EM 算法简介

chwang@microsoft.com

*注意：这是一个**非正式**的简介，所以没有列出相关的参考文献，以后我会陆续完善。写 EM 简介的文章很多，我这里希望能对初学者有所帮助。水平有限，如有错误，敬请指正。*

Expectation-Maximization (EM) 算法是一个在含有隐含变量的模型中常用的算法，最常见的是用于高斯混合模型 (Mixtures of Gaussians), 但 EM 算法的应用决不仅仅限于此 。

假设我们要估计联合概率密度$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\Phi})$，其中$\boldsymbol{x}$为可以观测的变量，$\boldsymbol{y}$为隐含变量，$\boldsymbol{\Phi}$为参数。令$\{\boldsymbol{X}, \boldsymbol{Y}\} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$。$(\boldsymbol{x}_i, \boldsymbol{y}_i), i \in \{1,2, \dots, N\}$是相互独立的。给定$\boldsymbol{X}$，考虑$\boldsymbol{\Phi}$的最大似然估计(Maximum Likelihood),也就是最大化下面式子:

$\mathcal{L}(\boldsymbol{\Phi}|\boldsymbol{X}) = \log \prod_i p(\boldsymbol{x}_i|\boldsymbol{\Phi}) = \sum_i \log p(\boldsymbol{x}_i|\boldsymbol{\Phi}) = \sum_i log \sum_{\boldsymbol{y}_i} p(\boldsymbol{x}_i, \boldsymbol{y}_i|\boldsymbol{\Phi})$.

写的更一般一些，可以为：

$\mathcal{L}(\boldsymbol{\Phi}|\boldsymbol{X}) = \log p(\boldsymbol{X}|\boldsymbol{\Phi}) = \log \sum_{\boldsymbol{Y}} p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\Phi})$ 。如果$\boldsymbol{Y}$是连续变量，$\mathcal{L}(\boldsymbol{\Phi}|\boldsymbol{X}) = \log \int p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\Phi}) d\boldsymbol{Y}$

基本的 EM 算法：

假设$\boldsymbol{Y}$可观测，那么$\mathcal{L}(\boldsymbol{\Phi}|\boldsymbol{X}, \boldsymbol{Y}) = \log p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\Phi})$, 通常来说比较容易求解，特别是如果$p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\Phi})$是指数族的情况下。

定义Q函数: $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{old}) = E_{p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Phi}^{old})}[\mathcal{L}(\boldsymbol{\Phi}|\boldsymbol{X}, \boldsymbol{Y})]$

E-step：Compute $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Phi}^{old})$

M-step: $\boldsymbol{\Phi}^{\text{new}} = \text{argmax}_{\boldsymbol{\Phi}} Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{old})$

可以证明每次 EM 都可以使得$\mathcal{L}(\boldsymbol{\Phi}|\boldsymbol{X})$增加，保证收敛到局部极大值。

上述的 EM 算法虽然描述简单，但不是很直观，下面从另外一个角度来导出 EM 算法。

从优化的角度导出的 EM 算法：

首先介绍 Jensen 不等式:

假设$f(\boldsymbol{x})$为凸函数函数（对于凹函数，不等号方向相反），（对于二阶可导的情况，可以考虑证明 Hessian 矩阵是非负定的，一维情况$f''(x) \geq 0$）, 对于随机变量$\boldsymbol{X}$,有

$$E[f(\boldsymbol{X})] \geq f(E\boldsymbol{X})$$

恒成立，等号成立当且仅当$\boldsymbol{X} = E\boldsymbol{X}$, 以概率 1 成立（$\boldsymbol{X}$是常数）。如图 1 所示。

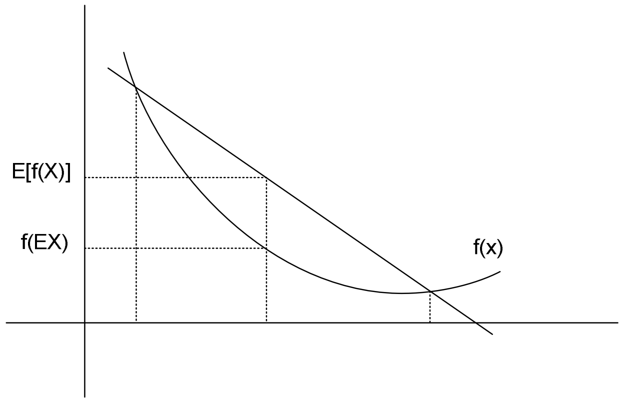

图 1. 凸函数

很容易验证$\log(x)$是凹函数。由于直接优化$\mathcal{L}(\mathbf{\Phi}|\mathbf{X}) = \log p(\mathbf{X}|\mathbf{\Phi}) = \log \sum_{\mathbf{Y}} p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi})$很难，考虑$\mathcal{L}(\mathbf{\Phi}|\mathbf{X})$的下界，假设$q(\mathbf{Y})$为任意概率密度：

$$\mathcal{L}(\mathbf{\Phi}|\mathbf{X}) = \log \sum_{\mathbf{Y}} p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi}) = \log \sum_{\mathbf{Y}} q(\mathbf{Y}) \frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi})}{q(\mathbf{Y})} = \log E_{q(\mathbf{Y})}\left[\frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi})}{q(\mathbf{Y})}\right]$$

由 Jensen 不等式:

$$\log E_{q(\mathbf{Y})}\left[\frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi})}{q(\mathbf{Y})}\right] \geq E_{q(\mathbf{Y})}\left[\log \frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi})}{q(\mathbf{Y})}\right]$$

等号成立当且仅当$\frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi})}{q(\mathbf{Y})} = E_{q(\mathbf{Y})}\left[\frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi})}{q(\mathbf{Y})}\right] = p(\mathbf{X}|\mathbf{\Phi}) = constant$，所以有$q(\mathbf{Y}) = p(\mathbf{Y}|\mathbf{X},\mathbf{\Phi})$。可见，这就是在基本 EM 算法中的 E-step。

而 M-step 则变为:

$$\mathbf{\Phi}^{\text{new}} = \operatorname{argmax}_{\mathbf{\Phi}^*} \log E_{q(\mathbf{Y})}\left[\frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi}^*)}{q(\mathbf{Y})}\right] = \operatorname{argmax}_{\mathbf{\Phi}^*} E_{q(\mathbf{Y})}\left[\log \frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi}^*)}{q(\mathbf{Y})}\right]$$

总结，考虑最大化$\mathcal{L}(\mathbf{\Phi}|\mathbf{X})$下界$E_{q(\mathbf{Y})}\left[\log \frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi})}{q(\mathbf{Y})}\right] = \sum_{\mathbf{Y}} q(\mathbf{Y}) \log \frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi})}{q(\mathbf{Y})}$的问题，从而解决了最大化$\mathcal{L}(\mathbf{\Phi}|\mathbf{X})$的问题。利用 Jensen 也很容易证明$\mathcal{L}(\mathbf{\Phi}|\mathbf{X})$是递增的。假设$\mathbf{\Phi}^n, \mathbf{\Phi}^{n+1}$为迭代第$n, n+1$的值, 并且 $q^n(\mathbf{Y}) = p(\mathbf{Y}|\mathbf{X},\mathbf{\Phi}^{\boldsymbol{n}})$有：

$$\mathcal{L}(\mathbf{\Phi}^{\boldsymbol{n+1}}|\mathbf{X}) = \log E_{q^n(\mathbf{Y})}\left[\frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi}^{\boldsymbol{n+1}})}{q^n(\mathbf{Y})}\right]$$

$$\geq E_{q^n(\mathbf{Y})}\left[\log \frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi}^{\boldsymbol{n+1}})}{q^n(\mathbf{Y})}\right]$$

$$\geq E_{q^n(\mathbf{Y})}\left[\log \frac{p(\mathbf{X},\mathbf{Y}|\mathbf{\Phi}^{\boldsymbol{n}})}{q^n(\mathbf{Y})}\right]$$

$$= \mathcal{L}(\mathbf{\Phi}^{\boldsymbol{n}}|\mathbf{X})$$

上式中第一个不等式是由于 Jensen 不等式，第二个是由于 M-step 求最大的性质。

有些时候，$q(\mathbf{Y}) = p(\mathbf{Y}|\mathbf{X},\mathbf{\Phi})$也不是能够显示的计算出来的，这个时候最大化$\mathcal{L}(\mathbf{\Phi}|\mathbf{X})$就显得相当困难。这个时候，可以考虑不一定保证 Jensen 不等式一定取等号。如果给定$q(\mathbf{Y})$某种形式，就得到 Variational EM 算法。

To be continued …