期望最大

期望最大算法的目的是解决具有隐变量的混合模型的参数估计（极大似然估计）。MLE 对 $p(x|\theta)$ 参数的估计记为：$\theta_{MLE}=\underset{\theta}{argmax}\log p(x|\theta)$。EM 算法对这个问题的解决方法是采用迭代的方法：

$$\theta^{t+1}=\underset{\theta}{argmax}\int_z\log[p(x,z|\theta)]p(z|x,\theta^t)dz=\mathbb{E}_{z|x,\theta^t}[\log p(x,z|\theta)] \tag{1}$$

这个公式包含了迭代的两步：

1. E step：计算 $\log p(x,z|\theta)$ 在概率分布 $p(z|x,\theta^t)$ 下的期望
2. M step：计算使这个期望最大化的参数得到下一个 EM 步骤的输入

> 求证：$\log p(x|\theta^t)\le\log p(x|\theta^{t+1})$
>
> 证明：$\log p(x|\theta)=\log p(z,x|\theta)-\log p(z|x,\theta)$，对左右两边求积分：
>
> $$Left:\int_zp(z|x,\theta^t)\log p(x|\theta)dz=\log p(x|\theta) \tag{2}$$
>
> $$Right:\int_zp(z|x,\theta^t)\log p(x,z|\theta)dz-\int_zp(z|x,\theta^t)\log p(z|x,\theta)dz=Q(\theta,\theta^t)-H(\theta,\theta^t) \tag{18}$$
>
> 所以：
>
> $$\log p(x|\theta)=Q(\theta,\theta^t)-H(\theta,\theta^t) \tag{4}$$
>
> 由于 $Q(\theta,\theta^t)=\int_zp(z|x,\theta^t)\log p(x,z|\theta)dz$，而 $\theta^{t+1}=\underset{\theta}{argmax}\int_z\log[p(x,z|\theta)]p(z|x,\theta^t)dz$，所以 $Q(\theta^{t+1},\theta^t)\ge Q(\theta^t,\theta^t)$。要证 $\log p(x|\theta^t)\le\log p(x|\theta^{t+1})$，需证：$H(\theta^t,\theta^t)\ge H(\theta^{t+1},\theta^t)$：
>
> $$\begin{aligned}H(\theta^{t+1},\theta^t)-H(\theta^t,\theta^t)&=\int_zp(z|x,\theta^t)\log p(z|x,\theta^{t+1})dz-\int_zp(z|x,\theta^t)\log p(z|x,\theta^t)dz\\&=\int_zp(z|x,\theta^t)\log\frac{p(z|x,\theta^{t+1})}{p(z|x,\theta^t)}=-KL(p(z|x,\theta^{t+1}),p(z|x,\theta^t))\le0\end{aligned} \tag{5}$$
>
> 综合上面的结果：
>
> $$\log p(x|\theta^t)\le\log p(x|\theta^{t+1}) \tag{6}$$

根据上面的证明，我们看到，似然函数在每一步都会增大。进一步的，我们看 EM 迭代过程中的式子是怎么来的：

$$\log p(x|\theta)=\log p(z,x|\theta)-\log p(z|x,\theta)=\log\frac{p(z,x|\theta)}{q(z)}-\log\frac{p(z|x,\theta)}{q(z)} \tag{7}$$

分别对两边求期望 $\mathbb{E}_{q(z)}$：

$$
\begin{aligned}
&Left:\int_z q(z)\log p(x|\theta)dz=\log p(x|\theta) &(8)\\
&Right:\int_z q(z)\log\frac{p(z,x|\theta)}{q(z)}dz-\int_z q(z)\log\frac{p(z|x,\theta)}{q(z)}dz=ELBO+KL(p(z|x,\theta),q(z)) &(9)
\end{aligned}
$$

上式中，Evidence Lower Bound(ELBO)，是一个下界，所以 $\log p(x|\theta)\ge ELBO$，等于号取在 KL 散度为0是，即：$q(z)=p(z|x,\theta)$，EM 算法的目的是将 ELBO 最大化，根据上面的证明过程，在每一步 EM 后，求得了最大的ELBO，并根据这个使 ELBO 最大的参数代入下一步中：

$$
\hat{\theta}=\mathop{argmax}_{\theta}ELBO=\mathop{argmax}_{\theta}\int_z q(z)\log\frac{p(x,z|\theta)}{q(z)}dz \tag{10}
$$

由于 $q(z)=p(z|x,\theta^t)$ 的时候，这一步的最大值才能取等号，所以：

$$
\begin{aligned}
\hat{\theta}=\mathop{argmax}_{\theta}ELBO=\mathop{argmax}_{\theta}\int_z q(z)\log\frac{p(x,z|\theta)}{q(z)}dz&=\mathop{argmax}_{\theta}\int_z p(z|x,\theta^t)\log\frac{p(x,z|\theta)}{p(z|x,\theta^t)}dz \quad (11)\\
&=\mathop{argmax}_{\theta}\int_z p(z|x,\theta^t)\log p(x,z|\theta)
\end{aligned}
$$

这个式子就是上面 EM 迭代过程中的式子。

从 Jensen 不等式出发，也可以导出这个式子：

$$
\begin{aligned}
\log p(x|\theta)=\log\int_z p(x,z|\theta)dz&=\log\int_z\frac{p(x,z|\theta)q(z)}{q(z)}dz \quad (12)\\
&=\log\mathbb{E}_{q(z)}[\frac{p(x,z|\theta)}{q(z)}]\ge\mathbb{E}_{q(z)}[\log\frac{p(x,z|\theta)}{q(z)}]
\end{aligned}
$$

其中，右边的式子就是 ELBO，等号在 $p(x,z|\theta)=Cq(z)$ 时成立。于是：

$$
\begin{aligned}
\int_z q(z)dz=\frac{1}{C}\int_z p(x,z|\theta)dz=\frac{1}{C}p(x|\theta)=1 \quad (13)\\
\Rightarrow q(z)=\frac{1}{p(x|\theta)}p(x,z|\theta)=p(z|x,\theta)
\end{aligned}
$$

我们发现，这个过程就是上面的最大值取等号的条件。

广义 EM

EM 模型解决了概率生成模型的参数估计的问题，通过引入隐变量 z，来学习 θ，具体的模型对 z 有不同的假设。对学习任务 $p(x|\theta)$，就是学习任务 $\frac{p(x,z|\theta)}{p(z|x,\theta)}$。在这个式子中，我们假定了在 E 步骤中，$q(z)=p(z|x,\theta)$，但是这个$p(z|x,\theta)$ 如果无法求解，那么必须使用采样（MCMC）或者变分推断等方法来近似推断这个后验。我们观察 KL 散度的表达式，为了最大化 ELBO，在固定的 θ 时，我们需要最小化 KL 散度，于是：

$$
\hat{q}(z)=\mathop{argmin}_{q}KL(p,q)=\mathop{argmax}_{q}ELBO \tag{14}
$$

这就是广义 EM 的基本思路：

1. E step：

$$
\hat{q}^{t+1}(z)=\mathop{argmax}_q\int_z q^t(z)\log\frac{p(x,z|\theta)}{q^t(z)}dz, fixed\,\theta \tag{15}
$$

2. M step：

$$
\hat{\theta}=\mathop{argmax}_\theta\int_z q^{t+1}(z)\log\frac{p(x,z|\theta)}{q^{t+1}(z)}dz, fixed\,\hat{q} \tag{16}
$$

对于上面的积分：

$$
ELBO=\int_z q(z)\log\frac{p(x,z|\theta)}{q(z)}dz=\mathbb{E}_{q(z)}[p(x,z|\theta)]+Entropy(q(z)) \tag{17}
$$

因此，我们看到，广义 EM 相当于在原来的式子中加入熵这一项。

EM 的推广

EM 算法类似于坐标上升法，固定部分坐标，优化其他坐标，再一遍一遍的迭代。如果在 EM 框架中，无法求解 z 后验概率，那么需要采用一些变种的 EM 来估算这个后验。

1. 基于平均场的变分推断，VBEM/VEM
2. 基于蒙特卡洛的EM，MCEM