Expectation Maximization 01 Algorithm Convergence

Chen Gong

17 December 2019

Expectation Maximization (EM) 算法，是用来解决具有隐变量的模型的概率计算问题。在比较简单的情况中，我们可以直接得出我们想要求得的参数的解析解，比如：MLE: $p(X|\theta)$。我们想要求解的结果就是：

$$\theta_{MLE} = \arg\max_{\theta} \log p(X|\theta) \tag{1}$$

然而一旦问题变得复杂起来以后，就不是这么简单了，特别是引入了隐变量之后。

1 EM 算法简述

实际上，EM 算法的描述也并不是很难，我们知道，通常我们想求的似然函数为 $p(X|\theta)$。引入隐变量之后，原式就变成了：

$$p(X|\theta) = \int \log p(X,Z|\theta) p(Z|X,\theta^{(t)}) dZ \tag{2}$$

EM 算法是一种迭代的算法，我们的目标是求：

$$\begin{aligned} \theta^{(t+1)} &= \arg\max_{\theta} \int_Z \log p(X,Z|\theta) p(Z|X,\theta^{(t)}) dZ \\ &= \arg\max_{\theta} \mathbb{E}_{Z\sim p(Z|X,\theta^{(t)})}[\log p(X,Z|\theta)] \end{aligned} \tag{3}$$

也就是找到一个更新的参数 θ，使得 $\log p(X,Z|\theta)$ 出现的概率更大。

2 EM 算法的收敛性

我们想要证的是当 $\theta^{(t)} \longrightarrow \theta^{(t+1)}$ 时，有 $\log p(X|\theta^{(t)}) \leq \log p(X|\theta^{(t+1)})$。这样才能说明我们的每次迭代都是有效的。

$$\log p(X|\theta) = \log \frac{p(X,Z|\theta)}{p(Z|X,\theta)} = \log p(X,Z|\theta) - \log p(Z|X,\theta) \tag{4}$$

下一步，则是同时对两边求关于 $p(Z|X,\theta^{(t)})$ 的期望。

左边：

$$\begin{aligned} \mathbb{E}_{Z\sim p(Z|X,\theta^{(t)})}[\log p(X|\theta)] &= \int_Z p(Z|X,\theta^{(t)} \log p(X|\theta) dZ \\ &= \log p(X|\theta) \int_Z p(Z|X,\theta^{(t)}) dZ \\ &= \log p(X|\theta) \cdot 1 = \log p(X|\theta) \end{aligned} \tag{5}$$

右边：

$$\underbrace{\int_Z p(Z|X,\theta^{(t)})\log p(X,Z|\theta)dZ}_{Q(\theta,\theta^{(t)})} - \underbrace{\int_Z p(Z|X,\theta^{(t)})\log p(Z|X,\theta)dZ}_{H(\theta,\theta^{(t)})} \tag{6}$$

大家很容易就观察到，$Q(\theta,\theta^{(t)})$ 就是我们要求的 $\theta^{(t+1)} = \arg\max_\theta \int_Z p(X,Z|\theta)p(Z|X,\theta^{(t)})dZ$。那么，根据定义，我们可以很显然的得到：$Q(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta,\theta^{(t)})$。当 $\theta = \theta^{(t)}$ 时，等式也是显然成立的，那么我们可以得到：

$$Q(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta^{(t)},\theta^{(t)}) \tag{7}$$

这时，大家想一想，我们已经得到了 $Q(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta^{(t)},\theta^{(t)})$ 了。如果，$H(\theta^{(t+1)},\theta^{(t)}) \leq H(\theta^{(t)},\theta^{(t)})$。我们就可以很显然的得出，$\log p(X|\theta^{(t)}) \leq \log p(X|\theta^{(t+1)})$ 了。

证明：

$$\begin{aligned} H(\theta^{(t+1)},\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) &= \int_Z p(Z|X,\theta^{(t)})\log p(Z|X,\theta^{(t+1)})dZ - \int_Z p(Z|X,\theta^{(t)})\log p(Z|X,\theta^{(t)})dZ \\ &= \int_Z p(Z|X,\theta^{(t)})\log \frac{p(Z|X,\theta^{(t+1)})}{p(Z|X,\theta^{(t)})}dZ \\ &= -KL(p(Z|X,\theta^{(t)})||p(Z|X,\theta^{(t+1)})) \leq 0 \end{aligned} \tag{8}$$

或者，我们也可以使用 Jensen inequality。很显然，log 函数是一个 concave 函数，那么有 $\mathbb{E}[\log X] \leq \log[\mathbb{E}[X]]$，那么：

$$\begin{aligned} \int_Z p(Z|X,\theta^{(t)})\log \frac{p(Z|X,\theta^{(t+1)})}{p(Z|X,\theta^{(t)})}dZ =& \mathbb{E}_{Z\sim p(Z|X,\theta^{(t)})}\left[\log \frac{p(Z|X,\theta^{(t+1)})}{p(Z|X,\theta^{(t)})}\right] \\ \leq& \log\left[\mathbb{E}_{Z\sim p(Z|X,\theta^{(t)})}\left[\frac{p(Z|X,\theta^{(t+1)})}{p(Z|X,\theta^{(t)})}\right]\right] \\ =& \log\left[\int_Z p(Z|X,\theta^{(t)})\left[\frac{p(Z|X,\theta^{(t+1)})}{p(Z|X,\theta^{(t)})}\right]dZ\right] \\ =& \log \int_Z p(Z|X,\theta^{(t+1)})dZ \\ =& 0 \end{aligned} \tag{9}$$

所以，从两个方面我们都证明了，$\log p(X|\theta^{(t)}) \leq \log p(X|\theta^{(t+1)})$。那么，经过每次的迭代，似然函数在不断的增大。这就证明了我们的更新是有效的，也证明了算法是收敛的。

Expectation Maximization 02 Derived Formula

Chen Gong

18 December 2019

机器学习中，所谓的模型实际上就可以看成是一堆的参数。根据极大似然估计的思想，我们要求解的对象的是：

$$\theta_{MLE} = \log P(X|\theta) \tag{1}$$

其中，X 为 observed data；Z 为 latent data；(X, Z) 为 complete data；θ 为 parameter。

那么，EM 公式就被我们描述为：

$$\theta^{(t+1)} = \arg\max_{\theta} \int_Z \log P(X, Z|\theta) P(Z|X, \theta^{(t)}) dZ \tag{2}$$

EM 算法可以被我们分解成 E-step 和 M-step 两个部分。

E-step:

$$P(Z|X, \theta^{(t)}) \longrightarrow \mathbb{E}_{Z\sim P(Z|X,\theta^{(t)})} [\log P(X, Z|\theta)] \tag{3}$$

M-step:

$$\theta^{(t+1)} = \arg\max_{\theta} \mathbb{E}_{Z\sim P(Z|X,\theta^{(t)})} [\log P(X, Z|\theta)] \tag{4}$$

前面我们已经证明了 EM 算法的收敛性了，也就是：

$$\log P(X|\theta^{(t+1)}) \geq \log P(X|\theta^{(t)}) \tag{5}$$

收敛性告诉了我们算法确实是有效的，我们可以放心的去使用它。而大家会不会觉得这个公式的得来有点懵逼？懵逼就对了，那么下一步，我们的目标就是要推导出 EM 算法究竟是怎么来的，给出一个理论的证明。

1 从 KL Divergence 进行分析

这是个什么东西呢？中文名字叫做“证据下界”。这个名字读起来似乎有一点点奇怪。我们首先看看它是怎么来的。首先，我们定义一个有关于表示层 Z 的表示层变量 $q(Z)$，$q(Z)$ 可以表示任何一个变量。

$$\begin{aligned} \log P(X|\theta) &= \log P(X, Z|\theta) - \log P(Z|X, \theta) \\ &= \log \frac{P(X, Z|\theta)}{Q(Z)} - \log \frac{P(Z|X, \theta)}{Q(Z)} \end{aligned} \tag{6}$$

两边同时对于 $Q(Z)$ 求期望，我们可以得到：

左边:

$$
\begin{aligned}
\int_Z Q(Z)\log P(X|\theta)dZ &= \log P(X|\theta)\int_Z Q(Z)dZ \\
&= \log P(X|\theta)\cdot 1 \\
&= \log P(X|\theta)
\end{aligned} \tag{7}
$$

右边:

$$
\underbrace{\int_Z Q(Z)\log\frac{P(X,Z|\theta)}{Q(Z)}dZ}_{ELBO} - \underbrace{\int_Z Q(Z)\log\frac{P(Z|X,\theta)}{Q(Z)}dZ}_{KL} \tag{8}
$$

所以，实际上，$\log P(X|\theta) = ELBO + KL(Q||P)$。其中，$P(Z|X,\theta)$ 为后验分布 (Posterior)。并且，KL 散度的值一定是大于零的。所以，$\log P(X|\theta) \geq ELBO$，当且仅当 $P(Z|X,\theta) = Q(Z)$ 时等号成立。

EM 算法的一个想法就是想让 ELBO 不断的增加，从而使 $\log P(X|\theta)$ 不断的变大的一种攀爬的迭代方法。

那么，我们对下界进行优化，使下界尽可能的变大，就可以使目标函数不断的上升，那么我们可以得到:

$$
\hat{\theta} = \arg\max_{\theta} ELBO = \arg\min_{\theta} - \int Q(Z)\log\frac{P(X,Z|\theta)}{Q(Z)}dZ \tag{9}
$$

而这里的 $Q(Z)$ 的分布我们怎么得到呢? 这里我们就要来讲一讲 EM 算法的一个核心的理解了。首先我们给出这个理解的图示结果，再对这个图来进行讲解:

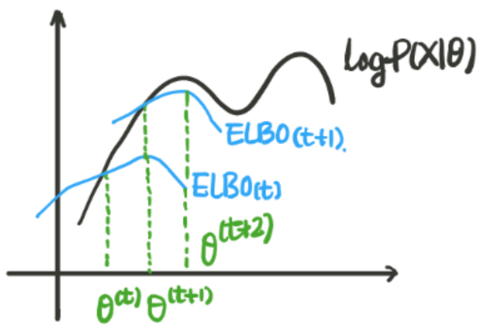

图 1: EM 算法迭代流程图

由于我们的目标是最大化 ELBO。这个下界我们怎么优化? 因为我们需要优化的是 ELBO 的参数 θ。那么，对于某一个时刻的 $\theta^{(t)}$，我们可以的得到一个关于 θ 的函数:

$$
\log P(X|\theta^{(t)}) = \int_Z Q(Z)\log\frac{P(X,Z|\theta)}{Q(Z)}dZ - \int_Z Q(Z)\log\frac{P(Z|X,\theta^{(t)})}{Q(Z)}dZ \tag{10}
$$

由于想让 $\int_Z Q(Z)\log\frac{P(X,Z|\theta)}{Q(Z)}dZ$ 更大。由于 $\log P(X|\theta^{(t)})$ 是一个定值，那么也就是想让 KL 散度的值越小。所以，我们想让 KL 散度的值为零，也就是让 $Q(Z) = P(X,Z|\theta^{(t)})$。这样我们在固定了 $\theta^{(t)}$ 之后就得到了一个 ELBO 关于 θ 的函数。然后我们找到这个函数令值最大的 $\theta^{(t+1)}$ 后开始进行下

一步迭代。实际上我们的目的就是在不断的优化 ELBO，使 ELBO 不断的变大，那么我们想要的结果自然也就变大了，这是一个间接优化的方法。所以，我们紧接着公式 (9) 进行推导：

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \int Q(Z) \log \frac{P(X,Z|\theta)}{Q(Z)} dZ \\ &= \arg\max_{\theta} \int P(X,Z|\theta^{(t)}) \log \frac{P(X,Z|\theta)}{P(X,Z|\theta^{(t)})} dZ \\ &= \arg\max_{\theta} \int P(X,Z|\theta^{(t)}) \log P(X,Z|\theta) - P(X,Z|\theta^{(t)}) P(X,Z|\theta^{(t)}) dZ\end{aligned} \tag{11}$$

由于，$P(X,Z|\theta^{(t)})P(X,Z|\theta^{(t)})$ 与 θ 的求解无关。所以我们可以直接省略掉。那么下一步的 $\theta^{(t+1)}$ 的表达自然也就是：

$$\begin{aligned}\theta^{(t+1)} &= \arg\max_{\theta} \int_Z P(X,Z|\theta^{(t)}) \log P(X,Z|\theta) dZ \\ &= \arg\max_{\theta} \mathbb{E}_{Z\sim P(Z|X,\theta^{(t)})} [\log P(X,Z|\theta)]\end{aligned} \tag{12}$$

而这个公式 (12)，实际上就是我们之前直接给出的公式 (3) 和公式 (4)。

2 从 Jensen Inequality 的角度进行分析

首先，我们介绍一下什么是 Jensen Inequality。实际上，进行过一些机器学习理论研究的同学，都应该听说过这个概念。在这里我们做一个简述。首先我们需要保证函数是一个凸函数，下面我们来画一个凸函数：

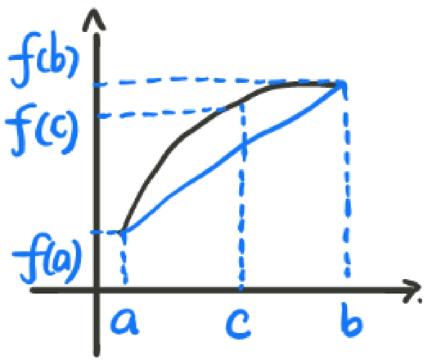

图 2: 凸函数示意图

那么对于一个 $t \in [0,1]$，$c = ta + (1-t)b$，我们都可以得到：

$$f(c) = f[ta + (1-t)b] \geq tf(a) + (1-t)f(b) \tag{13}$$

当 $t = \frac{1}{2}$ 时，我们可以得到：

$$f(\frac{(a+b)}{2}) \geq \frac{1}{2}[f(a) + f(b)] \qquad f[E] \geq E[f] \tag{14}$$

所以，我们可以利用 Jensen Inequality 进行推导：

$$\begin{aligned}
\log P(X|\theta) =& \log \int_z P(X,Z|\theta)dZ \\
=& \log \int_z Q(Z)\frac{P(X,Z|\theta)}{Q(Z)}dZ \\
=& \log \mathbb{E}_{Z\sim Q(Z)}\left[\frac{P(X,Z|\theta)}{Q(Z)}\right] \\
\geq& \mathbb{E}_{Z\sim Q(Z)}\left[\log \frac{P(X,Z|\theta)}{Q(Z)}\right]
\end{aligned} \tag{15}$$

根据 Jensen Inequality 的定义，当 $\frac{P(X,Z|\theta)}{Q(Z)} = C$ 时可以取得等号。不知道，大家有没有发现这里的 $\mathbb{E}_{Z\sim Q(Z)}\left[\log \frac{P(X,Z|\theta)}{Q(Z)}\right]$ 实际上就是 $\int_Z Q(Z)\log \frac{P(X,Z|\theta)}{Q(Z)}dZ$，也就是之前在 KL Divergence 角度进行分析时得到的 ELBO。

毫无疑问，当我们取等时，可以达到最大。所以有，

$$\frac{P(X,Z|\theta)}{Q(Z)} = C \tag{16}$$

$$Q(Z) = \frac{1}{C}P(X,Z|\theta) \tag{17}$$

$$\int_Z Q(Z)dZ = \frac{1}{C}\int_Z P(X,Z|\theta)dZ \tag{18}$$

$$1 = \frac{1}{C}P(X|\theta) \tag{19}$$

所以，我们就可以得到：

$$Q(Z) = \frac{P(X,Z|\theta)}{P(X|\theta)} = P(Z|X,\theta) \tag{20}$$

所以，大家有没有惊奇的发现，这个 $Q(Z)$ 实际上就是 Posterior。当时我们随便引入的一个分布 $Q(Z)$，没想到当它取等的时候就是后验分布。那么像不断去优化这个 ELBO，从而使得 $\log P(X|\theta)$ 的值不断的增加。由于，我们是迭代式的上升，这里的 $Q(Z) = P(Z|X,\theta^{(t)})$，而 $\theta^{(t)}$ 是上一次迭代得到的，我们可以认为是一个常数。所以，

$$\mathbb{E}_{Z\sim Q(Z)}\left[\log \frac{P(X,Z|\theta)}{Q(Z)}\right] = \mathbb{E}_{Z\sim Q(Z)}\left[\log \frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t)})}\right] \tag{21}$$

所以，

$$\theta^{(t+1)} = \arg\max_\theta \mathbb{E}_{Z\sim Q(Z)}\left[\log \frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t)})}\right] \tag{22}$$

所以，从 Jensen Inequality 的角度，我们仍然可以得到 EM 算法的核心表达式。

3 小结

在最后，我们再来来梳理一下 EM 算法的实现思想。我们的目标是想使 $P(X|\theta)$ 似然函数值最大。但是，很不幸，我们直接优化非常的难。所以，我们想到了一个优化下降的方法。对于，每一个 $\theta^{(t)}$ 时，我们可以计算得到下界为：$\mathbb{E}_{Z\sim Q(Z)}\left[\log \frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t)})}\right]$，令这个值最大我们就得到了，想要求得的 $\theta^{(t+1)}$。然后，按这个思路，不断的进行迭代。

Expectation Maximization 03 Generalized Expectation Maximization

Chen Gong

19 December 2019

本小节中，我们想要介绍三个方便的知识点。1. 从狭义的 EM 算法推广到广义的 EM 算法；2. 狭义的 EM 实际上只是广义的 EM 的一个特例；3. 真正的开始介绍 EM 算法。

X：Observed Variable $\longrightarrow X = \{x_i\}_{i=1}^N$；

Z：Latent Variable $\longrightarrow Z = \{Z_i\}_{i=1}^N$；

(X, Z)：Complete Model；

θ：Model Parameter。

我们希望得到一个参数 θ，可以来推导出 X，也就是 $P(X|\theta)$。而这个参数怎么求得呢？所以，这就是一个 learning 的问题了。

1 极大似然估计

所以，根据极大似然估计法的思路，我们要求的最优化参数 $\hat{\theta}$ 为：

$$
\begin{aligned}
\hat{\theta} =& \arg\max_{\theta} P(X|\theta) \\
=& \arg\max_{\theta} \prod_{i=1}^N P(x_i|\theta) \\
=& \arg\max_{\theta} \sum_{i=1}^N \log P(x_i|\theta)
\end{aligned} \tag{1}
$$

好像，我们这样做就可以解决问题了呀。为什么要多此一举的来引入隐变量 Z 呢？这是因为，我们实际观察的输入空间 $\mathcal{X}$ 的分布 $P(X)$，是非常复杂的。可能什么规律都找不出来，这时我们就想到了一个很好的解决办法。我们引入了一个隐变量 Z，这个变量中包含了我们自己的一些归纳总结，引入了内部结构。而 $P(X) = \int_Z P(X, Z)dZ$，实际上就是对 X 进行了分解处理。

2 广义的 EM 算法

EM 算法是为了解决参数估计问题，也就是 learning 问题：

$$
\hat{\theta} = \arg\max_{\theta} P(X|\theta) \tag{2}
$$

但是，$P(X|\theta)$ 可能会非常的复杂。那么，在生成模型的思路中，可以假设一个隐变量 Z。有了这个生成模型的假设以后，我们就可以引入一些潜在归纳出的结构进去。通过 $P(X) = \frac{P(X,Z)}{P(Z|X)}$，就可以把问题具体化了。

这里说明一下，我们习惯用的表达是 $\log P(X|\theta)$，但是也有的文献中使用 $P(X;\theta)$ 或者 $P_\theta(X)$。这三种表达方式代表的意义是等价的。

前面我们已经说过了，我们的目标是：

$$\log P(X|\theta) = \underbrace{ELBO}_{L(Q,\theta)} + KL(Q||P) \geq L(Q,\theta) \tag{3}$$

$$\begin{cases} ELBO = \int_Z Q(Z) \log \frac{P(X,Z|\theta)}{Q(Z)} dZ \\ KL(Q||P) = \int_Z Q(Z) \log \frac{Q(Z)}{P(Z|X,\theta)} dZ \end{cases} \tag{4}$$

但是，问题马上就上来了，那就是 $P(Z|X,\theta)$ 非常有可能求不出来。那么我们怎么来求解这个方程呢？也就是使下界变得更大。

首先第一步，我们把 θ 给固定住。那么，$P(Z|X,\theta)$ 的结果就是一个定值。那么 KL 越小，ELBO 就会越大。由于，$Q(Z)$ 是我们引入的一个中间变量，那么我们的第一步就是得到：

$$\hat{Q}(Z) = \arg\min_Q KL(Q||P) = \arg\max_Q L(Q,\theta) \tag{5}$$

当 Q 被我们求出来以后，我们就可以将 Q 固定了，再来求解 θ：

$$\hat{\theta} = \arg\max_\theta L(\hat{q},\theta) \tag{6}$$

那么，广义的 EM 算法，就可以被我们定义为：

$$\begin{aligned} & E-step: \quad Q^{(t+1)} = \arg\max_Q L(Q(Z),\theta^{(t)}) \\ & M-step: \quad \theta^{(t+1)} = \arg\max_\theta L(Q(Z)^{(t+1)},\theta) \\ & L(Q,\theta) = \mathbb{E}_Q\left[\log P(X,Z) - \log Q\right] = \mathbb{E}_Q\left[\log P(X,Z)\right] - \mathbb{E}_Q\left[\log Q\right] \end{aligned} \tag{7}$$

看到这里，我估计大家已经可以理解上一小节中，为什么有的 θ 带 (t) 有的不带。因为，首先第一步中是固定 θ 求 Q，这里的 θ 就是来自于上一次迭代的 $\theta^{(t+1)}$。第二次，是将上一步求得的 Q 固定，将 θ 看成参数，来求最优的表达结果的 $\theta^{(t+1)}$。另一个方面，从等式 (7) 的第三行，我们可以可以看出实际上：

$$ELBO = \mathbb{E}_{Q(Z)}[\log P(X,Z|\theta)] + H(Q(Z)) \tag{8}$$

我们对比一下上一节讲到的 EM 算法，就会惊奇的发现，ELBO 中最后那个 $H(Q(Z))$ 竟然不见了。这是为什么呢？其实也很好理解的。因为在 M-step 中，我们假定 $Q(Z)$ 已经是固定的了，那么显然 $H[Q(Z)]$ 就是一个定值了，并且与我们的优化目标 θ 之间并没有任何的关系，所以就被我们给省略掉了。

所以，本小节中引出了广义 EM 算法，也说明了原来的 EM 算法是广义 EM 算法的一种特殊情况。

3 坐标上升法

EM 算法的整体描述如下所示：

$$\begin{cases} E-step: & Q^{(t+1)} = \arg\max_Q L(Q(Z), \theta^{(t)}) \\ M-step: & \theta^{(t+1)} = \arg\max_\theta L(Q(Z)^{(t+1)}, \theta) \end{cases} \tag{9}$$

这个坐标上升法 (SMO) 是个什么东西呢？具体的描述，大家可以去网上找找资料看一看。两者都是迭代的思路，在这里我们将它和梯度下降法的优化思路放在一起，做一个小小的对比。大家就会发现有什么不一样的地方，

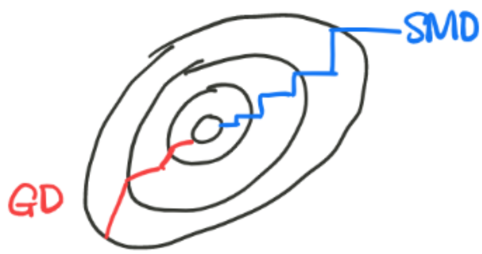

图 1: 坐标上升法和梯度上升法的优化思路对比

我们发现坐标上升法的优化方向基本是恒定不变的，而梯度下降法的优化方向是随着梯度方向而不断发生改变的。

讲到这里，好像一切都很完美，可以圆满的结束了。但是，很不幸的是，问题马上又来了。因为，现实生活中，并没有那么的容易，一切都没有我们想的那么的简单。实际上，有关 $P(Z|X,\theta)$ 的计算，有可能会非常的复杂。所以，我们将采用变分推断 (Variable Inference) 或者马尔可夫蒙特卡罗采样 (Markov Chain Monte Carlo) 的方法来求解。结合起来以后就是，VBEM/VEM 和 MCEM。这里注意一下，Variable Inference 和 Variable Bayes 实际上都是一种东西。

当然，虽然 EM 算法看上去好像很厉害的样子。但是，没有一种算法可以一劳永逸的解决所有的问题。它一定存在优点，也一定有无法解决的问题。具体描述，大家可以去网上寻找相关的资料，我这里就不做过多的描述了。