Variational Inference

> *Since there are some problems about mathjax, I also upload the pdf version 「Algorithm」2 Variational Inference.*

背景

变分推断用于解决参数估计以及有缺失数据的问题，它可以看作是EM算法在贝叶斯框架下的一个推广。EM算法通过选择最大化完全似然条件在观测数据的参数来对参数进行推断。而变分推断则尝试用一个函数族对后验来进行逼近。

变分推断能够：

- 为没有观测到的变量（包含参数）的后验得到一个解析的近似，从而能够对这些变量进行推断。
- 可以得到观测变量边际分布的一个下界，这个值能够反映对应变量的重要程度，从而能够用于模型选择。

WHY?

假设X为观测数据，$Z=\{Z_1,\cdots,Z_n\}$为未观测数据或者参数，由于后验$\Pr(Z|X)$的常数项通常是一个难以求解，我们的目标是找到variational distribution $Q(Z)$使得：

$$\Pr(Z|X)\approx Q(Z)$$

Kullback-Leibler Divergence

Introduction of KL divergence

KL散度起源于信息论中，常用于衡量两个分布的相似性。

$$\mathrm{KL}(q\|p)=\mathrm{E}_q\left[\log\frac{q(X)}{p(X)}\right]$$

通过Jensen's Inequality，我们很容易得到$\mathrm{KL}(q\|p)\geq 0$并且等于0当且仅当p和q几乎处处相等。

在上述描述中，我们可以知道KL散度越小，两个分布越接近。实际上，我们可以将KL散度分为三种情况：

- 当q较高，p也比较高；
- 当q较高，p却较低；
- q较低。

第一种情况是我们乐于见到的，因为这代表了p和q比较接近；而第二种情况是我们不希望看到的；最后一种情况在求期望的时候影响较小，因此我们并不关心这一部分。

KL divergence in variational inference

在变分推断中，需要度量的KL散度为

$$\mathrm{KL}(q\|p) = \mathrm{E}_q\left[\log \frac{q(Z)}{p(Z \mid x)}\right]$$

Evidence Lower Bound

由于我们无法直接对观测数据的对数似然进行估计，我们利用Jensen's Inequality得到：

$$\begin{aligned}
\log p(x) &= \log \int_z p(x, z) \\
&= \log \int_z p(x, z) \frac{q(z)}{q(z)} \\
&= \log \left(\mathrm{E}_q\left[\frac{p(x, Z)}{q(z)}\right]\right) \\
&\geq \mathrm{E}_q[\log p(x, Z)] - \underbrace{\mathrm{E}_q[\log q(Z)]}_{Entropy}
\end{aligned}$$

其中，$\mathrm{E}_q[\log p(x, Z)] - \mathrm{E}_q[\log q(Z)]$为ELBO，为观测数据的对数似然的一个下界。于是，我们的目标从最大化$\log(p)$转移为最大化ELBO。

Another viewpoint

接着，我们可以建立ELBO和KL散度之间的联系，利用

$$p(z \mid x) = \frac{p(z, x)}{p(x)}$$

得到

$$\begin{aligned}
\mathrm{KL}(q(z)\|p(z \mid x)) &= \mathrm{E}_q\left[\log \frac{q(Z)}{p(Z \mid x)}\right] \\
&= \mathrm{E}_q[\log q(Z)] - \mathrm{E}_q[\log p(Z \mid x)] \\
&= \mathrm{E}_q[\log q(Z)] - \mathrm{E}_q[\log p(Z, x)] + \log p(x) \\
&= -\left(\mathrm{E}_q[\log p(Z, x)] - \mathrm{E}_q[\log q(Z)]\right) + \log p(x) \\
&= -\mathrm{ELBO} + \log p(x)
\end{aligned}$$

由于$p(x)$是固定的，从这个角度来说，我们要最小化$q(z)$(approximator)和后验之间KL散度等价于最大化ELBO。

Mean Field Variational Inference

接下来，我们介绍着mean field方法，一种常用的求解variational inference的方法。

之前提到之所以要使用variational inference，是因为后验的归一化常数难以估计。因此，我们假设我们的approximator是一族由简单函数的乘积构成的函数：

$$q\left(z_1, \ldots, z_m\right) = \prod_{j=1}^{m} q\left(z_j\right)$$

注：这里$q(z_j)$并不代表marginal distribution，这是用于估计approximator的一个整体。

接着，我们采用坐标下降推断，也就是迭代优化每个vatiational distribution的时候，固定其他vatiational distribution，一次只对一个进行估计。

梯度下降法

首先，分解联合分布：

$$p\left(z_{1:m}, x_{1:n}\right)=p\left(x_{1:n}\right) \prod_{j=1}^{m} p\left(z_j \mid z_{1:(j-1)}, x_{1:n}\right)$$

接着，分解variational distribution的熵：

$$\mathrm{E}\left[\log q\left(z_{1:m}\right)\right]=\sum_{j=1}^{m} \mathrm{E}_j\left[\log q\left(z_j\right)\right]$$

于是，ELBO可以写为：

$$\mathcal{L}=\log p\left(x_{1:n}\right)+\sum_{j=1}^{m} \mathrm{E}\left[\log p\left(z_j \mid z_{1:(j-1)}, x_{1:n}\right)\right]-E_j\left[\log q\left(z_j\right)\right]$$

注意到在这一步只更新$q(z_k)$，因此我们可以提出相关的项，得到：

$$\mathcal{L}_k=\int q\left(z_k\right) \mathrm{E}_{-k}\left[\log p\left(z_k \mid z_{-k}, x\right)\right] d z_k-\int q\left(z_k\right) \log q\left(z_k\right) d z_k$$

求导得到：

$$\frac{d \mathcal{L}_k}{d q\left(z_k\right)}=\mathrm{E}_{-k}\left[\log p\left(z_k \mid z_{-k}, x\right)\right]-\log q\left(z_k\right)-1=0$$

于是可以得到：

$$q^*\left(z_k\right) \propto \exp\{\mathrm{E}_{-k}\left[\log p\left(z_k \mid Z_{-k}, x\right)\right]\}$$

但是由于(Z_{-k}, x)与$q^\star(z_k)$无关，所以

$$q^*\left(z_k\right) \propto \exp\{\mathrm{E}_{-k}\left[\log p\left(z_k, Z_{-k}, x\right)\right]\}$$

参考资料

- Variational Bayesian methods wiki
- A Tutorial on Variational Bayesian Inference
- Variational Inference