The Expectation Maximization (EM) algorithm

Alessandro Chiuso

January 16, 2017
The Expectation Maximization (EM) algorithm

The EM algorithm (see the famous paper ”Maximum Likelihood from incomplete data via the EM algorithm, by Dempster-Laird-Rubin, JRSS-B, 1977) has been developed to solve the Maximum Likelihood problem

\[ \hat{\theta} := \arg \max_{\theta} p_{\theta}(x) = \arg \max_{\theta} \log(p_{\theta}(x)) \]

It is often useful, to this purpose, to introduce some auxiliary (non-observable) variable \( z \) (“the missing data”) so that the problem

\[ \arg \max_{\theta} p_{\theta}(x, z) = \arg \max_{\theta} \log(p_{\theta}(x, z)) \]

becomes “simple”
The Expectation Maximization (EM) algorithm

The EM algorithm is an alternating algorithm which provides a sequence $\hat{\theta}^{(k)}$, $k \in \mathbb{N}$, satisfying the following properties:

1. $\log(p_{\hat{\theta}(k+1)}(x)) \geq \log(p_{\hat{\theta}(k)}(x))$
2. $\hat{\theta}^{(k)} \to \theta^*$ where $\theta^*$ is a stationary point of $\log(p_{\theta}(x))$

To do so, the algorithm alternates between an *Expectation step* and a *Maximization step*.
Expectation step

Since the variable $z$ is not observed, one needs to “integrate it out”. Intuitively, this can be done defining the following function:

$$Q(\theta, \theta') := \mathbb{E}_{p_{\theta'}(z|x)} \log(p_{\theta}(x, z))$$

The following result holds.

**FACT:**

$$Q(\theta, \theta') \leq \log(p_{\theta}(x)) + C$$

where $C$ does NOT depend on $\theta$, and equality holds for $\theta = \theta'$. This shows that the function $Q(\theta, \theta') - C$ provides a (tight at $\theta = \theta'$) lower bound for $\log(p_{\theta}(x))$. 
The proof is based on properties of the Kullback-Leibler (KL) divergence:

\[
Q(\theta, \theta') = \mathbb{E}_{p_{\theta'}}(z|x) \log(p_\theta(x, z)) \\
= \mathbb{E}_{p_{\theta'}}(z|x) \log \left( \frac{p_\theta(z|x)p_\theta(x)}{p_{\theta'}(z|x)} \right) + \mathbb{E}_{p_{\theta'}}(z|x) \log(p_{\theta'}(z|x)) \\
= \mathbb{E}_{p_{\theta'}}(z|x) \log \left( \frac{p_\theta(z|x)}{p_{\theta'}(z|x)} \right) \\
\leq 0 \quad (=0 \text{ if } \theta = \theta') \\
+ \log(p_\theta(x)) + \mathbb{E}_{p_{\theta'}}(z|x) \log(p_{\theta'}(z|x)) \\
\leq \log(p_\theta(x)) + \mathbb{E}_{p_{\theta'}}(z|x) \log(p_{\theta'}(z|x)) \\
= C
\]
Maximization step

Given a “current” estimate \( \hat{\theta}^{(k)} \) and having performed the Expectation step to compute \( Q(\theta, \hat{\theta}^{(k)}) \), the Maximization step is as follows:

\[
\hat{\theta}^{(k+1)} = \arg \max_{\theta} Q(\theta, \hat{\theta}^{(k)})
\]

**REMARK**

This implies that \( Q(\hat{\theta}^{(k+1)}, \hat{\theta}^{(k)}) \geq Q(\hat{\theta}^{(k)}, \hat{\theta}^{(k)}) \). Using now the fact that \( Q(\theta, \theta') \leq \log(p_\theta(x)) + C \) and \( Q(\theta', \theta') = \log(p'_\theta(x)) + C \) it is clear that

\[
\begin{align*}
\log(p_{\hat{\theta}^{(k+1)}}(x)) & \geq Q(\hat{\theta}^{(k+1)}, \hat{\theta}^{(k)}) - C \\
& \geq Q(\hat{\theta}^{(k)}, \hat{\theta}^{(k)}) - C \\
& = \log(p_{\hat{\theta}^{(k)}}(x))
\end{align*}
\]

proving that the likelihood increases along the sequence \( \hat{\theta}^{(k)} \).
Let us now consider the Gaussian Mixture Model

\[ x \sim p_\theta(x) \]

where

\[ p_\theta(x) = \sum_{\ell=1}^{K} \pi_\ell p_\ell(x) \]

where \( \pi_\ell \geq 0 \), \( \sum_{\ell=1}^{K} \pi_\ell = 1 \) and \( p_\ell(x) \) is the density of a Gaussian random vector with mean \( \mu_\ell \) and variance \( \Sigma_\ell \). The parameter vector \( \theta \) contains all the means \( \mu_\ell \), the variances \( \Sigma_\ell \) as well as the mixing probabilities \( \pi_\ell \).
Let us now introduce the indicator variable $z \in \{1, \ldots, K\}$ which takes value $\ell$ if $x$ comes from the $\ell$–th Gaussian so that:

$$p_\theta(x|z = \ell) = p_\ell(x)$$

With this notation the density of $x$ is of the form

$$p_\theta(x) = \sum_{\ell=1}^{K} p_\theta(x|z = \ell) p_\theta(z = \ell) = p_\ell(x) \pi_\ell$$
EM Algorithm for Gaussian Mixtures Models (GMM) (II)

Now, given i.i.d. observations \( \{x_i\}_{i=1,\ldots,N} \) from the Gaussian Mixture Model, their joint density takes the form:

\[
p_\theta(x_1, \ldots, x_N) = \prod_{i=1}^{N} \sum_{\ell=1}^{K} \pi_\ell p_\ell(x_i)
\]

Estimation of \( \theta := (\mu_\ell, \Sigma_\ell, \pi_\ell, \ell = 1, \ldots, K) \) in the GMM can be performed using the EM algorithm, using as “hidden variable” the indicator variables \( z_i, i = 1, \ldots, N \) alternating between the following steps:

- Given \( \hat{\theta}^{(k)} \) compute \( Q(\theta, \hat{\theta}^{(k)}) \) as above
- Maximize \( Q(\theta, \hat{\theta}^{(k)}) \) over \( \theta \) to obtain \( \hat{\theta}^{(k+1)} \)

**Intuition behind the introduction of the variables** \( z_i \): *if one knew from which component of the mixture each observation \( x_i \) came from, then it would be simple to estimate the parameters of the corresponding component of the mixture*
EM Algorithm for Gaussian Mixtures Models (GMM) (E-Step)

Need to compute:

\[
Q(\theta, \hat{\theta}^{(k)}) := \mathbb{E}_{p_{\hat{\theta}^{(k)}}(z|x)}[\log(p_\theta(x|z)p_\theta(z))]
\]

\[
= \mathbb{E}_{p_{\hat{\theta}^{(k)}}(z|x)}\left[ \sum_{i=1}^{N} \log(p_\theta(x_i|z_i)p_\theta(z_i)) \right]
\]

\[
= \sum_{i=1}^{N} \mathbb{E}_{p_{\hat{\theta}^{(k)}}(z_i|x_i)}[\log(p_\theta(x_i|z_i)p_\theta(z_i))]
\]

\[
= \sum_{i=1}^{N} \left\{ \sum_{\ell=1}^{K} \log(p_\theta(x_i|z_i=\ell)p_\theta(z_i=\ell)) p_{\hat{\theta}^{(k)}}(z_i=\ell|x_i) \right\}
\]

\[
= \sum_{i=1}^{N} \left\{ \sum_{\ell=1}^{K} \log(p_\theta(x_i|z_i=\ell)\pi_\ell) w_{\ell i} \right\}
\]
EM Algorithm for Gaussian Mixtures Models (GMM) (E-Step, II)

Now, using the fact that

$$\log(p_{\theta}(x_i|z_i = \ell)) = \text{const} - \frac{1}{2} \log(\det(\Sigma_\ell)) - \frac{1}{2}(x_i - \mu_\ell)^\top \Sigma_\ell^{-1}(x_i - \mu_\ell)$$

we obtain:

$$Q(\theta, \hat{\theta}^{(k)}) := \text{const} - \frac{1}{2} \sum_{\ell=1}^{K} \log(\det(\Sigma_\ell)) \sum_{i=1}^{N} w_{\ell i} - \frac{1}{2} \sum_{\ell=1}^{K} \sum_{i=1}^{N} (x_i - \mu_\ell)^\top \Sigma_\ell^{-1}(x_i - \mu_\ell) w_{\ell i}$$

$$+ \left\{ \sum_{\ell=1}^{K} \log(\pi_\ell) \sum_{i=1}^{N} w_{\ell i} \right\}$$
EM Algorithm for Gaussian Mixtures Models (GMM) (E-Step, III)

Observation:

\[ w_{\ell i} := p_{\hat{\theta}(k)}(z_i = \ell | x_i) = \frac{\mathcal{N}(\hat{\mu}_{\ell}^{(k)}, \hat{\Sigma}_{\ell}^{(k)})}{\sum_{\ell=1}^{K} p_{\hat{\theta}(k)}(x_i | z_i = \ell) p_{\hat{\theta}(k)}(z_i = \ell)} \]
EM Algorithm for Gaussian Mixtures Models (GMM) (M-Step, I)

To maximise w.r.t. $\pi_\ell$ under the constraint $\sum_{\ell=1}^{K} \pi_\ell = 1$ we use Lagrange multipliers

$$\Lambda(\theta, \lambda) = Q(\theta, \hat{\theta}^{(k)}) + \lambda \left( \sum_{\ell=1}^{K} \pi_\ell - 1 \right)$$

setting to zero the partial derivatives

$$\frac{\partial \Lambda(\theta, \lambda)}{\partial \pi_\ell} = \frac{1}{\pi_\ell} \sum_{i=1}^{N} w_{\ell i} + \lambda = 0$$

which, under the condition $\sum_{\ell=1}^{K} \pi_\ell = 1$ has the unique solution

$$\hat{\pi}_\ell^{(k+1)} = \frac{\sum_{i=1}^{N} w_{\ell i}}{\sum_{j=1}^{K} \sum_{i=1}^{N} w_{ji}} = \frac{1}{N} \sum_{i=1}^{N} w_{\ell i}$$
EM Algorithm for Gaussian Mixtures Models (GMM) (M-Step, I)

Similarly, taking derivatives w.r.t. $\mu_\ell$ we obtain:

$$\frac{\partial \Lambda(\theta, \lambda)}{\partial \mu_\ell} = \frac{\partial Q(\theta, \hat{\theta}^{(k)})}{\partial \mu_\ell} = \sum_{\ell}^{-1} \sum_{i=1}^{N} (x_i - \mu_\ell)w_{\ell i} = 0$$

which admits the unique solution

$$\hat{\mu}_\ell^{(k+1)} = \frac{\sum_{i=1}^{N} x_i w_{\ell i}}{\sum_{i=1}^{N} w_{\ell i}}$$
EM Algorithm for Gaussian Mixtures Models (GMM) (M-Step, I)

Last, it is possible to show (HOMEWORK) that the solution for $\Sigma_\ell$ is given by the equation:

$$
\Sigma_{(k+1)}^{(\ell)} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu}_{(k+1)}^{(\ell)})(x_i - \hat{\mu}_{(k+1)}^{(\ell)})^\top w_{\ell i}
$$