_n, \mu_j^t, \Sigma_j^t)}$$

Maximization step:

$$\mu_k^{t+1} = \frac{\sum\limits_{n=1}^{N} \gamma_k^t(x_n) x_n}{\sum\limits_{n=1}^{N} \gamma_k^t(x_n)}$$

$$\Sigma_k^{t+1} = \frac{\sum\limits_{n=1}^{N} \gamma_k^t(x_n)(x_n - \mu_k^{t+1})(x_n - \mu_k^{t+1})^T}{\sum\limits_{n=1}^{N} \gamma_k^t(x_n)}$$

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

end

Algorithm 1: EM algorithm for GMM

Comments: μ is a m-dimensional mean vector, Σ is a m $\times$ m covariance matrix and π is K-dimensional weight vector, t is the current time step or iteration. One can choose various stopping criteria :

- A fixed number of iteration ($t \in \{1...300\}$ for instance)
- A threshold of variation of the parameters
- A threshold of variation of log-likelihood