Notes on Expectation Maximization and Variational Inference

https://hustwj.github.io/notes/
https://github.com/hustwj/notes

Mathematic Framework

Expectation Maximization (EM) and Variational Inference (VI) 开始都是基于下面的基本框架：

$$
\begin{aligned}
\ln p(X) &= \mathbb{E}_q\left[\ln \frac{p(X,Z)}{q(Z)} - \ln \frac{p(Z|X)}{q(Z)}\right] \\
&= \mathbb{E}_q\left[\ln \frac{p(X,Z)}{q(Z)}\right] - \mathbb{E}_q\left[\ln \frac{p(Z|X)}{q(Z)}\right] \\
&= \int q(Z) \ln \frac{p(X,Z)}{q(Z)} \,\mathrm{d}Z - \int q(Z) \ln \frac{p(Z|X)}{q(Z)} \,\mathrm{d}Z \\
&= ELOB(q(Z)) + KL(q(Z)||p(Z|X))
\end{aligned}
$$

虽然EM和VI都在各自的计算过程中希望最大化ELOB和最小化KL，有些相似，但是实质上在两者基本目标和思路还是有很大区别的。

EM的根本目标是观察到的数据X的likelihood最大化，也就是$\ln p(X) = \ln p(X;\theta)$的最大化。对能够用EM来求解的这类问题，当给定特定的θ值，可以方便evaluate和计算出$p(Z|X;\theta)$的值，通过选择$q(Z) = p(Z|X;\theta)$就可以使得两者之间的KL为0。所以可以通过不断调整θ来逐渐实现$\ln p(X;\theta)$最大化。这里θ是一个参数而不是随机变量，对θ求解是点估计，获得某个确定的值。

而在VI中，尤其是在fully Bayesian model中，所有的未知的参数都给予了先验概率，被吸收为latent variables Z的一部分。这时候$\ln p(X)$与Z无关，不管Z如何变化，对其没有影响，所以可以看作是一个常量。VI求解的根本目标求解出Z给出X的后验概率$p(Z|X)$，而用VI来求解的这类问题，是无法方便计算出$p(Z|X)$，所以只能通过$q(Z)$尽可能近似和逼近$p(Z|X)$，然后通过通过调整$q(Z)$最小化两者的KL来实现这样的目标的，而最小化KL等价于最大化ELOB。

实例

典型的例子是Topic model的两种最常用的方法：pLSA和LDA。

pLSA用EM求解，其中的参数Θ和Φ都不是随机变量，而是具有特定的值（对应于前面框架中提到的

$\ln p(X;\theta)$中的参数θ），也不属于latent variables。只有记录每个词对应的topic的随机变量是latent variables。

LDA用VI求解，其中的参数Θ和Φ都是categorical随机变量，具有Dirichlet先验分布（超参数分别是α和β)。Θ，Φ和记录每个词对应的topic的随机变量这三部分一起构成了前面框架中提到的latent variables Z。

EM和VI的区别

EM的思路

在EM中，目标是求解一个特定的θ优化值，使得观测数据的**likelihood**值$\ln p(X;\theta)$ 最大。

EM的思路是，给定当前的$\theta^{(t)}$，将KL变为0，同时得到一个确定的概率分布$q(Z)^{(t)} = p(Z|X;\theta^{(t)})$，这就是E-step。在使用EM的模型中，例如pLSA和GMM，通常$p(Z|X;\theta^{(t)})$是容易计算的。然后将E-step获得确定的分布$q(Z)^{(t)}$代入到ELOB中，然后通过调节模型参数θ来最大化ELOB从而求解出下一轮的$\theta^{(t+1)}$，这就是M-step。

因为分布$q(Z)^{(t)}$是确定之后，$\int q(Z)^{(t)} \ln q(Z)^{(t)} \mathrm{d}Z$是个常数。所以求解ELOB的最大化等于求解$\int q(Z)^{(t)} \ln p(X,Z;\theta) \mathrm{d}Z$的最大化。

VI的思路

在VI中，目标是给出当前观察数据，求解latent variables Z的后验概率$p(Z|X)$。不同于EM，因为latent variables Z包含的未知参数具有先验概率，导致Z的后验概率$p(Z|X)$往往很难直接计算。

VI的思路是，KL等于0，意味着$q(Z)$与$p(Z|X)$完全相同，KL越小越接近0，也意味着$q(Z)$与$p(Z|X)$越近似。因为$\ln p(X)$是一个固定的上界，如果能将ELOB不断最大化，就意味着将KL不断最小化，也就意味着可以$q(Z)$用近似$p(Z|X)$。VI中的变分就体现在调整输入到ELOB中的函数$q(Z)$来使得ELOB最大化，因为$q(Z)$是一个概率分布，当然也是一个函数。

在VI中，通常会选择某种限定类型的函数（概率分布）来作为$q(Z)$。分布类型限定之后，就将优化调整函数的问题转化为优化调整限定函数的参数的问题，这就可以通过常规的优化方法来求解（不同于EM中优化模型参数θ）。

对于常用的mean-field VI方法，对$q(Z)$做了进一步的限定和简化，设定$q(Z) = \prod_{i=1}^{n} q_i(Z_i;\lambda_i)$。VI优化调整每个函数（分布）$q_i$的参数$\lambda_i$。

对连续的随机变量来说，限定分布类型很重要，不过对于LDA里的latent variables都是exponential family，所以根据这一特性来简化计算。