EXPECTATION MAXIMIZATION

JIAN ZHANG
JIANZHAN@STAT.PURDUE.EDU

The *Expectation Maximization* (EM) algorithm [1, 2] is one of the most widely used algorithms in statistics. Suppose we are given some observed data X and a model family parametrized by θ, and would like to find the θ which maximizes $p(X|\theta)$, i.e. the *maximum likelihood estimator*. The basic idea of EM is actually quite simple: when direct maximization of $p(X|\theta)$ is complicated we can augment the data X by introducing some "*hidden variable*" Z such that

$$p(X, Z|\theta)$$

can be computed easily (for example when you observe both X and Z it can be easily maximized with respect to θ).

General Derivation

Suppose we have a guess of the parameter value θ^t and want to find θ such that $p(X|\theta) \geq p(X|\theta^t)$. This can be done by considering the difference between *observed-data* log-likelihood

$$\Delta L = L(\theta) - L(\theta^t) = \log \frac{p(X|\theta)}{p(X|\theta^t)}.$$

Now we introduce the hidden variable Z such that $p(X, Z|\theta)$ is easy to compute (usually in a product form so that $\log p(X, Z|\theta)$ can be factorized). We have

$$\begin{aligned}
L(\theta) - L(\theta^t) &= \log \frac{\int p(X, Z|\theta) dZ}{p(X|\theta^t)} \\
&= \log \left[\int \frac{p(Z|\theta^t, X)}{p(Z|\theta^t, X)} \frac{p(X, Z|\theta)}{p(X|\theta^t)} dZ \right] \\
&\geq \int \left[p(Z|\theta^t, X) \log \frac{p(X, Z|\theta)}{p(Z|\theta^t, X) p(X|\theta^t)} \right] dZ \\
&\triangleq \underline{\Delta} L(\theta; \theta^t).
\end{aligned}$$

where the last inequality is due to Jensen's inequality and the fact that $\log(.)$ is concave. Note that equivalently we have $L(\theta) \geq L(\theta^t) + \underline{\Delta} L(\theta; \theta^t)$, which says that $L(\theta^t) + \underline{\Delta} L(\theta; \theta^t)$ is a global lower bound of $L(\theta)$ for any θ. Consequently we can maximize $\underline{\Delta} L(\theta; \theta^t)$ wrt θ to obtain θ^{t+1}, and as long as $\underline{\Delta} L(\theta^{t+1}; \theta^t) \geq 0$ we have $L(\theta^{t+1}) \geq L(\theta^t)$ (and verify that $\underline{\Delta} L(\theta^t; \theta^t) = 0$).

Now back to the problem of maximizing $\underline{\Delta} L(\theta; \theta^t)$ wrt θ:

$$\begin{aligned}
\theta^{t+1} &= \arg\max_\theta \underline{\Delta} L(\theta; \theta^t) \\
&= \arg\max_\theta \int \left[p(Z|\theta^t, X) \log \frac{p(X, Z|\theta)}{p(Z|\theta^t, X) p(X|\theta^t)} \right] dZ \\
&= \arg\max_\theta \int p(Z|\theta^t, X) \log p(X, Z|\theta) dZ.
\end{aligned}$$

Define

$$Q(\theta; \theta^t) \triangleq \int p(Z|\theta^t, X) \log p(X, Z|\theta) dZ = \mathbb{E}_{Z|\theta^t, X}[\log p(X, Z|\theta)].$$

Finally we derived the EM algorithm:

- *E-step*: compute $Q(\theta;\theta^t)$, which is the **expectation** of *complete-data* log-likelihood $\log p(X,Z|\theta^t)$ and the expectation is wrt $p(Z|\theta^t,X)$.
- *M-step*: **maximize** $Q(\theta;\theta^t)$ wrt θ to obtain θ^{t+1}.

Mixture of Normal Distributions

We now apply EM to fit a mixture of two normal distributions. Suppose we observe $x_1,\dots,x_n$ from a mixture of normal distributions

$$p(x) = \lambda N(\mu_1,\sigma_1^2) + (1-\lambda)N(\mu_2,\sigma_2^2).$$

So in our case the observed data is $\{x_1,\dots,x_n\}$ and the $\theta = \{\lambda,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2\}$. We introduce hidden variables $Z_1,\dots,Z_n$ where $Z_i = 0$ if x_i comes from the first mixture component and 1 otherwise. The complete-data log-likelihood can be written down easily as (due to our introduction of hidden variables):

$$\begin{aligned}
\log p(x_i,z_i|\theta) &= \log\left\{\left[\lambda\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right)\right]^{z_i}\left[(1-\lambda)\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right)\right]^{1-z_i}\right\} \\
&= z_i\log\left[\lambda\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right)\right] + (1-z_i)\log\left[(1-\lambda)\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right)\right]
\end{aligned}$$

and $Q(\theta;\theta^t)$ can be written as (for simplicity we discard constants and group parameters together):

$$\begin{aligned}
Q(\theta;\theta^t) &= \mathbb{E}\left[\sum_{i=1}^n (z_i\log\lambda + (1-z_i)\log(1-\lambda))\right] + \mathbb{E}\left[\sum_{i=1}^n(-z_i\log\sigma_1 - (1-z_i)\log\sigma_2)\right] \\
&+ \mathbb{E}\left[\sum_{i=1}^n\left(-z_i\frac{(x_i-\mu_1)^2}{2\sigma_1^2} - (1-z_i)\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right)\right] \\
&= \sum_{i=1}^n(\mathbb{E}[z_i]\log\lambda - (1-\mathbb{E}[z_i])\log(1-\lambda)) + \sum_{i=1}^n(-\mathbb{E}[z_i]\log\sigma_1 - (1-\mathbb{E}[z_i])\log\sigma_2) \\
&+ \sum_{i=1}^n\left(-\mathbb{E}[z_i]\frac{(x_i-\mu_1)^2}{2\sigma_1^2} - (1-\mathbb{E}[z_i])\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right).
\end{aligned}$$

Define $m_i^1 = \mathbb{E}[z_i]$ and $m_i^2 = 1-\mathbb{E}[z_i]$ and we first work out the M-step assuming that we already know m_i^1, m_i^2's (which depend on the value of θ^t). By maximizing $Q(\theta;\theta^t)$ wrt θ we have

$$\begin{aligned}
\lambda^{t+1} &= \frac{1}{n}\sum_{i=1}^n m_i^1 \\
\mu_j^{t+1} &= \frac{\sum_{i=1}^n m_i^j x_i}{\sum_{i=1}^n m_i^j},\quad (j=1,2) \\
\sigma_j^{t+1} &= \frac{\sum_{i=1}^n m_i^j(x_i-\mu_j^{t+1})^2}{\sum_{i=1}^n m_i^j},\quad (j=1,2).
\end{aligned}$$

Note that the M-step makes perfect sense if we split each x_i into two particles, the first comes from mixture component one with weight m_i^1, etc. The quantity $m_i^1 = \mathbb{E}[z_i]$ which is needed in the E-step can be computed as

$$\begin{aligned}
\mathbb{E}[z_i] &= 1\cdot p(z_i=1|\theta^t,x_1,\dots,x_n) + 0\cdot p(z_i=0|\theta^t,x_1,\dots,x_n) \\
&= \frac{p(x_i,z_i=1|\theta^t)}{p(x_i,z_i=0|\theta^t)+p(x_i,z_i=1|\theta^t)} \\
&= \frac{\lambda^t N(x_i|\mu_1^t,(\sigma_1^t)^2)}{\lambda^t N(x_i|\mu_1^t,(\sigma_1^t)^2) + (1-\lambda^t)N(x_i|\mu_2^t,(\sigma_2^t)^2)}.
\end{aligned}$$

To extend the idea to a mixture of m distributions we can introduce hidden variables $z_{i,j}$ with $i=1,\dots,n$ and $j=1,\dots,m$. Define $z_{ij}=1$ if x_i is generated from the j-th mixture component and 0 otherwise. The rest follows straightforwardly.

EM in the Exponential Family

Let X be the observed data and Z be the hidden variable [2]. Suppose the augmented data $Y = (X, Z)$ are distributed as

$$p(Y|\theta) = \frac{b(Y)}{a(\theta)} \exp(\theta^T s(Y)),$$

i.e. the regular *exponential family*, where $\theta \in \mathbb{R}^d$ is the parameter vector and $s(Y) \in \mathbb{R}^d$ is the vector of *sufficient statistics*.

The $Q(\theta; \theta^t)$ can be written as

$$\begin{aligned} Q(\theta;\theta^t) &= \int p(Z|\theta^t, X) \log p(X, Z|\theta) dZ \\ &= \int p(Z|\theta^t, X) \log b(Y) dZ + \theta^T \int p(Z|\theta^t, X) s(Y) dZ - \log a(\theta). \end{aligned}$$

Notice that the first term does not depend on θ and thus can be thrown away. So the E-step reduces to just compute the *expected sufficient statistics*

$$\mathbb{E}_{Z|\theta^t, X}[S(Y)] \quad \triangleq \quad \mathbf{s}.$$

In the M-step we maximize $\theta^T \mathbf{s} - \log a(\theta)$, where $a(\theta) = \int b(Y) \exp(\theta^T s(Y)) dY$. Compute the derivative wrt θ and set it to zero we have

$$\mathbb{E}_{Y|\theta}[S(Y)] = \int p(X, Z|\theta) S(Y) dY = \mathbf{s}.$$

In other words, the M-step is reduced to find the root θ^{t+1} of the above equation.

EM Generalizations

There are many ways to generalize the standard EM algorithm, and here we just mention a few.

Generalized M-step.

Sometimes it may be difficult or expensive to find $\theta^{t+1} = \arg\max Q(\theta; \theta^t)$. Since all we need is to find θ^{t+1} such that $Q(\theta; \theta^t) \geq Q(\theta^t, \theta^t) = 0$, we may use an easy or cheap method to just maximize $Q(\theta; \theta^t)$ approximately. Note that this change usually results in more iterations to converge and may not slow down your algorithm especially when E-step is expensive to compute.

Generalized E-step.

In practice the E-step is often more complicated than the M-step, and sometimes the exact E-step is difficult to obtain. Recall that in the derivation of EM we have

$$L(\theta) - L(\theta^t) = \log \frac{\int p(X, Z|\theta) dZ}{p(X|\theta^t)} = \log \left[\int \frac{p(Z|\theta^t, X)}{p(Z|\theta^t, X)} \frac{p(X, Z|\theta)}{p(X|\theta^t)} dZ \right] \geq \int \left[p(Z|\theta^t, X) \log \frac{p(X, Z|\theta)}{p(Z|\theta^t, X) p(X|\theta^t)} \right] dZ.$$

In fact in order for the Jensen's inequality to hold we can replace $p(Z|\theta^t, X)$ with any valid distribution $q(Z|\gamma)$ in the above derivation. Thus we have

$$L(\theta) - L(\theta^t) = \log \left[\int \frac{q(Z|\gamma)}{q(Z|\gamma)} \frac{p(X, Z|\theta)}{p(X|\theta^t)} dZ \right] \geq \int \left[q(Z|\gamma) \log \frac{p(X, Z|\theta)}{q(Z|\gamma) p(X|\theta^t)} \right] dZ \triangleq \underline{\Delta} L_{q(.|\gamma)}(\theta, \theta^t).$$

Note also that when $q(Z|\gamma)$ is the true posterior $p(Z|\theta, X)$ the above bound is exact. So the generalized EM works as follows:

- E-step: compute the expectation of complete-data log-likelihood $\log p(X, Z|\theta)$ where the expectation is wrt $q(Z|\gamma)$. You want to use a distribution $q(Z|\gamma)$ which is a good approximation[1] to $p(Z|\theta^t, X)$.
- M-step: maximize $Q_{q(.|\gamma)}(\theta; \theta^t) = \mathbb{E}_{q(Z|\gamma)}[\log p(X, Z|\theta)]$ to obtain θ^{t+1}.

[1] This can be done, for example, by choosing its parameter γ in some parametric family $q(.|\gamma) \in \mathcal{F}$.

Monte Carlo E-step.

Instead of computing $Q(\theta;\theta^t) = \mathbb{E}_{Z|\theta^t,X}[\log p(X,Z|\theta)]$, one may apply the method of Monte Carlo [2] to approximate the Q function. In particular, the Monte Carlo E-step can be computed as:

(1) Draw $z_1,\ldots,z_m \overset{iid}{\sim} p(Z|\theta^t,X)$.
(2) Let $\hat{Q}(\theta;\theta^t) = \frac{1}{m}\sum_{j=1}^{m} \log p(X,z_j|\theta)$.

References

[1] Dempster, A., Laird, N. and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39 (Series B)*, 1-38, 1977.
[2] Martin A. Tanner. *Tools for Statistical Inference*, 3rd edition. Springer, 1996.