Variational Inference 03 Algorithm Solution

Chen Gong

01 December 2019

在上一小节中，我们介绍了 Mean Field Theory Variational Inference 的方法。在这里我需要进一步做一些说明，z_i 表示的不是一个数，而是一个数据维度的集合，它表示的不是一个维度，而是一个类似的最大团，也就是多个维度凑在一起。在上一节中，我们得出：

$$\log q_j(z_j) = \mathbf{E}_{\prod_{i \neq j} q_i(z_i)} \left[\log p(X, Z|\theta)\right] + C \tag{1}$$

并且，我们令数据集为 $X = \{x^{(i)}\}_{i=1}^N$，$Y = \{y^{(i)}\}_{i=1}^N$。variation 的核心思想是在于用一个分布 q 来近似得到 $p(z|x)$。其中优化目标为，$\hat{q} = argmin\ KL(q||p)$。其中：

$$\log p(X|\theta) = ELBO(\mathcal{L}(q)) + KL(q||p) \geq \mathcal{L}(q) \tag{2}$$

在这个求解中，我们主要想求的是 $q(x)$，那么我们需要弱化 θ 的作用。所以，我们计算的目标函数为：

$$\hat{q} = argmin_q KL(q||p) = argmax_q \mathcal{L}(q) \tag{3}$$

在上一小节中，这是我们的便于观察的表达方法，但是我们需要严格的使用我们的数学符号。

1 数学符号规范化

在这里我们弱化了相关参数 θ，也就是求解过程中，不太考虑 θ 起到的作用。我们展示一下似然函数，

$$\log p_\theta(X) = \log \prod_{i=1}^N p_\theta(x^{(i)}) = \sum_{i=1}^N \log p_\theta(x^{(i)}) \tag{4}$$

我们的目标是使每一个 $x^{(i)}$ 最大，所以将对 ELBO 和 $KL(p||q)$ 进行规范化表达：

ELBO：

$$\mathbf{E}_{q(z)}\left[\log \frac{p_\theta(x^{(i)}, z)}{q(z)}\right] = \mathbf{E}_{q(z)}\left[\log p_\theta(x^{(i)}, z)\right] + H(q(z)) \tag{5}$$

KL:

$$KL(q||p) = \int q(z) \cdot \log \frac{q(z)}{p_\theta(z|x^{(i)})} dz \tag{6}$$

而，

$$\begin{aligned} \log q_j(z_j) =& \mathbf{E}_{\prod_{i \neq j} q_i(z_i)} \left[\log p_\theta(x^{(i)}, z)\right] + C \\ =& \int_{q_1} \int_{q_2} \cdots \int_{q_{j-1}} \int_{q_{j+1}} \cdots \int_{q_M} q_1 q_2 \cdots q_{j-1} q_{j+1} \cdots q_M dq_1 dq_2 \cdots dq_{j-1} dq_{j+1} \cdots dq_M \end{aligned} \tag{7}$$

2 迭代算法求解

在上一步中，我们已经将所有的符号从数据点和划分维度上进行了规范化的表达。在这一步中，我们将使用迭代算法来进行求解：

$$\hat{q}_1(z_1) = \int_{q_2} \cdots \int_{q_M} q_2 \cdots q_M \left[\log p_\theta(x^{(i)}, z)\right] dq_2 \cdots dq_M \tag{8}$$

$$\hat{q}_2(z_2) = \int_{\hat{q}_1(z_1)} \int_{q_3} \cdots \int_{q_M} \hat{q}_1 q_3 \cdots q_M \left[\log p_\theta(x^{(i)}, z)\right] \hat{q}_1 dq_2 \cdots dq_M \tag{9}$$

$$\vdots$$

$$\hat{q}_M(z_M) = \int_{\hat{q}_1} \cdots \int_{\hat{q}_{M-1}} \hat{q}_1 \cdots \hat{q}_{M-1} \left[\log p_\theta(x^{(i)}, z)\right] d\hat{q}_1 \cdots d\hat{q}_{M-1} \tag{10}$$

如果，我们将 $q_1, q_2, \cdots, q_M$ 看成一个个的坐标点，那么我们知道的坐标点越来越多，这实际上就是一种坐标上升的方法 (Coordinate Ascend)。

这是一种迭代算法，那我们怎么考虑迭代的停止条件呢？我们设置当 $\mathcal{L}^{(t+1)} \leq \mathcal{L}^{(t)}$ 时停止迭代。

3 Mean Field Theory 的存在问题

1. 首先假设上就有问题，这个假设太强了。在假设中，我们提到，假设变分后验分式是一种完全可分解的分布。实际上，这样的适用条件挺少的。大部分时候都并不会适用。

2. Intractable。本来就是因为后验分布 $p(Z|X)$ 的计算非常的复杂，所以我们才使用变分推断来进行计算，但是有个很不幸的消息。这个迭代的方法也非常的难以计算，并且

$$\log q_j(z_j) = \mathbf{E}_{\prod_{i \neq j} q_i(z_i)} \left[\log p(X, Z|\theta)\right] + C \tag{11}$$

的计算也非常的复杂。所以，我们需要寻找一种更加优秀的方法，比如 Stein Disparency 等等。Stein 变分是个非常 Fashion 的东西，机器学习理论中非常强大的算法，我们以后会详细的分析。