EM Algorithm

> *Since there are some problems about mathjax, I also upload the pdf version 「Statistics Model」 1 Generalized Linear Model.*

背景

EM算法在1977年由 Arthur Dempster, Nan Laird, 和Donald Rubin在《Maximum Likelihood from Incomplete Data via the EM Algorithm》中首次正式提出。

EM算法通常在频率学派中用于处理缺失数据的一种很广泛的算法，其主要思想是采用迭代的算法，不断通过条件在观测数据上的对缺失数据的似然进行估计，并且选取参数极大化似然，从而得到对参数的推断。

也就是说，EM算法对Missing Data的处理方式是通过取期望移除，而这个期望是条件在观测数据上的。

EM算法

令Y代表观测数据，U代表未观测数据，我们的目标是在如下模型中对参数θ进行推断：

$$f(y;\theta)=\int f(y|u;\theta)f(u;\theta)\mathrm{d}u$$

这里，我们并不直接计算$f(y;\theta)$，而是通过计算完全数据的对数似然(*complete-data log likelihood*)的期望

$$\log f(y,u;\theta)=\underbrace{\log f(y;\theta)}_{l(\theta)}+\log f(u\mid y;\theta)$$

但是由于u未知，这个式子仍然无法进行计算。因此，我们计算其条件在观测数据和当前参数$(Y=y,\theta')$上的期望并将其记为$Q(\theta;\theta')$：

$$\mathrm{E}\left\{\log f(Y,U;\theta)\mid Y=y;\theta'\right\}=\ell(\theta)+\mathrm{E}\left\{\log f(U\mid Y;\theta)\mid Y=y;\theta'\right\}$$

算法

1. E步：计算$Q(\theta;\theta')$;
2. M步：固定θ'，选取θ^+最大化$Q(\theta;\theta')$，令$\theta'=\theta^+$;
3. 重复上面两步直到收敛。

为什么EM算法可行？

注意到我们想要最大话的目标是$\ell(\theta)$，但是我们实际上在每一步最大化的目标是$Q(\theta;\theta')$。但是，通过推导可以得到

$$Q\left(\theta;\theta'\right) \geq Q\left(\theta';\theta'\right) \text{ implies } \ell(\theta) - \ell\left(\theta'\right) \geq C\left(\theta';\theta'\right) - C\left(\theta;\theta'\right) \geq 0$$

因此，每一步迭代都使得$\ell(\theta)$不减，从而最大化。

算法加速

尽管EM算法的理论性质良好，但是它的收敛比较慢，有时候可以通过直接将M步改为直接最大化，考虑如下近似：

$$\frac{\partial \ell(\theta)}{\partial \theta} = \left.\frac{\partial Q\left(\theta;\theta'\right)}{\partial \theta}\right|_{\theta'=\theta} \qquad \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}} = \left.\left\{\frac{\partial^2 Q\left(\theta;\theta'\right)}{\partial \theta \partial \theta^{\mathrm{T}}} + \frac{\partial^2 Q\left(\theta;\theta'\right)}{\partial \theta \partial \theta'^{\mathrm{T}}}\right\}\right|_{\theta'=\theta}$$

这样就可以直接进行牛顿二阶梯度下降法。

参考资料

- Statstical Models
- Expectation-Maximization algorithm wiki